Order Number 8914803

The effect of job information and validity generalization information on expert judgments of employment test validities

Weinberg, Karen, Ph.D.

City University of New York, 1988

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

THE EFFECT OF JOB INFORMATION AND VALIDITY GENERALIZATION INFORMATION ON EXPERT JUDGMENTS OF EMPLOYMENT TEST VALIDITIES

by

KAREN WEINBERG

A dissertation submitted to the Graduate Faculty
in Psychology in partial fulfillment of the
requirements for the degree of Doctor of Philosophy,
The City University of New York

1988

This manuscript has been read and accepted for the Graduate Faculty in Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

8/3/88

Date

Chair of Examining Committee

August 16, 1988

Date

Executive Officer

Walter Reichman

Roger Millsap

Lawrence Zeitlin

Supervisory Committee

The City University of New York

Abstract

THE EFFECT OF JOB INFORMATION AND VALIDITY GENERALIZATION INFORMATION ON EXPERT JUDGMENTS OF EMPLOYMENT TEST VALIDITIES

by

Karen Weinberg

Advisor: Professor Joel M. Lefkowitz

A sample of Industrial/Organizational Psychologists (N=114) participated in a judgment task which required them to estimate the true criterion-related validity of 28 different predictor/criterion combinations for jobs with "clerical" and "sales" job titles. The judges were also required to estimate the 90% confidence interval around each of their validity estimates. Two types of task-related information were manipulated to study the effects on estimated validities: (1) the congruency/incongruency of job title and job SKAPs; and (2) the presence/absence of validity generalization (VG) information.

It was hypothesized that (1) validity judgments for jobs with identical SKAPs would be significantly different when different job titles were provided; (2) validity judgments for jobs with identical job titles (irrespective of SKAP information) would be significantly different when VG information was provided versus not provided; (3)

confidence intervals would be smaller (i.e., confidence greater) when job information was congruent; and (4) confidence intervals would be smaller when VG information was provided.

MANOVA analyses were conducted to determine if the sets of judged validities and confidence judgments were significantly affected by these two types of manipulated information. Then, ANOVA analyses were used to identify the effects of job and VG information on individual validity or confidence judgments.

When sales SKAP information was provided, judges' validity estimates were significantly higher when the "clerical" versus "sales" job title was provided. This test of Hypothesis #1 was not supported for clerical SKAP judgments, however. Validity generalization information had a significant effect on the "clerical" job title validity judgments such that average judgments were higher in the presence of VG information. A smaller, non-significant effect on "sales" job title judgments was observed. Therefore, Hypothesis #2 also received only partial support. Neither of the manipulated factors significantly affected the experts' confidence judgments.

The results were discussed with respect to (a) why some types of task-related information may be more salient when experts estimate criterion-related validities; (b) how information-processing biases may affect the expert judgment process; and (c) why rational validity estimates are not as meaningful as empirically-derived validity estimates. It was suggested that validity estimates derived from this expert judgment process should be interpreted cautiously because (a) rational judgments are not comparable to empirical estimates of criterion-related validity; and (b) experts' validity estimates may not be free of unknown information-processing biases.

TABLE OF CONTENTS

LIST OF TABLES

List of Figures

CHAPTER I

Introduction

As early as 1795 when F.W. Bessel, a German astronomer, found systematic differences in judgments by astronomers of the time it took for a star to pass from one point to another, scientists have known about and considered the effects of human judgments on the outcomes of scientific research (Schultz, 1981). At first, scientists considered the role of the human observer and the nature of judgment errors as methodological issues which, unresolved, would limit the validity of their conclusions about the physical world. However, approximately 100 years ago, Ernst Weber, Wilhelm Wundt and Gustav Fechner made psychophysiological phenomena the prime focus of their research and human judgments the methodology by which to measure covert cognitive processes (Schultz, 1981). Thus, these earliest psychologists identified human judgments as an important tool in psychological research.

The use of introspection to collect data to estimate just-noticeable-differences, average errors, reaction time and other sensory thresholds established a tradition in psychology of using subjective judgments to measure psychological processes and to test hypotheses regarding the structure of cognitive activities and the nature of human behavior. Today, researchers in the field of psychology use introspection only as an ancillary method for collecting data (c.f., Hogarth, 1974; Cook & Stewart, 1975; Payne, 1976; Payne, Braunstein, & Carroll, 1978; Svenson, 1985), primarily because of the development and refinement of more objective methods for measuring human behavior and cognitive processes. In addition, data from introspective-like techniques have come to be viewed as suspect because research on human judgment processes conducted in the past two decades has strongly suggested that human judgments are subject to a plethora of systematic biases and errors that are difficult to avoid (cf., Tversky & Kahneman, 1974; Slovic, 1972a; Edwards, 1968) and that insight into one's judgment processes is unlikely (Nisbett & Wilson, 1977).

Despite the prevalence of non-judgmental methodologies such as empirical decision-making models, and the caveats against total reliance on subjective judgments, Industrial and Organizational Psychologists must frequently resort to the use of human judgments in many areas of psychological research and practice. The judgments of subjects are frequently collected and used to test hypotheses regarding cognitive functioning in such areas as career decision making (Soelberg, 1967), performance appraisal (Schein, 1978), and social perception (Feldman, 1981). Judgments from individuals of varying degrees of expertise and training are frequently used to determine the skill and ability demands of different jobs or tasks (Primoff, 1974; Hogan & Fleishman, 1979), predict the job performance of individuals (Roose & Doherty, 1976), determine the appropriateness of compensation systems (Roose & Doherty, 1978) and identify optimal methods for setting passing scores for employment tests (Reilly, Zink & Israelski, 1984). At every stage of research, psychologists (and possibly other types of experts) are required to make judgments regarding the design, execution and interpretation of research (Faust, 1982). Finally, throughout all areas of professional practice, Industrial and Organizational Psychologists incorporate subjective judgments into decisions about the value of an area of research, the efficacy of one intervention strategy over another, the "meaning" of empirical data, and the proper use and application of their expertise (c.f., Standards for educational and psychological testing, 1985).

A primary, unavoidable subjective element in psychology, as in all other areas of scientific inquiry, is the scientist because "our knowledge of the cognitive processes that the scientist uses or is capable of using lags so far behind our philosophical knowledge [of the scientific method]" (Faust, 1982). The constraints that are imposed by the scientific method on the research design and methodology used to collect judgments from subjects insure the information value (i.e., validity) of this type of data.

And, because the conclusions that can be drawn from these judgments are further defined by statistical, measurement and sampling theories, researchers are provided with stringent rules that guide the processes used to extract meaning from judgmental data.

On the other hand, the information value of judgments made by psychologists and other experts[1] in the course of conducting research and in the professional practice of Industrial and Organizational Psychology can not always be known in the same way. First, the processes scientists use to make judgments may be no more effective than the processes used by untrained individuals, and may be ineffective for providing meaningful information (Mahoney & DeMonbreun, 1977). Second, observations made by scientists in the course of making judgments are never independent of the observer's beliefs, biases, preconceptions and previous experiences; perception, itself, is an interpretative act (Mahoney, 1977). Third, "expert" judgments in applied situations are usually made under circumstances where there is little control imposed over the judgment task: the judgments are rarely elicited or measured under the same rigorous conditions used to collect the judgments of experimental subjects, and psychologists are not sampled in

[1] Throughout this document psychologists and other individuals with special skills or training will be referred to as "experts". The distinction between judgments made by naive individuals and those made by experts is that it is assumed that experts incorporate their special skills or knowledge into their judgments. In addition, it is usually assumed that experts are somewhat homogeneous with respect to certain skills, knowledge, training, etc.

such ways that their judgments can be generalized to other psychologists (Brandt, 1968).

The issue is not whether psychologists should rely on their own or others' expert judgments when the validity of these judgments cannot be rigorously evaluated. Clearly, the incorporation of subjective judgments into research and the practice of Industrial and Organizational Psychology is a necessity and an important component of professional behavior and responsibility. Rather, the important issues are what types of expert judgments are potentially valid and what is the basis for determining the validity of a particular type of judgmental strategy.

To address these issues it is necessary to understand what different types of judgments are made by individuals and how these judgments are used in the field of Industrial and Organizational Psychology.

Types and Uses of Judgments in Industrial and Organizational Psychology

While Guilford (1967) conceptualized all judgment tasks as evaluations involving comparisons[2], his definition of judgment fails to take into account many of the types of judgments that individuals may be required to make or the level of complexity of each type of judgment. Judgment

[2] It has also been suggested that judgment may be confounded with other, more complex cognitive abilities such as problem solving and general reasoning (Vernon, 1950; Johnson, 1955; Northrop, 1976).

tasks that are found outside of laboratory settings probably require other, more complex types of judgments in addition to evaluative judgments. Therefore, to fully understand the value of human judgments, it is necessary to examine the types of judgments that individuals may be required to make in both personal and professional arenas.

Individuals are capable of making many different types of judgments which vary in complexity based on the types of mental processes required and the type and amount of information that needs to be processed. In general, one can determine how complex a particular judgment is by comparing the information required to make the judgment to the judgment, itself. As the necessary transformations to the information input increase both in kind and amount, so does the complexity of the cognitive processes required to make the judgment. In addition, as the amount of information needed to make the judgment increases, or the availability of the information decreases, the complexity of the judgment task increases.

Following is a discussion of classes or types of judgments that are made in both research and applied settings. This classification system was developed by the author to help organize and highlight the important differences in the types of judgments made by individuals and, more specifically, by Industrial and Organizational Psychologists. (An extensive review of the judgment and decision-making literature did not yield any similar type of classification system.) The discussion also includes examples of judgments made in the field of Industrial and Organizational Psychology that fall into each category.

Descriptive Judgments

Descriptive judgments are those that require an individual to perceive and make a judgment about a directly observable characteristic of a stimulus or cue. For example, psychophysiological experiments on reaction time often involve very simple judgments about the presence or absence of a color, simple symbol, or sound. The judgment is a gross determination of whether the targeted stimulus is perceived. The mental processes required to make this type of judgment are not very complex, as the judgment task in this case is almost wholly perceptual. (Descriptive judgments are those that are frequently studied by psychologists interested in basic cognitive abilities such as perception, reaction time, etc.)

In descriptive judgment tasks, complexity is associated with the amount of information that must be perceived and retained in short term (iconic) memory (Neisser, 1967). Descriptive judgments become more complex as a function of the complexity of the stimulus material. A more complex version of the reaction time experiment might require the subject to judge the presence or absence of a class attribute based on an observable characteristic such as color, shape, tone, etc. This type of judgment requires one to compare the stimulus to a category and to make a judgment about the match between the stimulus and the category. Matching/judgment tasks like this are fairly simple because they involve simple distinctions (i.e., red vs. blue vs. green). However, the tasks become increasingly complex as the stimulus object or the type of distinction that must be made become more complex.

The degree to which the judgment must reflect absolute characteristics of the stimulus object also affects the complexity of a descriptive judgment. For example, a subject may be asked to compare two tones or the distance of two objects from a third object and judge which is louder or further away, respectively. In a more difficult version of these tasks, the subject may need to judge the actual notes or distance.

Examples of judgments in Industrial and Organizational Psychology which are similar to the descriptive judgments described above can be found in the areas of job analysis, job evaluation, and performance appraisal. When job incumbents or supervisors are asked to judge whether a task is performed on a particular job or when the task is learned (i.e., prior to or after job entry) they are making descriptive-type judgments (Fine & Wiley, 1971). In addition, global judgments about the similarity of jobs (on the basis of tasks that are performed) are descriptive judgments. Global judgments such as these can be performed fairly accurately by both job incumbents and supervisors (Sackett, Cornelius & Carron, 1981). It appears as if familiarity with the job in question is all that is needed to make such a judgment since the presence/absence of the task or tasks on the job is an observable phenomena. (It should be noted that memory for these observations plays an important role in the accuracy of these "applied" descriptive judgments.) However, for tasks that are performed infrequently or are primarily covert in nature (i.e., cognitive processes), the judgment is more difficult and probably would result in less accuracy.

Judgments made in certain types of performance appraisal systems could also be considered descriptive judgments when those judgments are based on observations rather than evaluations of incumbent job performance. Kane (1984) has proposed a "Performance Distribution Assessment" appraisal system that relies on supervisory ratings of the frequency of subordinate job performance at specified levels for specific tasks. Although many of the judgments in this system are more complex (e.g., judgment of performance standards), it is possible to develop the system so that supervisors need only make descriptive judgments about subordinates' performance behaviors. Descriptive judgments are also made in rating systems that utilize some form of behavioral checklist. The checklist may be presented in the form of a mixed-standard scale (cf., Blanz & Ghiselli, 1972) or forced-choice scales (cf., Bernardin, Morgan & Winne, 1980), in which supervisors indicate if the subordinate has performed a specific activity. Because the value of performance is imbedded in the statement of the activity, the supervisor need only make a descriptive judgment based on his or her observation of the subordinate.

Many job evaluation systems rely on descriptive judgments about the presence or absence of a specific characteristic of a job. For example, point systems require individuals to make judgments about which level of a factor (as defined by specific, observable job features) best represents the characteristics of a job (Treiman, 1979). When the factor is an observable characteristic such as work environment, the judge need only compare the features of the job to the features defining the levels of each factor to judge how the job should be rated on a particular compensable factor. This type of judgment would be considered descriptive.

Evaluative Judgments

Some judgment tasks are focused on characteristics of stimuli that are not directly observable. Before one can make the judgment it is first necessary to transform the input stimulus into information that is in a usable form (i.e., assign a value to it or give it meaning). This transformation, therefore, requires an evaluation of certain properties of the object. These types of judgments require perceptual and short term memory processes to process the raw stimuli and other, higher-level processes to analyze and interpret what was observed. For example, judges have been used frequently in psychological research to rate behaviors. This judgment may be the focal task, or it may be used as a measure of a variable of interest. To judge behaviors as being "aggressive", "altruistic", or "directive" requires an understanding of these categories and the ability to make an association between the observed behaviors and the construct or category it represents. The behavior is directly observed (as it was in a descriptive judgment task), but the behavioral category must be judged by evaluating the match between the definition of the category and the raw stimulus behavior.

Evaluative judgments become more complex when the task requires an evaluative comparison of more than one stimulus or when a single stimulus must be evaluated along multiple dimensions. In comparison types of evaluative judgments, judges are required to perceive the raw stimuli, evaluate the class, category, or value of each stimulus, and then make a comparison judgment of amount, degree, similarity, etc. (Guilford, 1967). If judges are required to make more rigorous evaluations of stimuli (e.g., absolute amount of a characteristic rather than relative amount), the judgment task increases in complexity. Also, if the characteristics that distinguish category membership or differences between stimuli, are very subtle, then the evaluative judgment may be more difficult to make because the interpretation of the of the stimuli with respect to the dimensions under consideration becomes more difficult.

Evaluative judgments about traits or personal characteristics on the basis of observed behaviors have frequently been used to assess managerial potential (Bray & Grant, 1966), appraise job incumbents (Guion, 1961), or determine the qualifications needed to perform specific jobs (Trattner, Fine & Kubis, 1955; Cornelius & Lyness, 1980). In addition, in much of the research on job attitudes, individuals have been asked to make evaluative judgments about characteristics of the work they perform or the organization in which they work, feelings about their jobs, characteristics of their supervisors and work-related personal tendencies or predispositions (c.f., Scarpello & Campbell, 1983; Fried & Ferris, 1986).

Evaluative judgments that are made for the purpose of evaluating personnel or analyzing jobs may be especially complex because the behavioral referents of the dimensions being rated may be difficult to operationalize and, therefore, to identify and evaluate. Also, since behaviors such as these are observed in actual work settings, the context within which the behaviors occur becomes a relevant part of the stimulus. The result is that there is potentially more information in the situation to process in order to provide meaning to the stimulus and, therefore, it is more difficult to identify, analyze and evaluate the stimulus behavior.

Job analysis strategies that depend on judgments about the skills, abilities or personal characteristics needed to perform the job require evaluative judgments. Researchers have frequently found that the source of this information (i.e., job incumbents, supervisors, job analysts) affects the accuracy and consistency with which these types of judgments are made. For example, differences in mean scores, the reliability of judgments, and correlations of judgments with external criteria have been attributed to the type of rater used (Cornelius, DeNisi & Blencoe, 1984; Smith & Hakel, 1979). Not only do different raters have different perspectives and, therefore, different perceptions of the job in question, they may possess different levels of skill for making evaluative judgments of this type (Cornelius & Lyness, 1980).

Inferential Judgments

In both descriptive and evaluative judgments, information that is available from perceptual and memory processes is directly used to make a judgment. <u>How</u> the information should be used is determined by the task. In some types of judgment tasks, judgments can not be made directly from the information or cues provided. Rather, judges must manipulate the information and add to it in a meaningful way to make the judgment. Judgments that require judges to reason from available information are classified as "inferential judgments." The judge may be provided with some of the additional information that is needed to make the judgment (e.g., prior probabilities, frequencies, etc.), or it may be assumed that this information is otherwise available to the judge on the basis of his or her experience or expertise (c.f., research on problem-solving/ judgment in semantically-rich domains [Bhaskar & Simon, 1977; Simon, 1979]). Although information that is necessary to make the judgment is available, how the information should be combined or used is not provided directly to the judge.

Examples of inferential judgment tasks are plentiful in psychological research, especially in the study of judgment processes. Judgments about the diagnostic category of patients (Goldberg, 1965, 1971; Wiggins & Hoffman, 1968), the guilt of a defendant (Kaplan & Kemmerick, 1974), the likelihood of disease alternatives (Johnson, Hassebrock, Duran & Moller, 1982), the quality of pigs (Phelps & Shanteau, 1978), parole risk (Carroll & Payne, 1976); the likelihood of graduate school success (Grinnell, Keeley, & Doherty, 1971), the growth potential of stocks (Slovic, 1969), and the likelihood of success of insurance salesmen (Roose & Doherty, 1976) are the content of a few of the many inferential judgments studied. In addition, the types of judgments, themselves, may vary. The judgments may be predictions of events (Einhorn, 1972), probability/likelihood estimates of events (Johnson et al., 1982), or judgments about how stimuli are related (John & Miller, 1957; Ward & Jenkins, 1965).

In many of these examples of inferential judgments, the judgment tasks are structured similarly: raw cues are provided to the judge in the form of cases which vary along different dimensions. The judge may also be provided with additional information such as base rates or prior probabilities. However, because many of these judgment tasks utilize expert judges, it is frequently the case that such additional information is assumed to be provided by their expertise (Slovic, 1969). The cues are "diagnostic" (i.e., contain information) and the judges must not only identify the relevant cues, they must assign values to the different cues (both "evaluative" judgments) (Einhorn, 1974). Then, they must integrate the various types of information into a unitary value or judgment through an informal, implicit process or by applying some of the general rules which they may have explicitly formulated. Slovic (1969) summarized the later steps in this process in the following way:

> The key to the expert's success resides in his ability to interpret and integrate information appropriately. This means that he must weigh items of information differentially, according to

> their relevance, and must be able to qualify his interpretations of a given fact when other considerations make such qualification necessary. (p. 255)

The complexity of an inferential judgment may also be affected by several different characteristics of the stimulus input: the amount of information that needs to be processed (Oskamp, 1965; Einhorn, 1971), the reliability of the stimuli, the degree to which the cues are interrelated (Phelps & Shanteau, 1978), and the degree to which cues are easily differentiated (i.e., reflect major versus subtle differences in categories or events).

These characteristics affect the ease and accuracy with which information can be extracted from the stimuli and evaluated, as well as the value of such information. For example, as the cues become more interrelated, judges extract and use fewer dimensions (Phelps & Shanteau, 1978). Slovic and MacPhillamy (1974) found that cues were erroneously given more weight when they were common across all stimuli than when they were unique. Also, as cues become more difficult to extract from stimuli, the reliability and validity of the information which goes into the judgment may be affected (Brehmer, 1970). Likewise, when stimuli are themselves unreliable, the information extracted from the stimuli has less value and makes the judgment more errorful (Sawyer, 1966).

Inferential judgment tasks also increase in difficulty as a function of the complexity of the processes needed to combine information and the type of judgment that is required. Judgments for which universal rules for combining information are widely known or easily determined (such as in solving simple math problems) are less complex than judgments that require the judge to devise and/or apply complex algorithms[3]. Research has consistently demonstrated, for example, that judges have difficulty optimally integrating different types of information (Shanteau & Anderson, 1972; Kaplan & Kemmerick, 1974; Ebbesen & Konecni, 1975) and in applying optimal "problem solving" strategies in judgment tasks (John & Miller, 1957; Payne, 1976; Mahoney & DeMonbreun, 1977).

Inferential judgment tasks may differ in difficulty as evidenced by differences in the levels of accuracy with which such judgments are made (Pitz & Sachs, 1984). For example, Lichtenstein and Fischhoff (1976) found that probability assessments made by students were no more accurate when the substance of the judgments were in students' areas of expertise than when students were less familiar with the content of judgment. After compiling results from many studies on judgments of probabilities,

---

3 Of course, expertise is an important factor in determining how difficult such a task is (c.f., Phelps & Shanteau, 1978; Gruppen & Wolf, 1986; Christensen-Szalanski, Beck, Christensen-Szalanski & Koepsell, 1983). DeGroot (1966) suggested that Grandmasters of chess do not even have to consciously think about strategies but, rather, respond almost automatically to the composition of the chessboard.

researchers have concluded that individuals, even trained statisticians, do not make optimal Bayesian judgments (c.f., Slovic & Lichtenstein, 1971). In addition, different types of judgments are subject to different sets of heuristics and biases (Tversky & Kahneman, 1974). Although training is important, training does not always decrease the difficulty of certain types of judgment tasks (c.f., the difficulty with which trained psychologists apply sampling theory to actual sampling problems [Tversky & Kahneman, 1971]). In fact, familiarity with the judgment task or feedback and training may have little effect on overcoming the ubiquity of certain judgment biases and errors (Ebbesen & Konecni, 1975; Alpert & Raiffa, 1968, respectively).

Both in research and in the practice of Industrial and Organizational Psychology there are numerous examples of inferential judgments. For example, in the areas of performance measurement and job evaluation, psychologists must combine and integrate different types of job-relevant information to make judgments about the dimensionality of job performance or compensable job factors, respectively. Even the development of the most elementary supervisory rating system is founded on inferential judgments regarding the relationships between actual tasks, aspects of performance (e.g., quality, productivity, etc.) and overall job performance.

The relationship between assessment center exercises[4] and dimension ratings is generally based on the rational judgments of psychologists. Inferential judgments or empirical analyses (e.g., factor analysis) are used to determine the skill/ability requirements of (usually supervisory or managerial) work. Then, exercises are developed which are designed to elicit behaviors which are indicative of these skill/ability dimensions. The relationships between exercises and behaviors and behaviors and skill/ability dimensions are inferred by psychologists.

In developing criteria for performance appraisal or measurement, psychologists frequently must make inferential judgments. For example, psychologists must judge the dimensionality of performance in order to develop appraisal systems or criterion measures. To do this judges must consider what elements contribute to "overall" job performance across a variety of tasks and activities and determine which of these elements are under the control of job incumbents. Expert judgments are frequently used to determine the importance of each of the dimensions of performance. To make such judgments, it is not simply the case that judges make evaluative judgments. Judges must determine how information (i.e., consequences, cost of errors, etc.) about the different dimensions of job

[4] In many ways, the techniques used to develop work simulations, work samples and other types of performance-related measures of work performance are similar to the techniques used to develop assessment center exercises. Therefore, the discussion here is relevant to these other techniques for measuring job-related performance.

performance for a variety of tasks and activities should be integrated to judge "importance".

Job evaluation systems also frequently rely on inferential judgments by psychologists or other job experts. These judgments are used to determine the dimensionality of compensable performance or to rank jobs to determine their comparative worth (Treiman, 1979). Judging the dimensions of compensable performance is similar, in most respects, to the discussion of the judgment of performance dimensionality above. When jobs are compared in ranking systems, for example, judges must make comparative judgments about the worth of particular jobs. However, these judgments are not evaluative judgments. Jobs are multidimensional stimuli which judges must evaluate across (inherent) dimensions and they must evaluate each dimension to determine its value. Then, the judges must find some system for combining this information into an overall, integrated judgment of job worth. Considering the complexity of this process, and the lack of structure generally provided by this type of job evaluation system, it is not surprising that these systems are subject to many biases and that more objectively-based, standardized systems have been developed.

Inductive Judgments

Inferential judgment tasks require judges to weigh and combine available information to arrive at an ultimate judgment. An assumption of this type of judgment is that given a set of informational stimuli, judges can discover, recall, or apply rules for extracting, weighing and combining the cues (Goldberg, 1968). Although this procedure for eliciting judgments is commonplace in psychological research, in many real life judgment situations it is unlikely that judges will be provided with all of the necessary raw material on which to base their judgments. For example, Edwards (1983) suggested that in the real world of decisions, information is almost always inadequate.[5] In fact, much of the research on judgments of probability is unconcerned with how information is acquired and, therefore, how attention and/or memory processes and biases in information acquisition affect judgment outcomes (Einhorn & Hogarth, 1981).

When the judge must search the environment or through his or her memory for relevant raw data from which to make a judgment, judgments may be based on incomplete information or on potentially irrelevant stimuli. Therefore, one major way in which inductive judgments differ from inferential judgments is that when information must be sought as part of the judgment process, the judgment task is substantially changed. Ebbesen and Konecni (1980), for example, have

[5] Hogarth (1980) concluded that many of the prescriptive judgment models have been found to be appropriate because they have only been tested in laboratory-like settings. He suggested that optimal judgment models probably cannot work in "continuous" environments where multiple sources of uncertainty are operating (e.g., what information is needed, how accurate is the information, etc.).

found major differences in results of similar judgment tasks (e.g., setting of bail, driving a car) carried out within laboratory versus natural settings. Inductive judgments are labeled as such because, through their search procedures, judges gain access to only some of the stimuli that are relevant to the judgment task and must make judgments on the basis of this limited information. The judgments they output contain more information than the original set of stimuli on which the judgment was based.

In inferential judgment tasks, judges may decide information has little value and either not use it or give it a low weight in the course of integrating information. This is the only way in which judges evaluate information for inclusion in inferential judgment tasks. However, one important way in which inductive judgments differ from inferential judgments is with respect to the means of acquiring and evaluating information. First, judges must decide what information is appropriate so that search activities can focus on extracting relevant information from memory or from the environment. Second, because judges must search for the stimuli on which the judgments are based, they must determine what type of search procedures are appropriate and when the activities should stop (i.e., when sufficient raw information has been sampled and/or what is the marginal utility of additional information) (Upshaw, 1975).

Inductive judgments differ from inferential judgments not only in how stimulus information is acquired, but in how information must be integrated and combined. When using information to produce a judgment, judges must determine the adequacy and representativeness of the stimuli. This includes determining if information is systematically missing and, if so, how this information could be expected to affect judgment outcomes (Upshaw, 1975). The judgment task, therefore, involves more than simply weighing and integrating available information, it involves extrapolating from this information to what the information might actually be. Inductive judgment tasks are, inherently, probabilistic because the information from which judgments are made is not known with certainty and, therefore, the information must be used in a way which takes this uncertainty into account. Unlike inferential judgments, where uncertainty may only be present when integrating information, inductive judgments require that the uncertainty of information be taken into consideration both when weighing and combining information.

To determine how inductive judgments increase in complexity one must focus on the source(s) of stimulus information and the factors that introduce greater uncertainty into the processes of searching for and using relevant information. As the number and diversity of events within an environment increase, the more difficult it is to identify relevant stimulus information (Neisser, 1964). Also, if information in memory is not coded in such a way that different stimuli cannot be easily recalled, then the more difficult it is to identify and extract relevant information (Neisser, 1964). (It has been suggested that another way in which laboratory versions of tasks differ from those made in the real world is that laboratory judgments generally do not require judges to have and apply large amounts of prior, potentially task-related information [Bhaskar & Simon, 1977]). Information acquisition activities are also plagued by systematic biases in deciding what data are relevant and in sampling cases (Crocker, 1981). For example, individuals tend to seek out instances that confirm their own beliefs or the proposition they are attempting to prove (Wason & Johnson-Laird, 1972; Mynatt, Doherty, & Tweney, 1977; Crocker, 1982) and the sample cases they accumulate tend to be non-representative (Crocker, 1981). In addition, even trained individuals have been shown to generalize from too few cases (Kahneman & Tversky, 1972); i.e., they do not take into consideration the probabilistic nature of the information.

Examples of inductive judgments are available in the area of employment test development. Consider, for example, the setting of passing scores on training mastery tests or job knowledge tests in order to define qualified versus unqualified examinees. One technique is to use expert judgments to determine the score at which minimally competent examinees would score by estimating their likely success on each item (cf., Angoff, 1971; Nedelsky, 1954) or by statistically analyzing the scores of individuals who are independently judged to be qualified versus unqualified (c.f., Cronbach & Gleser, 1965). Both of these approaches involve inductive judgments. For example, to judge whether minimally qualified examinees can answer an item correctly, or the probability of them choosing each item alternative (two different methods of expert judgment), requires a judge to acquire and process information concerning the factors that would affect examinees responses. The information should optimally include the range and variability of test-specific skill or ability and general test-taking ability of the examinees, difficulty level of the item, difficulty level of the alternatives, the degree to which these sources intercorrelate, etc. All of this information must be weighed and combined and a judgment made. Clearly, if judges are carrying-out the judgment task in a meaningful way, they must search for, accumulate, evaluate and make judgments from limited, often inaccurate, unrepresentative or unreliable information.

Another example of the use of inductive judgments in test development is in the area of judgmental analyses of item bias. A common procedure for determining the possibility of bias in tests is to have subject matter experts judge each test item for the presence or absence of bias (Tittle, 1982). These judgments can take different forms. For example, judges may subjectively determine whether an item may be offensive or apparently biased

(Tittle, 1982). (Judgments by experts about the potential offensiveness of a test item are more characteristic of evaluative than inductive judgments, however.) In addition, they may be asked to judge whether items have characteristics that would make them irrelevant to the intended construct for different racial, sexual, or socio-economic groups.

The results of such judgmental processes regarding the construct validity of test items have been disappointing. Judges are usually fairly inaccurate at determining which items were more difficult for minority students. Also, inter-rater agreement is fairly low and even lower among judges of different ethnic backgrounds (Sandoval & Whelan-Mille, 1980). The complexity of this type of judgment is evident when one considers the types of information that judges must be able to process to arrive at accurate judgments of item bias. They need to have knowledge of the intended construct, a detailed understanding of the special group(s) for which bias is being evaluated, information regarding the skill, ability or knowledge characteristics of the specific sample for which the test would be used, etc. To analyze each item this information would have to be weighed and integrated to determine whether the apparent bias would actually be realized in different subgroup performance on that item.

Inductive judgments are not as prevalent in the professional and technical practices of the field of

Industrial and Organizational Psychology as other types of judgments. In general, these types of tasks are more likely to be found in practical decision-making situations than in formal, standardized methods or practices. One reason that psychologists may not have adopted inductive judgments into their technical practices is that inductive judgments are especially subject to error. Unlike inferential judgments which involve the application of one's expertise to already available information, inductive judgments require experts to decide what the stimuli should be before they can apply their expertise to processing it. The data collection step and the probabilistic nature of the information both add additional, important sources of error into the judgment process.

Error in inductive judgments is the result of the fact that expertise is probably more homogeneous with respect to using information than it is to acquiring information. Einhorn (1974) has suggested that for individuals to be considered "experts" at a particular type of judgment, they need to be able to accurately extract the relevant cues from their (noisy) context. While this is a reasonable standard for defining expertise in inductive judgment tasks, it is likely that the number of experts that could be so identified are quite few for many types of inductive judgment tasks.

The following chapter describes an inductive judgment task that has recently received attention in the professional research literature. Because of the difficulties inherent in making inductive judgments, there is some question as to whether this particular type of inductive judgment should be incorporated into the professional methods within the field of Industrial and Organizational Psychology. This paper and the research being proposed herein was designed to consider this question, and to explore the utility of this judgmental strategy.

CHAPTER II

Expert Judgment of Test-Job Validities

In addition to local studies of criterion-related validity, the 1978 Uniform Guidelines offer several other strategies for establishing the value of an employment selection procedure (Equal Employment Opportunity Commission, 1978). One option--the demonstration of content similarity between the test and the job performance domain--is only useful when one can clearly document the direct link between test and job content. A second alternative is to engage in a cooperative validation effort with other employers. This strategy can only be reasonably applied when jobs and employment needs are comparable across a sufficiently large number of employers. The third alternative--transporting validity--has recently become more feasible with the development of validity generalization methodology which may allow inferences from test scores to be transported across different situations (Schmidt & Hunter, 1977).

Validity generalization not only provides a methodology for inferring whether between-study differences may be due to situation-specific moderators, it allows one to estimate the "true" relationship between a test and a measure of job performance. This technique is particularly useful where observed validities for a particular test and job are subject to a number of attenuating influences (i.e., sampling error, unreliability, range restriction, etc.) and, therefore, provide a less than perfect estimate of test validity. Although validity generalization is an important, new strategy for obtaining validity evidence, it can be applied only when sufficient validity data are available.

Recently, Schmidt, Hunter, Croll, and McKenzie (1983) suggested that experienced experts within the fields of testing or personnel psychology might be used to estimate validities where empirical validation strategies are not possible or where existing estimates of validity have little information value (i.e., are based on small samples). They proposed that if experts are capable of judging validities with sufficient accuracy, then their estimates would contain more information than validities derived from typical small sample, local validity studies[6].

Two issues are critical to determining whether such a rational estimation methodology should be used to establish the validity of tests: (1) whether judges can accurately estimate validities; and (2) whether rational estimates of test-job relationships contain the same information as empirically-derived validities. Clearly, for expert

[6] Schmidt et al. (1983) and Hirsh, Schmidt & Hunter (1986) described small sample validity studies as having sample sizes less than or close to the average (n=68) found in Lent, Aurbach & Levin (1971). However, since the advent of federal regulation, the average sample size of local validity studies may have increased.

judgment to be a useful replacement for empirical validation studies, judges' validity estimates must be comparable in accuracy to empirically-derived validities. However, accuracy is only a necessary condition for establishing the value of this methodology, it is not by itself sufficient for concluding that such estimated validities are as meaningful as the empirical estimates they would replace. Therefore, to fully justify the use of judged validities, one must also establish that experts' validity estimates allow inferences about the meaning of test scores comparable to those justified by empirical validities.

The Accuracy of Expert-Estimated Validities

To date, three studies have been conducted which were specifically designed to examine the accuracy of rational judgments of validity (Parry, 1968; Schmidt et al., 1983; Hirsh, Schmidt & Hunter, 1986). Each of these research efforts was concerned with establishing the judgment accuracy of samples of judges with professional training in personnel psychology or testing. Although none of these studies focused on establishing the evidential basis for using rational validity estimates, they did acknowledge some of the questions pertinent to understanding the meaning of judges' estimates. A discussion of the inferences regarding expert accuracy that can be drawn from these studies follows.

Sources of Error in Experts' Validity Judgments

As Schmidt et al. (1983) have pointed out, there are two sources of error that can affect the accuracy of rational judgments of validity from groups of experts. As can be expected, the mean judgment of validity from any size group of judges for a specific test-job combination will deviate somewhat from the true validity or population correlation. This is the systematic error in the judges' mean estimate. Also, individual judges' estimates will deviate from the mean judgment for the population of judges. This random error can be reduced by averaging the judgments of more judges as evidenced in the formula for the standard error (Schmidt et al., 1983):

$$SE = \sigma_j / n^{1/2}$$

where σ_j is the standard deviation for a single judge and n is the number of judges contributing to the mean judgment.

The systematic error in validity judgments is unaffected by the number of experts used to estimate a validity because increasing the number of judges for whom validity estimates are averaged does not alter the population average. It is important to point out, however, that in measuring systematic error, one must be able to assume that samples of experts are randomly selected from the population[7] of expert judges. The accuracy of rational judgments is, therefore, "fundamentally bounded" by the degree to which experts' average validity estimates can be expected to deviate from the actual validity values they are attempting to estimate (Schmidt, et al., 1983)[8]. The result is that the accuracy of expert judges and, therefore, the utility of their validity estimates, is determined by whether their validity judgments contain a small or large amount of systematic error.

Research on other types of judgment tasks, particularly performance rating, has suggested that the accuracy of an individual rater on a single rating is comprised of: (1) the overall level of ratings for the particular rater; (2) the overall level of ratings for each of the dimensions being rated for the particular rater; and (3) the stimulus (e.g., person) being rated (Cronbach, 1955; Becker & Cardy, 1986). If one extends this model to judgments of test-job

---

[7] Whereas components of expertise that are critical to accuracy should be used to identify populations of experts in research concerned with "expertise" (Einhorn, 1974), it is unlikely that initially this information will be available to researchers. Because researchers will be most concerned with generalizing results to the population of individuals who are likely to be used in applications of this methodology, the population can be defined by the same professional characteristics that would define samples of judges selected by testing organizations (i.e., academic credentials, job experience, etc.).

[8] Schmidt et al. (1983) argued that judges should be asked to estimate observed validities, primarily because they probably "think" in terms of observed rather than true validity. However, their assumption is unfounded. In fact, one might easily argue that for experts to substantiate their expertise in judging validities, they should be able to estimate observed, true, and even incremental validities accurately.

validities, one would identify the following components of judgment accuracy for a single judge: (1) the overall level of validity estimates for the particular judge; (2) the overall level of judged validity for a particular test; (3) the overall level of judged validity for a particular criterion; and (4) the job for which validity judgments are being made.

If the concern is with the systematic error in experts' judgments, then the unit of measurement for identifying sources of judgment error is the sample of expert judges rather than the individual judge. Errors in experts' average judgments of validity may result from any of the components of judgment accuracy listed above; however, these errors would be due to population effects as opposed to the effects of individual raters. If one is attempting to predict whether the systematic error in experts' judgments of validity will be high or low, one must identify the ways in which experts' judgments may be adversely affected in a systematic way.

Expertise implies that judges are fairly homogeneous with respect to relevant training, experience, and skills. As a result, experts may share not only a core knowledge and set of skills that will allow them to perform similarly on a judgment task relevant to their expertise, they may also share similar beliefs, biases, misinformation, or less-than-optimal information processing strategies. One might expect, therefore, that validity judgments from groups of judges may be systematically affected by some of the characteristics associated with their expertise.

For example, when the average judged validity deviates from the validity value being estimated, judges' estimates may have been affected by a belief about the general level of validity of selection strategies (#1, above), stereotypes associated with the job, test, and/or performance measure for which validity is being judged (#2, 3, and 4, above), similar biases when processing information, or access to or reliance on similar information when making these complex judgments. In addition, it is also important to note that a plausible explanation for the level of systematic error found in experts' average estimates may be the non-representativeness of the expert sample. For example, expert judgments from samples in which judges with particular experiences, training, or task-relevant abilities are over- or under-represented will tend to reflect these sampling biases if such factors are critical to the types of judgments being made.

Measurement of Expert Accuracy

Parry (1968) attempted to measure the accuracy of validity estimates from ten "test experts" for three different jobs and six different tests (i.e., three tests measured cognitive abilities and three tests measured supervisory judgment). Two comparisons of rational and empirical validities were made: (1) experts' mean judgments for each test-job combination were compared to the observed validity to measure absolute accuracy; and (2) individual judges' rank-ordered validities were compared to the observed rank-order to measure relative accuracy. The validity standard against which judged validities were compared was the corresponding observed validity from each of three samples of applicants (N=24, 25, & 52). This methodology, because it utilizes observed, small sample validities (which themselves contain sampling error), relies on a less-than-perfect standard of validity and, therefore, does not provide a meaningful evaluation of expert accuracy. The result is that it is impossible to determine the accuracy of Parry's judges or to make conclusions about the information value of their validity estimates.

Schmidt et al. (1983) and Hirsh et al. (1986) examined expert accuracy in judgments of validities for nine different jobs by groups of highly experienced and less-experienced judges, respectively. Their predictors consisted of six subtests of the Navy Basic Test Battery - three of which are traditional cognitive ability tests and three of which "are to varying degrees measures of knowledge and information (mechanical knowledge, shop information, and knowledge of algebra and electronics, respectively)" (Schmidt et al., 1983, p.593). The criterion measures used for all of the validity judgments were written tests of training performance. To ameliorate the problem of sampling error in the set of validities to which judges' estimates would be compared, Schmidt et al. (1983) and Hirsh et al. (1986) used validation data which included validity samples of at least 2,000 examinees per test-criterion combination within each job. These large samples allow for a "sampling error-free validity standard against which the accuracy of judges' estimates could be assessed" (Hirsh et al., 1986, p.6).

The procedure developed by Schmidt et al. (1983) consisted of averaging the absolute deviations of each rational validity from each large-sample, empirical validity. Therefore, in both Schmidt et al. (1983) and Hirsh et al. (1986), accuracy was a summary measure of the average deviation of judges' estimates across 54 test-job combinations. The researchers were then able to compare the error in judges' estimates to the error found in typically-sized validation studies and determine the sample-size equivalence of any size group of expert judges. Where judges were as accurate or more accurate than empirical samples of typical sizes available to organizations, the researchers suggested that rational judgments would have more utility than empirically-derived validities.

Schmidt et al. (1983) and Hirsh et al. (1986) were able to overcome the problem of establishing an adequate validity standard against which judges' rational estimates could be compared by identifying large sample validities which were relatively free of sampling error. However, their strategy for measuring expert accuracy (i.e., obtaining the average deviation of judgments across 54 test-job combinations) obscures the accuracy with which judges estimate validities for each of the nine different jobs, for each of the six different tests, and, most importantly, for each of the 54 test-job combinations. The measurement of expert accuracy using that kind of averaging ignores the possibility that different sources of rating error may be affecting experts' judgments for different combinations of tests and jobs. In addition, the assumption that experts are homogeneous with respect to aspects of expertise related to this judgment task could not be examined because individual differences in accuracy between judges (measured across the 54 test-job combinations) could not be assessed.

The accuracy with which samples of judges estimate validities for jobs, tests, and job-test combinations is not an inconsequential question in view of the proposal for how expert judgments would be used and in light of the various sources of judgment error. More specifically, if rational judgments will be used as substitutes for empirical validity studies, then it would be essential that expert deviations from empiric validities are minimal for validity judgments of specific test-job relationships. The degree to which judges are accurate at this level of judgment (i.e., particular test-job combinations) will determine the value of using rational versus empirical procedures for obtaining evidence about the validity of a selection procedures.

However, this issue was not addressed by any of the three studies on expert accuracy of judged validities.

The Importance of Understanding the Expert Judgment Process

Empirical research is an important source of two types of information needed to understand experts' validity estimates: (1) descriptions of the processes that judges use to produce validity estimates; and (2) descriptions of the factors--judgment conditions and sampling issues--that affect judgment output. Of these two types of information, knowledge of the processes used by judges to produce validity estimates is more critical to establishing how experts' judgments should be interpreted. Even if judges' validity estimates are found to be accurate across a wide variety of judgment conditions, without an understanding of the processes they use we cannot reasonably conclude that their estimates are comparable in meaning to empirically-derived validity estimates. The reason for this is described in detail below.

Generalizing Expert Accuracy

Although researchers have not as yet provided sufficient evidence to allow the conclusion that judges can accurately estimate test validities, future efforts may substantiate that judges are accurate producers of validity estimates. In fact, for reasons that will be discussed at length in other sections of this paper, additional research efforts may indicate that judges are fairly accurate at estimating validities for certain tests and jobs.

To conclude that rational estimation procedures are effective strategies for obtaining validity evidence, it is necessary to demonstrate expert accuracy across a variety of tests, criteria, jobs, and research conditions (i.e., types of information provided, types of judgments being made, etc.). Even then, it is only possible to generalize the level of expert accuracy observed in these studies to those situations that share similar characteristics to the situations in which the validity judgments were collected. However, this requirement presents two dilemmas.

First, to generalize expert output one must exhaustively identify the characteristics of the judgment task and expert sample that are associated with their output. Yet, if we do not have an understanding of how judges produce validity estimates, we can not expect to be able to fully identify the factors within the judgment situation which are critical to the accuracy of experts' judgments. It is plausible that factors that are not obvious characteristics of the judgment task may be operating to inhibit or promote expert accuracy[9]. The

[9] As the previous discussion pointed out, such factors as shared biases, beliefs, or non-optimal judgment strategies may reduce judgment accuracy. If such factors are created by aspects of the judgment situation (i.e., the organization of the judgment task, the information that is provided, etc.), then one might expect systematic error to be

result is that ignorance about the processes judges use to produce validity estimates may lead to mistakes concerning conclusions about the accuracy of judges in particular judgment situations, and inappropriate inferences from experts' validity estimates.

Second, expert accuracy can only be substantiated where validity standards are available. Therefore, we can only document expert accuracy for those jobs, tests, and criteria for which validity evidence already exists; i.e., where expert judgment is not critical to establishing test validity. However, the existence of validity standards may affect the accuracy of experts' validity judgments because experts presumably can have access to and use this information when making their estimates. Therefore, an important condition associated with research in this area--the availability of empirical validity "standards"--is potentially confounded with expert accuracy. Without an understanding of how validity estimates are produced, it is not possible to understand how available information impacts on experts' judgments or, as a result, whether it is appropriate to generalize expert accuracy to those situations in which it is likely to be most useful and most

artifactually induced. However, if one is unaware of how judges are using the information or responding to the particular judgment task, then it might be assumed that their inaccuracy is "real". Generalizations to real applications from such studies will be invalid if they do not take into account those conditions which are critical to the judgment task.

frequently used--where existing validity information is not available.

Establishing the Meaning of Expert-Estimated Validities

Test validities provide the evidential basis for making inferences about the meaning of test scores (Messick, 1975). Validities that result from empirical studies derive their meaning from the means and methods used to collect, analyze, and make conclusions about the relationship between test and criterion scores. For expert-estimated validities to allow for the same type of inferences about the meaning of test performance, the processes that experts use to derive their validity estimates must also be identified and understood. Without such an understanding, it is impossible to conclude that judges' validity estimates provide useful information about the value of tests in selecting individuals for specific jobs.

It is necessary to establish that judges are using processes that are congruent witn assumptions concerning the meaningfulness of their judgments. Therefore, even if one has substantiated which information is associated with judgment accuracy, what skills, knowledge, or experience is required, and whether different types of tests, performance measures, and jobs affect the accuracy of experts' judgments, one must also substantiate that experts used meaningful processes in those conditions where accuracy was observed. This latter point deserves further clarification because it is fundamental to conclusions concerning the meaning of rational validity estimates.

The Expert Judgment Process

Based on their suggestion that rational judgments of test validity can be substituted for empirical validities of equal accuracy, one must assume that Schmidt et al. (1983) and Hirsh et al. (1986) have concluded that estimates from both processes are comparable in meaning. More specifically, the researchers imply that rational estimates, like those derived from empirical research, are capable of taking into account all of the numerous factors that affect test and job performance. Although empirical evidence describing how experts judge validities is not available, it is possible to infer what types of judgmental processes experts may and may not be using by considering what processes would plausibly result in accurate judgments.

Judgment of validities as an inductive judgment task. If, in fact, the processes judges use to estimate validities "mirror" the empirical validation process, then the validity judgment task would necessarily be "inductive". In the initial steps of the process, judges would be using information provided in the judgment task to identify all of the different factors that affect the particular test-job validities being judged. Then, they would have to access all of this information in memory, accurately code and evaluate the relevant raw data, determine the representativeness and completeness of the information, and, lastly, integrate all of the relevant information into a final validity estimate. Figure 1 is a simplified model that represents some of the different factors that influence true validities and, therefore, that are implicitly or explicitly taken into account when empirically measuring the relationship between test and job performance in any given job. This model, therefore, also depicts the types of information that judges must have access to if they derive validities in a manner that is analogous to the empirical validation process.

Clearly, psychologists need not refer to the research literature concerning complex judgment and decision-making to conclude that judges, no matter how expert, will not be able to accurately process all of the information contained in this model. Cumulative research on complex judgment processes has demonstrated that individuals are very limited in their ability to accurately attend to, collect, recall, and integrate information (John & Miller, 1957; Payne, 1976; Kassin, 1979a; 1979b; Crocker, 1981). In addition, because individuals tend to simplify cognitive tasks by limiting the information processing demands of these tasks, it is possible that judges would not even attempt to attend to or use all of the information in the model (Miller, 1956). Given the added complexity of judging the covariation between sets of factors in the model, and integrating this information to form a single determination of the criterion-related validity coefficient (Smedslund, 1963), it is highly unlikely that experts who produce accurate validity estimates could be using processes that are in any way isomorphic with the empirical validation process.

Several questions then remain: How are judges producing estimates of test-job validities? How are judges producing accurate estimates of test-job validities? And, more importantly: Do the processes judges use to estimate validities provide the type of validity evidence that is needed for making meaningful inferences about the nature of test performance?

Judgment of validities as an inferential judgment task. If cognitive limitations prevent personnel testing experts from using inductive processes as described above to estimate validities accurately, then accurate expert judgments most likely result from judgmental processes that do not place such a great demand on assembling, evaluating, and integrating relevant raw data (Lindblom, 1964; Shanteau & Anderson, 1972; Payne, 1976; Phelps & Shanteau, 1978; Hogarth, 1980). For example, if judges were provided with all of the necessary information for judging validities (i.e., were not required to assemble it themselves, make inferences about the value or exhaustiveness of the information, etc.), then to produce validity estimates they would only need to integrate the information into a single validity value.

Even if the validity judgment task was designed to be inferential rather than inductive, it would be highly unlikely that judges could accurately estimate validities. First, to ensure that judges' estimates were based on all of the relevant information, it would be necessary to provide judges with an extremely large amount of raw data (see Figure 1). As a result, it would be improbable that they could effectively process all of the information or combine it appropriately. Research has shown that when presented with multiple pieces of information, individuals generally cannot integrate them appropriately (Sawyer, 1966; Einhorn, 1971, 1972; Kaplan & Kemmerick, 1974; Phelps & Shanteau, 1978). Secondly, even if the information was not that overwhelming (i.e., if the task-related information was intentionally limited), their estimates would probably not be accurate approximations of the actual statistical relationships between predictors and criterion. This assumption is based on research that has systematically demonstrated that individuals--even those trained in statistics--have a great deal of difficulty and are rather inaccurate in making statistical inferences from available data (Smedslund, 1963; Jenkins & Ward, 1965; Ward & Jenkins, 1965; Peterson & Beach, 1967). Judges appear to be equally inaccurate in judging covariation no matter how the information is presented (i.e., in scattergrams or as raw data).

Clearly, if all of the information was available so that it could be provided to judges for making rational validity estimates, it would be more reasonable to combine the information statistically rather than to use judges to make inferential judgments (Sawyer, 1966). In addition, since it would be reasonable to conclude that judges could not use inferential judgments to produce accurate estimates of test-job validities, how they are producing accurate estimates remains unanswered.

Judgment of validities as an evaluative judgment task. It is very possible that judgmental processes used by judges to estimate validities do not require the complex information processing activities that are involved when making inductive or inferential judgments. Judges may, in fact, reduce what appears to be a very complex judgment task into one which requires merely processes that are somewhat similar to the "evaluative" judgment task described earlier. For example, it is reasonable to assume that expert judges may assign a validity value to a test-job combination solely on the basis of information that is directly recalled from memory or provided as part of the judgment task. Information about the predictor, criterion, and/or job provided in the judgment task would serve as cues for identifying the most appropriate validity value.

In addition, the judges may use evaluative-type judgments to determine whether they should alter or adjust initial validity estimates ("anchors") that have been provided or recalled (Lichtenstein & Slovic, 1971). If so, this would be accomplished by comparing the essential characteristics of the "anchor" validity to the characteristics of the validity being judged. This evaluative judgment would be used to locate important differences (e.g., skill or ability requirements of jobs, sample or research characteristics, etc.) that may, as judged by the expert, affect the level of validity expected in the judged validity.

It should be emphasized that, for a variety of reasons, judges may not get involved in an adjustment process when judging validities. For example, judges may reject the notion of "situational specificity" and may, therefore, assume that an appropriate "anchor" value is an accurate estimate of validity across many different situations. Secondly, judges may be unwilling to modify validities rationally because they place greater weight on empiric validities that they can recall or are directly provided in the judgment task. In other words, empiric validities may be more compelling than the situational, test, or criterion differences between the recalled and judged validity. Thirdly, it is possible that the amount of information that would need to be incorporated into the judgment, and its complexity, would present an impossible cognitive task--so they are "ignored."

Both of these judgmental processes (i.e., selecting an initial "anchor" and determining whether to "adjust" it) involve evaluative judgments because the expert merely needs to evaluate the category of the validity (i.e., what type of predictor, what type of criterion, etc.) to assign a value to it. However, this process may differ somewhat from other evaluative judgments if the validity values are not provided in the judgment task but, rather, need to be recalled from memory.

It is probable that judges use an "evaluative" judgment strategy (in comparison to the "inductive" and "inferential" strategies described above) when producing validity judgments and possible that this strategy could result in somewhat accurate estimates of validity. First, there is a strong tendency for individuals to attempt to minimize the cognitive complexity of judgment tasks by adopting strategies that help simplify the amount of information they must collect and integrate and the types of processes they use to derive their judgments (Lindblom, 1964; Shanteau & Anderson, 1972; Payne, 1976; Phelps & Shanteau, 1978; Hogarth, 1980). Recalling validities and possibly adjusting them based on their most salient features is cognitively simpler than collecting and integrating raw data, estimating covariations among all relevant factors, and combining these into an overall estimate of validity. Also, validities published in the literature and known based on professional experience will be more salient that the other types of information needed if judgments are "inductive" or "inferential." If judges use an evaluative judgment strategy, they may be able to produce fairly accurate "estimates" of validity. Judgment accuracy would be contingent on the availability and appropriateness of the validity "anchor" initially chosen and the appropriateness of the adjustments made to the anchor. One could imagine that if the range of true validity across many different types of predictors and criteria is fairly limited, then experts may be accurate if they base their estimates on a reasonable, general validity value[10].

Summary. Although it has not as yet been determined how experts produce their validity estimates, from the preceding discussion it is obvious that judges do not in fact inductively "estimate" validities in a manner which is analogous to the empirical estimation process. Rather, judges must rely on information provided in the judgment task to integrate or infer what validity values should be, or to recall actual validity values. It is likely that if either of these processes are used, that judges' validity

[10] Several experts who were contacted after their participation in the validity judgment task suggested that "if you just start with a value of .30, you can't be far off." It appears that experts may, in fact, have a generalized validity value that is used as an initial "anchor." Moreover, this "anchor" apparently derives from (recall of) empirical data, reviews of validity findings, etc., rather than "rational judgment" based on the particular situation.

estimates will be highly dependent on the informational features of the judgment task.

In addition, the type of judgment process employed by experts and, therefore, the type of judgment made, may vary across experts and designs of validity judgment tasks. While the empirical validity model is invariant across research studies (i.e., although the importance of different factors may change from one validity study to the next, the measurement, research, and statistical models are invariant), one can not presume that expert judges employ similar processes when they are faced with similar validity judgment tasks. Clearly, experts' validity estimates are not comparable in meaning to empirically-derived validities, no matter how accurate they are. And, more importantly, we do not know what their estimates mean with respect to the value of test scores.

Figure Caption

Figure 1. Model of the determinants of true validity

Population Characteristics

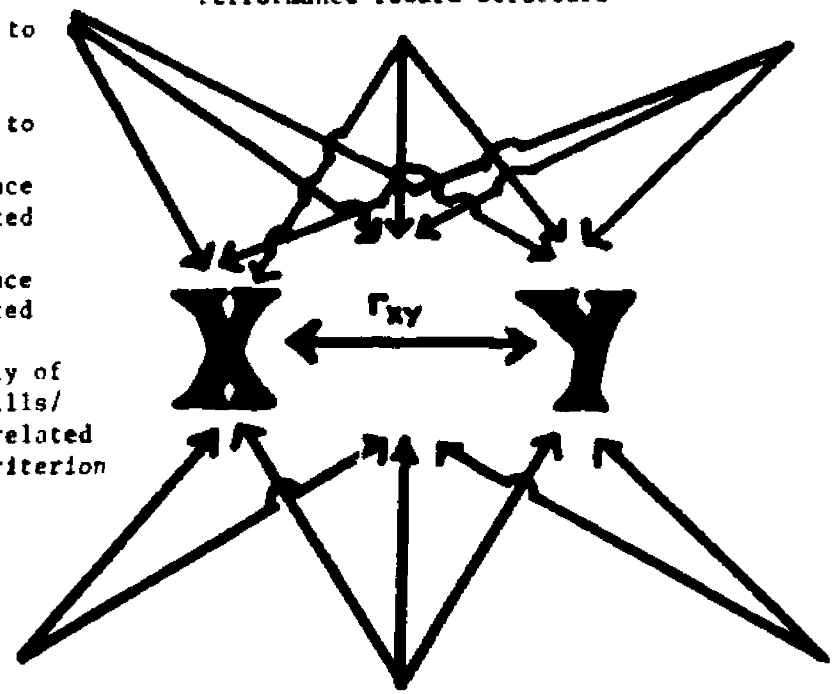
- Level of experience or skill/ability related to test content
- Level of experience or skill/ability related to criterion content
- Variability of experience or skill/ability related to test content
- Variability of experience or skill/ability related to criterion content
- Level and/or variability of other experiences, skills/abilities indirectly related to predictor and/or criterion measurement

Organizational Characteristics

- Amount and type of training
- Nature of supervision
- Performance reward structure

Research Study Characteristics

- Type of research study (e.g., predictive, concurrent)
- Characteristics of test administration
- Characteristics of criterion data collection
- Time between collection of predictor and criterion collection



Predictor Characteristics

- Type of predictor (e.g., cognitive ability, psychomotor skill, background & experience)
- Factorial composition
- Reliability
- Measurement characteristics (e.g., use of raters, difficulty level of items)
- Test characteristics (e.g., speededness, required reading level, length)

Job Characteristics

- SKAPs required for successful job performance
- Inhibitors/facilitators of job performance
- Variability in tasks and SKAPs required across job incumbents

Criterion Characteristics

- Type of criterion (e.g., work sample, supervisory ratings, production records)
- Factorial composition & relevance
- Reliability
- Measurement characteristics (e.g., use of raters, contamination, difficulty level)
- Criterion characteristics (e.g., when collected, used for multiple purposes)

CHAPTER III

PROBLEM STATEMENT AND HYPOTHESES

To meaningfully interpret the validity judgments produced by judges it is critical that we identify the factors that affect their judgments. By so doing, we can begin to understand the processes that may be involved in making such judgments. Although the role of judgment task structure has not previously received any attention by researchers interested in studying the value of rational estimation procedures, it would be important to understand how the validity estimates produced by judges are affected by aspects of the task environment, particularly the information that is provided in the judgment task.

The research presented in this dissertation differs from the research conducted by Schmidt et al. (1983) and Hirsh et al. (1986) primarily because of its focus on identifying conditions that affect judgment output rather than on measuring the accuracy of expert or less-expert judges. Whereas these researchers demonstrated that judges can produce fairly accurate estimates of validities, research from other areas of psychology has suggested that in complex judgment situations, accuracy is an unexpected outcome. This outcome raises some doubts as to whether judges are actually "estimating" validities in the same sense as these researchers assumed. If judges are not actually "estimating" validities, then their output is not comparable in meaning to empirical criterion-related validity estimates and can not be used to measure the value of test scores.

Because judges are not capable of estimating criterion-related validities using inductive strategies that are isomorphic with the empirical validation process, they must be relying on judgmental strategies that simplify the cognitive demands of the validity judgment task. These strategies, whether they rely on estimates of test validities that are drawn directly from memory or the judgment task materials, or are additionally adjusted by using an anchoring and adjustment strategy, are highly dependent on features of the task that serve as cues for recalling, selecting, and modifying validity values. How the judgment task is structured--what information is provided, how judges are instructed to make judgments, what types of judgments are required--should, therefore, be an important contributor to the outcomes of rational estimation methodologies.

Research has demonstrated that cognitive limitations result in individuals processing only a small subset of the information available to them when making complex judgments (Tversky & Kahneman, 1974; Hogarth, 1980). The task environment can affect the nature of the subset of information used to formulate validity judgments, and thereby, the output of judges, by making information more or less available, directing attention towards certain types of information, affecting the value and/or meaning attributed to different types of information, or determining how information should be used (Slovic et al, 1977; Tversky & Kahneman, 1981; Slovic, Fischhoff, & Lichtenstein, 1977). If experts' validity judgments are (merely) evaluative, as discussed previously, then task-related information has a direct bearing on the validity values that are recalled or are identified in the judgment material. In addition, if information in the judgment task serves as an influential source of validity "anchors", then how the task is structured may much more directly determine judges' validity estimates.

If judges are relying on evaluative judgment strategies, then the most critical elements within the validity judgment task environment are the different types of information that are available to judges when they are making their judgments. This information can be in the form of general descriptions of organizations, candidate populations, more specific descriptions of predictors, criteria, or jobs, or even summaries of relevant validation studies and study results. The present research is an attempt to determine the effects on judges' output of two different types of information: job descriptive information (i.e., job title and relevant skills/abilities) and validity generalization results (for the type of job in question). These types of information are of research interest because of their relevance and importance in selecting and modifying validity "anchor" values. Briefly, particular job descriptive information provided in the judgment task is needed to identify appropriate validity values. Validity generalization results--if available--are an important source of validity "anchors". If judges do not use these types of information appropriately, then conclusions about the meaning of experts' validity judgments may be specious.

Clearly, there are many unanswered questions about how judges produce validity estimates, which task-related factors affect their validity judgments, and whether such a level of accuracy should be expected. Also, because expert accuracy is necessary but not a sufficient condition for advocating the use of rational judgments of validity, to interpret judges' output in a meaningful way it is essential to develop an understanding of expert judgment processes. This research attempts to build on the previous research efforts by determining if judges use information appropriately in the process of judging validities. This is investigated by comparing judgment output under different conditions of task-relevant information. In addition, because the previous researchers obtained judgments of validities using only a training criterion, this research extends their design by including multiple criteria in the validity judgments being made. Lastly, judges' confidence in their validity judgments was also studied to determine whether confidence is also affected by task conditions.

The Effects of Job Information on Validity Judgments

No matter what type of judgmental strategy is used to produce rational estimates of validities, the task of estimating validities should be highly dependent on information about the skills, abilities, knowledge, or personal characteristics required by the job for which validities are being judged. If expert judgments of validity are to have any meaning, it is necessary to demonstrate that judges are deriving validity estimates by processing information about the ability requirements of the job and that they are using this information appropriately. Such information should be used by judges to discriminate among the levels of validity that would be expected for different types of predictors. For example, if it has been empirically determined that a job requires a high degree of quantitative skills but a very minimal degree of psychomotor skills (e.g., a bookkeeper or inventory clerk), validation research would demonstrate that tests of quantitative ability have a higher validity than tests of psychomotor ability. In order for experts to accurately judge the magnitude of validity of these two tests, they must have access to and appropriately use information regarding the ability requirements of the job.

Based on the evidence about expert accuracy presented by Schmidt et al. (1983) and Hirsh et al. (1986), the average mean deviation of judges' validity estimates from the large sample validity, given only a general job description, was rather small. This result seems surprising considering the amount of specific job information that should be needed to make accurate point estimates of employment validities[11].

It may be, however, that judges use only very general job information--such as a job title or job family label--to recall validity results for a particular test-criterion combination or to determine whether specific values (remembered or provided in the judgment task) are appropriate validity estimates. Using only general job information may be an explicit strategy (see footnote #11),

---

[11] Psychologists working in the area of validity generalization have asserted that the only information needed to accurately apply validity generalization results is general job information that would identify the job family to which the new job belongs (Pearlman, 1980). Using general information may be a sufficient strategy when one is seeking to conclude only that the new job is similar to jobs that are represented in the validity distributions. However, because expert judgments of validity constitute point estimates of employment validities, one would expect that judges require very specific information to distinguish nuances of the job and the situation which are relevant to the level of validity expected in the new job. This assumption is predicated on the fact that where observed variance cannot be totally accounted for by statistical artifacts, situational and job factors are contributing to the variance in observed and true validities. One cannot know apriori whether that is in fact the case. Consequently, one needs to understand the job well enough to determine how these job and situational factors may manifest themselves in the correlations between tests and job performance measures to produce accurate estimates of validity.

or may be an implicit way of coping with limitations in information processing capabilities (see the discussion of evaluative judgment strategies). By focusing on only the job information which is perceived as being most salient when recalling or using validity values, judges limit the cognitive demands of the validity judgment task (Wallsten, 1980; Simon, 1956). However, for judgments of validity to be both accurate and meaningful, they should be based on a thorough understanding of the job which requires more than just knowledge of general job information such as the job title. To conclude that expert judges are capable of providing meaningful estimates of employment validities, it is important to demonstrate that they effectively attend to, process, and integrate the relevant job information that is provided in the judgment task. And, more importantly, that the way in which they use this information is appropriate given the value of the information in determining the level of validity expected for a particular test and job.

The Effects of Validity Generalization Information on Validity Judgments

It is possible that judges produce estimates of validities by selecting an appropriate value based on general job information (i.e., the job title) rather than estimating the interrelationships of factors which comprise the validity model. When judges are making validity judgments in this manner, job information acts as a cue for remembering or identifying appropriate validity values. It has also been suggested that judges may treat this initial value as a baseline estimate (the "anchor"), and adjust it when there is not an exact one-to-one correspondence between the cues and the recalled or identified validity. (These adjustments tend to be gross and are not expected to reflect the potentially minute differences between characteristics of recalled/identified values and the provided cues.)

Numerical validity values used as validity "anchors" are made available from a number of different sources (e.g., personal experience, published and unpublished research reports, professional training, accepted "conventional wisdom" etc.). Although expert judges are expected to have been exposed to a large amount of information concerning test-job validities, recalling all of the relevant information is likely to be problematic, if not impossible. It is more plausible that experts will rely on strategies that do not place such a high demand on cognitive processes; i.e., they will use only a subset of the total amount of relevant information available in memory and will recall and use only that information which is most salient. For expert judges, summaries of empirical validity information in the form of validity generalization results, because of their information value, should have a greater effect on validity judgments, when available, than comparable results from individual validity studies.

Whether validity generalization results are recalled from memory or provided as part of the judgment task, judges should consider these empirical summaries more valuable and, therefore, should give them more weight than individual study results when judging validities. There are several reasons why this is true. First, validity generalization results for a single predictor type-criterion type are more comprehensive because validity generalization studies consolidate outcomes from a large number of published and unpublished validity studies. Second, estimates of test-job validities that result from validity generalization analyses are more accurate than results from single studies. The reason for this is that sampling error variance can be removed from the distribution of observed validities so that conclusions can be drawn about the importance of job and situational factors in determining validation outcomes. In addition, validities are corrected for statistical factors that are artifactually attenuating them (e.g., criterion unreliability, range restriction, differences in factor structure, etc.). Third, if judges are required to estimate true validities, the validity estimates that result from validity generalization analyses are more comparable to the judgments they are making than are observed (frequently uncorrected) validity estimates that result from individual studies. Without the corresponding estimates of range restriction, reliability, etc., these observed validities can not be accurately or readily transformed by experts into estimates of true validity. Fourth, validity generalization results are generally presented in a more readily usable form than are results from many individual studies. Predictors, jobs, and criteria are categorized in validity generalization studies so there is no need to evaluate this information when using it to judge validities. Therefore, validity results from validity generali ation studies should provide relatively unambiguous information regarding the true levels of validities of different tests, criteria, and jobs.

Although validity generalization results should provide more valuable information than that which results from individual validity studies, this is the case only if these results are relevant for the job for which validities are being judged. To be more specific, validity generalization results are appropriate estimates of test-criterion validities only if the job in question requires a similar profile of skills and abilities as the jobs that are represented in the validity distributions used in the validity generalization analyses. It is possible for validity generalization results to be misapplied and, therefore, for test-criterion combinations to be assigned inappropriate validity values. It would be important to determine, therefore, whether validity generalization information is used effectively when judges are estimating validity.

Theoretical Rationale and Hypotheses

To predict that expert judges' estimates of validity will depend both on general job information (rather than on specific predictor-job information) and on validity generalization results requires two different types of conceptual support. First, it is necessary to explain why some information (i.e., job titles versus particular skill and ability requirements; provided validity generalization information versus individual validity study results recalled from memory) will be more salient than other information when validity judgments are being made. Second, one must describe why judges will provide validity estimates that do not deviate substantially from these available validity generalization values.

Research from such areas as information processing heuristics and biases (cf., Tversky & Kahneman, 1974), statistical inferences (cf., Nisbett & Ross, 1980), and clinical judgment and decision-making (cf., Fischhoff, 1975), where applicable, will be reviewed to provide a justification for the hypotheses that will be tested. Although most of the research in these areas has centered on judgments of risk and probability, some of the findings should be generalizable to other types of complex judgments, namely, rational judgments of criterion-related validities.

Saliency of General Job Information and Validity Generalization Information

Information is considered "salient" when it has a disproportionate effect on cognitive processes relative to other information (Crocker, 1982). The effect of informational saliency has been evidenced in attention and perception (Neisser, 1967), memory, behavior (Fishbein & Ajzen, 1975) and judgment (Anderson, 1974). If multiple types of job information are presented in the judgment task, it is very probable that certain types of job information will be more salient and, therefore, will have a larger affect on validity judgments. In addition, if judges have access to a variety of sources of validity information, it is likely that the most salient information will be that which is considered more valuable and/or is easier to evaluate and use vis-a-vis the particular judgments being made.

Salient job or validity information may be expected to affect judgments of validity in two ways. First, salient information has a cuing function; it has a relatively larger effect on judgment task outcomes than less salient information because it has a relatively larger role in determining which information judges have access to when judging validities. Validity information may be organized and stored in memory in such a way that job information in one form (e.g., job title) serves as a better cue for recalling stored validity study outcomes than when it is in another form (e.g., ability requirements). In addition, general job information may be more useful in recalling more general information such as general beliefs, assumptions, and implicit theories about employment validities and, therefore, may play a more important role in determining the availability of this type of information when judging validities.

Second, the saliency of information may determine what information is directly used in producing validity judgments. More specifically, once relevant validity information is made available through memory processes and/or as part of the judgment task, the information must be integrated to produce an estimate of test-criterion validity. Information which is more salient is expected to be weighed more highly when judgments are being produced (cf., Ebbesen & Kenecni, 1975; Christensen-Szalanski, et al., 1983). This explains, for example, why validity generalization results would be expected to have a greater affect on judgments of validity than results from individual studies: in comparison to other sources of validity information, validity generalization results (provided or remembered) are expected to be more salient and, therefore, to be the more influential source of validity values.

Likewise, when a job title and specific job information are both available, the job title is expected to be more salient and to play a more influential role in the selection and use of validity values including validity generalization results.

Determinants of Information Saliency

Abstract vs. Concrete Information. Research that has explored how individuals use base rate information suggests that information may lack impact when it lacks realism (Nisbett & Ross, 1980) or when it is not presented exactly in the form in which it must be used (Slovic & Lichtenstein, 1968; Payne & Braunstein, 1971). "Concrete" information has been shown to have more impact than "abstract" information in a variety of judgment situations (Bruner, Goodnow & Austin, 1976; Slovic, 1972b; Olson, 1976; Borgida & Nisbett, 1977), suggesting that information that has to be analyzed, inferred from the explicit display, or transformed in some way tends to be discounted or is less salient.

"Abstract" information is operationally defined as information that lacks realism or does not directly correspond to an actual instance or specific object (e.g., statistical results such as event probabilities are considered more "abstract" than personal knowledge of an event occurrence). When information is abstract it is not perceived as being as directly relevant to the judgment at hand as is more concrete information. Therefore, its relevancy must be evaluated before it is used. If evaluating the usefulness of abstract information increases the information processing demands of the judgment task, this information is likely to be ignored relative to less abstract information (Simon, 1955).

"Concrete" information, because it is perceived by a judge as being isomorphic with an actual object or event, is much more vivid than abstract information. Within the context of a judgment task, concrete information may be perceived as being more relevant or valid than comparable abstract information, possibly because it is easier to attribute meaning to information that is "real" (Borgida & Nisbett, 1977), or because the meaning of the information is readily apparent (Slovic & Lichtenstein, 1968). As a result, concrete information is more directly available during the judgment process than is abstract information[12].

If two types of job information are made available to judges when estimating validities--the job title versus more specific information concerning the ability requirements of the job--the job title may be more likely to produce a characterized as being more concrete. The reason for this is that experts probably have mental representations of jobs

[12] Borgida & Nisbett (1977) have suggested that the difference in saliency of concrete vs. abstract information may explain why trained researchers are subject to the belief in the "law of small numbers" (Tversky & Kahneman, 1971). Researchers may be willing to readily generate inferences from small samples of data because the data are perceived as being more relevant to the judgment than is information regarding sampling theory.

in terms of the actual activities surrounding the work rather than the underlying skills, abilities, and personal characteristics associated with successful job performance. Therefore, when judges are processing information relevant to making validity judgments for a particular job, a job title may more easily evoke cognitive images of the job and how it is performed than a list of more abstract abilities and skills. It is likely, for example, that Industrial and Organizational Psychologists think in terms of the whole job rather than in terms of job components when processing relevant information about a job (i.e., "How well do tests of perceptual abilities or psychomotor skills work in clerical jobs?" versus "How well do tests of perceptual abilities or psychomotor skills work in jobs which require verbal communication, simple arithmetic and social skills?") and evaluate relevant information based on a more global understanding of the job.

Second, general job information, such as a job title, may also be more concrete than specific job information because relevant validity information is stored in memory based on more global characteristics of jobs. Therefore, when experts are thinking about validity information their memory for this information is more easily evoked by the more realistic image of the job. Such a strategy is likely because validity results (single study or validity generalization) are generally presented by job categories rather than for sets of different skills and abilities[13]. Recalling information based on cues comparable to those used to store the information also simplifies the information processing demands of the task (Neisser, 1967).

Whether judges are presented with validity generalization results in addition to other task-relevant information, or are able to recall these results on their own, this information is expected to have a large affect on judgment output. The reason for this is that validity generalization results will be perceived as being more concrete than other types or sources of validity information. There are several reasons for this. First, when validity generalization results are explicitly provided in the judgment task, this information is more concrete--by definition--than the validity information judges will be able to recall from memory[14]. Although judges will probably be able to recall the approximate level of some predictor-criterion correlations, relevant validity generalization results that are both available and in a readily usable form

[13] Pearlman (1980) has suggested that job families for validity generalization should be formed on the basis of underlying skill and ability requirements. However, this strategy is generally not used. Instead, validity generalization results are reported and transported across jobs on the basis of general job family/job title information.

[14] The only time that this might not be true is for those validity studies with which judges have had personal experience. In that case, the validity values from that source will be more concrete. However, except for a very few experts, it is unlikely that judges will have had personal experiences with validity studies similar to both types of jobs and with each of the 28 predictor-criterion combinations used in this judgment task.

are expected to be more concrete and, therefore, more salient than recalled validity values from individual studies (Slovic & Lichtenstein, 1968). This would even hold true if the remembered validity values are from validity generalization studies. Second, even if validity generalization results are not explicitly provided in the judgment task, the way in which this information is organized (i.e., general predictor-criterion categories, estimates of true validity, etc.) will make recalled validity generalization outcomes more concrete and, therefore, more salient, than recollections of individual study results (Slovic & Lichtenstein, 1968).

Effects of Outcome Knowledge. In a variety of judgment situations, individuals are affected by outcome knowledge even though they are largely unaware that this information has changed their perception (Fischhoff, 1975). Research concerning probability judgments, in particular, has found that reporting that an outcome has occurred (or the frequency of outcome occurrence) consistently increases its perceived likelihood. In addition, outcome knowledge alters the judged relevance of data describing the situation preceding the event (Fischhoff & Byeth, 1975).

In some judgment situations judges are well justified in using outcome knowledge to guide their judgments. This is certainly true if such information is based on results from a large sample of representative data[15]. However,

[15] For example, knowing the frequency of occurrence of a disease in the total population of individuals exposed to

because outcome knowledge is not always relevant, accurate, reliable, or appropriate to the particular judgment being made, it can negatively affect the accuracy of judges' output. If this bias is as strong as was suggested by Fischhoff (1975) and Fischhoff & Byeth (1975), then in many types of judgment tasks judges may be inappropriately influenced by information which they believe is relevant.

If this information-processing bias can be generalized from judgments of probability to other types of judgment tasks, then it can potentially explain why validity generalization results should have a greater effect on validity judgments than results from individual validity studies recalled from memory. More specifically, if validity generalization results are perceived to be more accurate estimates of validity research outcomes than individual studies, then judges may be more strongly influenced by this information when estimating test-criterion validities.

In some cases, weighing validity generalization results more highly is an effective judgment strategy. This would be true if validity generalization results are representative of the same types of jobs, tests, and criteria as the validities being judged. However, where validity generalization results should not be generalized

---

the disease is relevant and appropriate information on which to base one's probability estimate of the likelihood of contracting the disease subsequent to exposure. On the other hand, knowledge that one or ten individuals who are similar to you have contracted the disease should not affect one's probability estimate.

(i.e., where they represent jobs requiring a different profile of abilities than those required in the job for which validities are being judged), the research in this area suggests that judges will be inappropriately influenced by validity generalization outcomes.

Anchoring and Adjustment

Even if the job title or the validity generalization information is salient and, therefore, is an important source of information when validities are being judged, it is necessary to understand why other information concerning test-job validities (e.g., validity values recalled from memory, general beliefs about the magnitude of validities for different predictors, criteria, or jobs, etc.) may not have a large impact on validity estimates. Without such an understanding, it would be impossible to predict that judges' validity estimates will largely depend on the job title and on the validity generalization results.

As previously described, it is likely that judges are recalling or using validity values that they have identified as being relevant based on a limited set of salient cues present in the judgment task. A natural validity value or "anchor" for validity judgments is the most similar validity generalization value that is explicitly provided or recalled from memory. If such information is unavailable, then it is likely that judges will use as an "anchor" the most relevant individual validity study value that can be recalled from memory.

The "anchor" value may then be grossly adjusted if the judge determines that the characteristics of the validity being judged and the validity "source" do not match perfectly[16]. If these adjustments are inadequate, the final validity judgment will be very similar to the initial starting value and will not adequately take into account the relevant characteristics associated with the validity being judged. The question then is: how likely is it that judges would make accurate adjustments to the initial validity value that they identified?

Research studies of anchoring and adjustment processes have suggested that ". . . the adjustment is a crude and imprecise one which fails to do justice to the importance of additional information" (Slovic, 1972b, p.16). More specifically, the adjustment to the initial value is typically insufficient; individuals fail to take into account the full information value of additional information (Alpert & Raiffa, 1968; Lichtenstein & Slovic, 1971) or place too much value on the initial starting value (Tversky

[16] Validity values from validity generalization analyses are the best estimates of true validities in the new job only if jobs, situations, predictors, and criteria are randomly sampled in the validity generalization data base or if there is an exact match between the characteristics of the validity generalization sample and the characteristics of the job for which validities are being judged. If not, and these validity generalization values are used as anchors, they need to be altered to take into account these job, situational, test, and job performance measurement differences.

& Kahneman, 1972). For example, when individuals were given random starting values in a task involving the judgment of almanac values (i.e., the percentage of individuals in the United States under the age 55), those whose starting points were too high ended up with higher estimates than those whose starting points were too low (Tversky & Kahneman, 1972). Although it can be safely assumed that judges will not be using random values as starting points, this research does show that even given obviously erroneous "anchors", judges fail to make adequate adjustments to reflect the relevant and more valuable information to which they had access. As was previously mentioned (see footnote #16), true validity estimates from validity generalization analyses are "just right" only if they are very characteristic of the job, situation, tests, and criteria comprising the validities being judged.

The result of the anchoring and adjustment bias described above is that final judgments of criterion-related validities are expected to be strongly related to judges' starting values because the adjustments they may make are expected to be too conservative. These results are congruent with biases found in other decision-making situations (cf., "conservatism" [Edwards, 1968] and the "method of successive limited comparisons" [Lindblom, 1964]) and difficulties that individuals have in integrating information in other types of judgment tasks (cf., Roose &

Doherty, 1976; Kaplan & Kemmerick, 1974; Slovic & MacPhillamy, 1974; Phelps & Shanteau, 1975).

Taken together, these studies have demonstrated that individuals do not optimally combine information in the process of making judgments. For example, they tend to weigh information inappropriately (Roose & Doherty, 1976), they use information which is clearly irrelevant (Kaplan & Kemmerick, 1974), their strategies for combining information contradict what should optimally be done (Slovic & MacPhillamy, 1974), and they are ineffective at extracting cues from real stimuli and using multiple cues to make judgments (Phelps & Shanteau, 1978).

These research findings suggest that when relevant validity generalization results are available they will be used as the initial estimate of test-criterion validity for a particular job and that additional information (i.e., general beliefs about magnitudes of test validities in certain occupational groups, relevant differences between the judged and "anchor" validity, etc.) may be used to adjust these "anchor" values. However, the adjustments will be of small magnitude. Therefore, even when expert judges are provided with identical job information for two jobs, if validity generalization results are provided in one case but not in the other, or if the validity generalization results differ, then validity judgments made by expert judges will differ. Analogously, judges are likely to produce similar validity estimates for jobs which differ on important job dimensions if the judges are provided with identical, apparently relevant validity generalization results.

The preceding discussions have described the types of judgment processes that judges are likely to be using, which information in the judgment task will be most salient, and why anchoring and adjustment biases will result in judgments that do not accurately reflect information which is critical to judgment accuracy. Taken together, these shortcomings in judges' information processing strategies and skills have suggested what the effects of different types of job information and validity generalization results will be on judgments of validity. In consideration of these discussions, the following hypotheses are offered:

HYPOTHESIS #1: Judgments of criterion-related validity are significantly different for jobs with the same ability/skill requirements but having different job titles.

HYPOTHESIS #2: Judgments of criterion-related validity are significantly different for jobs with the same job title when validity generalization results are provided than when they are not provided.

Effects of Job Information and Validity Information on Judges' Confidence in Their Validity Judgments

While it is important to consider how the structure of the judgment task, primarily task information, affects judgments of validity, the task environment can also affect how judges' perceive their judgment output. In particular, it would be important to determine whether the availability of validity generalization results or differences in job titles or job information affect the confidence that judges have in their output.

Individuals are unable to make many different types of complex judgments optimally. However, they tend to be very confident about the appropriateness of their output (Goldberg, 1959; Oskamp, 1965; Kahneman & Tversky, 1973). Alpert and Raiffa (1968) and Tversky & Kahneman (1972) both found that when judges were asked to provide the lower and upper bounds of a 98% confidence interval around their point estimates, the interval failed to include the true value from 40% to 50% of the time. This result persisted even when feedback was given to judges about their overconfidence and after the judges were explicitly encouraged to widen their confidence intervals. The researchers concluded that given the judges' state of knowledge (about almanac-type information), these narrow confidence bands were indicators of gross overconfidence in judgment accuracy.

Certain characteristics of judgment tasks affect the level of confidence that judges have in their output. For example, despite the fact that judges are generally not able to use all of the information they are provided with, the presence of additional information appears to strengthen judges' confidence in their estimates (Oskamp, 1965). In addition, when different types of information provided in the judgment task are inconsistent or variable, judges have less confidence in their judgments (Kahneman & Tversky, 1973).

Researchers have found that in predictive judgments, individuals become overconfident when task-relevant information from two different sources is either consistent or extreme (Kahneman & Tversky, 1973; Fischhoff, Slovic, & Lichtenstein, 1977; Einhorn & Hogarth, 1978). However, under both of these judgment conditions judges' predictions tend to be less accurate. (Multiple correlation with a criterion is inversely related to the correlations among the inputs.) This "illusion of validity" suggests how easily the structure of the judgment task can lead judges to feel that complex judgments can be made with little error.

According to research by Oskamp (1965), the availability of additional, apparently relevant information -- for example, providing versus not providing validity generalization information--should be associated with a greater belief in one's judgment ability even if the additional information is actually irrelevant or incorrect. Oskamp's research findings suggest the following hypothesis:

> HYPOTHESIS #3: Confidence in judgments of validity is significantly greater when validity generalization information is provided than when it is not provided.

In addition, according to Kahneman & Tversky (1973), when the two types of job information provided in the judgment task are consistent, judges should be more confident in their validity judgments. With respect to the task of judging validities, job information is consistent (or congruent) when the actual skill and ability ratings associated with a job are those that are connoted by the job title (i.e., the job stereotype). In reality, judges should be no more confident in their output when such job information is consistent than when it is inconsistent because in this complex judgment task, mere job title has little information value and is independent of skill and ability ratings as sources of judgment information.

> HYPOTHESIS #4: Confidence in judgments of validity is significantly greater when the job title and skill/ability information are consistent than when this information is not consistent.

CHAPTER IV

Research Design and Methodology

Research Design

A 2x2x2 experimental design was employed to study judges' use of both job and validity generalization information in a task involving the judgment of test validities. The experimental design involved the following three factors: type of job (clerical versus sales), type of skill/ability information (congruent versus incongruent with job title/general job information), and availability of relevant validity generalization results (provided vs. not provided). The first factor was a within-subjects factor and the second and third factors were between-subjects factors. Judges were randomly assigned to one of the four experimental conditions which resulted from crossing the two levels of each of the job and validity generalization factors. So that potential, but unanticipated, order effects could be controlled, the order in which judges were required to judge validities for different predictors and jobs was counter-balanced within each of the four experimental conditions.

Experimental Manipulation of Job and Validity Generalization Information

Job information. Job information was manipulated by providing judges with skill/ability descriptions that were either congruent or incongruent with the job title of the job for which validities were being judged. In the "congruent" condition, the skill, ability, and knowledge areas that were given high ratings (i.e., were important) were similar to the human requirements that would generally be expected to be important in a job with that job title (i.e., that are congruent with the job stereotype). For example, in a job with a clerical job title, individuals would expect memory, verbal comprehension, and perceptual speed and accuracy to be moderately or very important, but reasoning, social skills, and spatial abilities to be less important. (These results have been observed in numerous job analysis studies of jobs in the clerical family.) Therefore, in the "congruent" job information condition, the information provided about skill/ability requirements was redundant with the requirements evoked by the job title alone.

In the "incongruent" job information condition, the judges were provided with skill, ability and knowledge requirements that differed from the requirements presumably evoked by the job title. For example, in the "incongruent" condition, for a job with a clerical job title, the judge was provided with skill and ability ratings that suggest that the job requires low to moderate memory, high verbal comprehension, high numerical facility, high social skills, low perceptual speed and accuracy, moderate reasoning ability, and low spatial ability, etc. These ratings are incongruent with the importance of these skills/abilities that would be connoted by the job title/general job description.

It should be noted that in the "incongruent" job information condition (i.e., when skill/ability ratings do not correspond to those evoked by the job title/general job information), the job information actually corresponds to the stereotype associated with the alternative job. Therefore, when the skill/ability information was incongruent with the sales job title, the job actually being described was a typical clerical job. When the skill/ability information was incongruent with the clerical job title, the job actually being described was a typical, (non-management) sales job. This manipulation suggests that if the job title does not play a significant role when judges are making their estimates, but the SKAP information does, then validity judgments should be comparable for the clerical job in the "congruent" condition and the sales job in the "incongruent" condition. Likewise, validity judgments should be comparable for the sales job in the "congruent" condition and the clerical job in the "incongruent" condition. If validity judgments from these corresponding conditions are significantly different (and, correspondingly, validity judgments are not significantly different when SKAPs differ but job titles are the same), then it can be presumed that job title information has had a significant influence on judges. This result would strongly suggest that judges were not using the task-related information optimally; i.e., they were unduly influenced by job titles. This conclusion was examined by testing hypotheses concerning significant differences in validity judgments where no such differences should have existed (i.e., where skill/ability information was identical but job titles differed).

This manipulation is considered to be highly relevant for two reasons. First, job titles adopted by organizations are frequently ambiguous and/or not generalizable to other organizations where similar job titles are used. Second, specific job information is clearly needed in any research involving employment selection, especially if criterion-related validities are being estimated. Unless job family/job title information is empirically derived, job titles and a general description of the job should, in actual practice, provide judges with little information on which to base judgments of validity. The relationship between test and job performance is based on the underlying skill, ability and knowledge requirements of the job, and should not be inferred solely from the label attached to the job by the organization. Therefore, judges should seek out and use job information that has the most value for the task at hand - information corresponding to the human requirements for work performance. This manipulation allowed us to discern whether judges attended to the more important information needed for judging validities.

Validity generalization information. Validity generalization information was manipulated so that validity judgments were made in one of two different conditions of information:

(1) **Relevant validity generalization results were provided** to judges as part of the judgment material. When provided, the validity generalization information was "matched" to the job title rather than to the SKAP information. For the clerical job title, validity generalization results from a study conducted by Nathan and Alexander (no date) using Pearlman, Schmidt, and Hunter's (1980) clerical data base was presented. For the sales job, results were synthesized from 76 validity studies involving sales jobs which were published in Personnel Psychology from 1918, or which appeared in the references cited by Churchill, Ford, Hartley, and Walker (1985)[17].

[17] When validity data were not available for a specific predictor-criterion combination, a fictitious, but plausible, estimate was used. These estimates were derived by maintaining the rank-order of validity magnitudes from the validity generalization results for clerical jobs. However, the magnitude of validities was made comparable to

(2) No validity generalization information was provided.

Support for this type of manipulation is based on the premise that the validity generalization results provided are actual reports of study outcomes. The methods and conclusions have been professionally scrutinized and accepted (i.e., the results were published in refereed journals or reviewed by dissertation committee members). In addition, many studies examining the accuracy of validity generalization procedures (focusing on clerical jobs, in particular) have been published and, therefore, have been circulated throughout the field of Industrial and Organizational Psychology.

The Judgment Task

The validity judgment task consisted of two different types of judgments (the dependent variables):

(1) **a judgment of the true validity for each of the 28 test-criterion combinations for each of the two jobs** (see Appendix C for a copy of the rating form). The judgment of validity for each predictor-criterion pair was made by indicating the validity value on a rating scale anchored with -.30 and .80 with intervals designated every .10 units; and

---

the sales validities. These fictitious estimates were computed for 64% (18 out of 28) of the validities represented in the matrix of sales validity generalization results.

(2) **a confidence judgment** was made for every validity judgment. This judgment consisted of defining an interval around each validity judgment for which the judge is 90% confident that the interval contains the true validity. The confidence judgment was made by placing brackets ([ ]) on the rating scale around the estimated validity. These brackets corresponded to P-values of .10 and .90.

Issues Related to Using Between-Subjects Research Designs

Between-subjects designs are generally not used in research on human judgments because individual differences in judgment processes (i.e., perceiving, categorizing, weighing and integrating informational cues) are usually the main focus of the research. Einhorn (1972) warns, for example, that grouped data may obscure trends that are occurring in individual data. These designs generally employ a limited number of judges (e.g., three to 20) who make many judgments from a preselected set of multidimensional stimuli. The stimuli are frequently artificially created to represent every combination of the cues under consideration (cf., Phelps & Shanteau, 1978). Analyses of these judgments are done separately for each judge, thereby providing information needed to investigate differences between judges as well as other factors of interest.

Although differences in the processes by which individuals make judgments of validity is an important area of research, this particular judgment task does not lend itself to examining effects of individual judges. On the practical side, it is not possible to have professionally trained psychologists judge validities several times for the same job under different conditions of job information and/or validity generalization information. These types of manipulations would be easily detected by such experts. Although the manipulations might go undetected if judgments were made for a large enough number of jobs, this would place unreasonable demands on the potential judges.

The more important justification for the use of a between-subjects research design is based on the theory underlying the rational estimation methodology and the desire to generalize results to samples of judges rather than individual decision makers. The strategy which has been advanced for obtaining estimates of validities is to pool judgments from samples of judges and to use the mean judgment as the estimate of validity (Schmidt et al., 1983). This strategy, therefore, assumes that judgment error associated with any individual judge (i.e., the judge's deviation from the mean judgment of the sample) is random error. Although random error can reduce the accuracy with which judges estimate validity, random error itself can be reduced by increasing the number of judges. Given a large enough sample of judges, the random error begins to approach zero and the mean estimate of validity approaches the population estimate. What primarily determines the accuracy of any group of judges, therefore, is the systematic error in their judgments (i.e., the deviation of the sample mean estimate from the true validity).

Because the validity judgment methodology conceptualizes the deviation of a judge's estimate from the sample mean as random error, to examine the practical utility of this process one can consider individual differences in judgments as having little information value. The focus, rather, is on systematic differences in judgments that occur under different conditions of measurement. The trends in the data that are of interest are not individual trends but, rather, group trends. The concern in this research effort, for example, is with identifying whether the presence of particular types of information affects validity judgments and, therefore, the value of this methodology.

The consequences of not being able to examine individual differences can potentially be significant. Because individual differences become a part of the (within cell) error term, main effects and, more likely, interactions, may be obscured. This would happen, ostensibly, if individual differences in judgment processes have a relatively large effect on validity judgments in comparison to the experimental factors within the judgment situation.

To avoid this situation, several strategies have been incorporated into the research design. First, the sample of judges that was used to assess the effect of job and validity generalization information was fairly homogeneous with respect to education and experience. This operationalization of "expert" judges helped to minimize the effects of individual variability on the grouped data. Second, care was taken when selecting the jobs for which validities would be judged, so that there would be real differences between jobs in the magnitude of validities of individual predictors. Based on a review of the validation research literature, it was concluded that the magnitudes of test-criterion validities for clerical and sales jobs differ. When judges were not provided with relevant validity generalization information, they were expected to rely on the job stereotype connoted by the job title/general job information when estimating validities. Again, these stereotypes would lead to judgments of predictor-criterion correlations that differ across jobs. Third, based on the results reported by Schmidt et al. (1983) and Hirsh et al. (1986), random error in judgments was appreciably less than systematic error in judges' estimates (.0208 vs. .0481 and .0332 vs. .0437, respectively), suggesting that judges are fairly homogeneous when making validity estimates under the same measurement conditions.

Sample

The sample of judges used in this research consisted of 114 professionals from the field of Industrial and Organizational Psychology who voluntarily responded to the judgment survey. Participants in the study were all members of Division 14, the Society for Industrial and Organizational Psychology, Inc. (SIOP), of the American Psychological Association (APA) during the 1986-1987 membership year. A randomly generated list representing three membership categories - Fellows (n=256), Members (n=336), and Associates (n=160) - was used to made random assignments of individuals to each of the four experimental conditions. These three membership categories reflect differences in levels of professional training and/or recognition within the field: Associates have graduate training in Psychology but have not attained the Ph.D., Members must have earned the Ph.D. in a field primarily psychological in nature, Fellowship status is elected and is based on contributions and service to the field.

To become a member of Division 14, in addition to being a member of APA, an individual must have received graduate training in psychology for at least two years ("Member" status requires the Ph.D in Industrial and Organizational Psychology or an equivalent field) and "must be engaged in professional activities, as demonstrated by research, teaching, and/or practice, related to the purpose of the

Society as stated in Article 1, Section 2 of the Bylaws. Such activities may be performed in a variety of settings, such as private business or industry, educational institution, consulting firm, government agency, public service foundation, or self. There must be at least one year of full-time service in these activities. (The Industrial-Organizational Psychologist, 1987)."

Of the 750 judgment surveys that were distributed, 139 (18.5%) responses were received and 14 (1.9%) were returned undelivered. Usable data were available from 114 individuals (15.2% of the total sample). (Only one of these 114 subjects did not provide a complete set of judgments.)

Sampling Issues

It is frequently the case in research on judgment processes that a distinction must be made between novices and experts. Judges familiar with the substance of the judgment task and trained in making similar types of judgments use and interpret information differently and arrive at different judgments than individuals who are novices (cf., deGroot, 1966; Johnson et al., 1982). The population of experts is usually identified by objectively measurable characteristics such as amount and type of training, licensing, professional recognition, etc. (cf., Johnson, et al., 1982; Hirsh et al., 1986; Phelps & Shanteau, 1978).

One could easily argue for a more rigorous identification of experts that focuses on skills and abilities that are related to effectiveness in the type of judgment task being studied. However, this strategy depends on researchers understanding and identifying the processes used by accurate judges. For example, Einhorn (1974) has suggested that the criteria for evaluating expertise should include accuracy in extracting information from the environment, accuracy and consistency in categorizing and measuring those stimuli, and the ability to use information according to its value. In more complex judgment tasks such as this one, where little is understood about how judges obtain and use information, Einhorn's guidelines would be difficult to implement.

The rationale for defining the population of judges according to the training and work experience criteria set forth for Division 14 membership is twofold. First, this research effort is concerned with being able to generalize results to the population of judges with whom actual applications of this validity estimation methodology would presumably be implemented. The strategy of defining experts as Division 14 members is similar to the strategy that would probably be employed by organizations interested in identifying judges. Second, it is necessary to establish that judges have at least a minimum level of task-relevant skills. Clearly, the value of rational estimation methodologies must be ascertained with respect to "expert" judges[18]. Results from samples of judges that are too heterogeneous with respect to skill-level (as a function of education, training or experience) will provide little information about the practical utility of using judges to estimate validities, although research using judges with less expertise may provide clues about the various processes used by judges. At the same time, results from a hand-picked sample of highly skilled experts (cf., Schmidt et al., 1983) will be inappropriate to generalize to the types of samples that would be likely to be employed in this task in real-world applications.

Procedures

Data Collection and Administration

All data collection took place by mail and was confidential and anonymous. Instructions, judgment information (i.e., predictor, criterion, job and validity generalization information), and survey forms were mailed directly to 750 qualified judges. In addition to the

[18] For example, Hamilton and Dickinson (telephone conversation with J.Hamilton, 2/87) used a sample of "experts" that consisted of one Ph.D. psychologist and four doctoral-level psychology graduate students to estimate the correlations between job elements and tests. Each of the methodologies employed to estimate the J-coefficient using these "expert" ratings was unable to accurately estimate test validities. Although the authors concluded that "experts" must have extensive experience in personnel selection to make such estimates accurately, clearly this is an empirical question which could have been more carefully addressed in their research.

judgment task materials, each judge also received an addressed and stamped envelope for returning the completed survey and an addresses, stamped postcard with which they could request a copy of the research results and/or indicate their willingness to participate in a follow-up study. All judges who returned surveys by June 30, 1987 were included in the research sample. (All surveys were mailed on April 27, 1987.)

Instructional set. Judges were provided with a general introduction to the judgment task and more specific, detailed instructions regarding how to complete the judgment process. (Appendix A contains a copy of the instructions that were provided to the judges.)

A separate sheet containing a brief review of the effects of criterion unreliability, restriction of range on the predictor, and sampling error on true validity was included along with the set of instructions. The purpose of this review was to provide all judges with the same minimal information needed to make their judgments and transform observed validities (where necessary) into true validities.

All judgment materials (i.e., predictor and criterion background information, SKAP information forms etc.) were briefly described to the judges and the rating form reviewed. Judges were asked to first browse through the general information corresponding to the judgment task (i.e., the introduction, instructions, review of observed versus true validity, predictor descriptions, criterion descriptions and the rating sheet) to familiarize themselves with the judgment materials that were referred to in the instructions. To ensure that order effects could be assessed properly and that experimental manipulations of job information were effective, judges were asked to review the specific information corresponding to each job (i.e., the job information and/or the validity generalization information) only at the time that they are making the judgments for that job.

Judges were not advised as to the amount of time they should use to make their judgments, although they were informed, in the cover letter, that their participation in the study would require less than one hour. The reason for allowing judges to use whatever time they needed to make their judgments was discussed by Einhorn et al. (1979). These researchers suggested that as time pressure increases, judges are forced to rely less on judgment and more on other cognitive "short-cuts" (e.g., non-compensatory choice strategies). Therefore, to ensure that judges would have sufficient time for deliberating over their judgments, no time limit was suggested.

Types of judgment information provided. In addition to the cover letter and instructions, judges were provided with the following types of information needed to complete the judgment task: (Copies of this information appear in Appendices B and C and a more detailed description of the general job description and validity generalization information appears in the review of the experimental manipulations.)

(1) **Predictor background information.** This information included a general description of each of the seven cognitive abilities for which validities were being judged and specific examples of the types of tests used to measure each ability;

(2) **Criterion background information.** This information included a general description of each of the four job performance measures and, where necessary, a more specific example of the criterion;

(3) **Job title.** This is a job title as might appear in an organizational job classification system;

(4) **General Job Description.** This general description of the job for which validities were estimated included a set of statements that described general job activities and job context. These descriptions were intentionally written to be ambiguous and interchangeable between the clerical and sales jobs;

(5) **Skill and Ability Requirements.** This information consisted of a list of skill and ability importance ratings such as would result from a job analysis; and

(6) **Validity generalization information.** Validity generalization information was provided in the form of a list of estimated mean true validities for each predictor-criterion pair for which validities were being judged.

Data Analyses

Data analyses were conducted to address the following research questions:

(1) What is the effect of congruent versus incongruent job information on level of judged validities across the 28 predictor-criterion combinations in the two different jobs? for each of the 28 predictor-criterion combinations in the two different jobs? (Hypothesis 1)

(2) What is the effect of the presence versus absence of validity generalization information on level of judged validity across the 28 predictor-criterion combinations in the two different jobs? for each of the 28 predictor-criterion combinations in the two different jobs? (Hypothesis 2)

(3) What is the effect of the presence versus absence of validity generalization information on level of confidence in judged validities across the 28 predictor-criterion combinations in the two different jobs? (Hypothesis 3)

(4) What is the effect of congruent versus incongruent job information on level of confidence in judged validity for each of the 28 predictor-criterion combinations in the two different jobs? (Hypothesis 4)

Effects of Job Information and Validity Generalization Information on Validity Judgments and Confidence Judgments. Several different types of statistical analyses were conducted to obtain data relevant to Hypotheses 1 - 4. First, descriptive statistics were computed to provide an initial description of judges' responses. Then, multivariate analysis of variance procedures were used to measure the effects of job information and validity generalization information on experts' validity judgments and confidence judgments. This analytical procedure was used because it allows one to measure the effects of multiple experimental factors on multiple response measures. To aid in understanding results obtained from the multivariate analysis of variance procedures, univariate analysis of variance procedures were used to assess the effects of the experimental conditions on individual validity and confidence judgments. Lastly, "exploratory" multivariate analysis of variance procedures were conducted to examine how predictor and criterion "factors" (types of tests and performance measures for which validities were judged) were related to judges' validity judgments and confidence judgments. These analyses also measured the interactions of these within-subject factors with the experimental conditions being studied.

Because multivariate analysis of variance was the main data analytic procedure employed in this research effort, is subsequently discussed in greater detail later in this section.

Analysis of Sample Demographics and of Order Effects. Additional data analyses were conducted to determine if order of presentation affected validity judgments and to describe the characteristics of the research sample. The former utilized multivariate analysis of variance procedures similar to those described above, the latter relied only on simple descriptive statistics.

Multivariate Analysis of Variance

Multivariate analysis of variance (MANOVA) procedures simultaneously analyze the effects of multiple experimental factors on sets of outcome variables. These procedures are thus an extension of univariate analysis of variance (ANOVA). In the case of MANOVA, the null hypothesis being tested is that population centroids from two or more groups are equal. These centroids are vectors of means of the response measures being studied.

The Statistical Analysis System (SAS) software, which was used to analyze data in this study, offers several different MANOVA tests including Wilks' Criterion, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Maximum Root Criterion (SAS, 1985). Although each of these is based on different computations of and approximations to the F-ratio, the results are identical because the experimental factors have only two levels.

Whenever the overall, multivariate null hypothesis of "no population differences" is rejected, post hoc comparisons between treatment groups can be made. These comparisons can be based on the original responses or the linear combination of the dependent measures that was used in the MANOVA analyses. In either case, univariate ANOVAs are computed. These analyses help to identify those outcome measures which have contributed to the rejection of the null hypothesis. The primary advantage of conducting ANOVAs on the original responses is that results are generally more easily interpreted. This is especially important if one is concerned with maintaining the same scale as the original responses.

MANOVA procedures are especially useful when subjects' responses constitute repeated measures on one or more factors. Repeated-measures ANOVA designs are based on fairly restrictive assumptions regarding the intercorrelations among subjects' responses across levels of the within-subject factor. Specifically, tests of within-subjects effects in ANOVA designs are only exact when the variance of subjects' responses to p-outcome measures are homogenous across different treatment levels. Therefore, one must either verify that this "homogeneity of treatment-difference variances" assumption is upheld, or use conservative tests of significance (Harris, 1975, p. 127).

MANOVA, on the other hand, does not require adherence to such an assumption. MANOVA can handle repeated measures by treating subjects' responses to different levels of the same factor or factors as a single outcome vector.

In any ANOVA-like procedure, the selection of the appropriate sums of squares for testing the significance of effects is important. When ANOVA or MANOVA designs are not balanced (i.e., when cell frequencies are unequal), to ensure that effects containing the same factors (e.g., A and A*B) are orthogonal, one must employ Type III or Type IV estimable functions. Type IV estimable functions assume that coefficients of any effect are equally distributed across higher-level effects that contain that effect. The result is that significance tests from different designs are based on the same sets of computations. When there are no missing cells in a design, Type III and Type IV sums of squares hypothesis tests are identical (SAS, 1985, p.91).

MANOVA Models Employed in this Research. The different MANOVA designs that were used to test the research hypotheses are described below. More in-depth discussions of the rationale for each analysis are contained in the section of results pertaining to each particular analysis.

(1) Job and predictor order effects were tested using a 2x2 factor MANOVA model containing these two main effects and their joint interaction effect. Two analyses were conducted, one each for validity judgments made for "clerical" and "sales" job title. Therefore, the vector of

dependent variables in each analysis consisted of 28 validity judgments. To investigate the interactions of the between-subject (job- and predictor-order) and within-subject (types of job, predictor, and criterion) factors, a second, exploratory analysis was conducted using a repeated-measures MANOVA model.

(2) <u>Job information effects on validity judgments</u> were tested using two separate MANOVA models. The first model was used to test Hypothesis #1. The 2x2 factor MANOVA model contained job information and validity generalization information main and interaction effects. The dependent variable vector consisted of all 28 judgments which were made based on identical SKAP information. Two MANOVA analyses were conducted, one each for clerical and sales SKAP judgments. Univariate ANOVAs were then conducted to identify which of the 28 validity judgments contributed to the rejection of the null hypothesis. It should be noted that when SKAP information was identical, experts in the "congruent" information condition were provided with a job title that matched the SKAP information, and experts in the "incongruent" condition had the alternative, non-matching title. Therefore, tests of the null hypothesis were tests that the job title/general job description had no effect. Repeated measures MANOVAs were also conducted on these same two vectors of validity judgments to investigate interactions of the job information manipulation with predictor and criterion factors.

The second MANOVA model used in the analysis of job information effects was identical to the first model with the exception that the vector of dependent variables consisted of validity judgments made under identical job title conditions (i.e., the SKAPs differed). Two MANOVA analyses were conducted, one each for "clerical" and "sales" job title judgments. These MANOVAs were used to determine whether judges' validity estimates reflected the relative importance of the different SKAP importance ratings. Taken together, the two sets of analyses provide a comprehensive description of the way in which judges used the job information which was provided in the judgment task.

(3) Validity generalization effects on validity judgments (Hypothesis #2), were investigated using two 2x2 MANOVAs - one each for validity judgments made for the "clerical" and "sales" job titles. The MANOVA model was identical to the second MANOVA model used in the analysis of the effects of job information. These analyses focused on judges' estimates that were made when job titles were held constant but SKAP information varied. The rationale for this model is that validity generalization results were "matched" to the job title, and that to assess the effects of this information, it is necessary to hold the job title constant. Univariate ANOVAs were again computed for each of the individual validity judgments in each of the two job titles. Repeated measures MANOVAs, identical to those described previously, were also conducted to examine

validity generalization and within-subject factor interactions.

(4) <u>The effects of job information and validity generalization information on confidence judgments</u> were assessed using a 2x2 MANOVA model. Two models were tested: one which utilized the 28 confidence judgments for "clerical" job title validities in the response vector and the other which utilized the 28 confidence judgments for "sales" job title validities. To test Hypothesis #3, the main effect for validity generalization information was examined. To test Hypothesis #4, the main effect for job information was examined. Subsequent to the MANOVA analyses, ANOVA analyses were performed to identify job and validity generalization main effects on individual validity judgments. Lastly, exploratory analyses were performed using a repeated measures MANOVA model identical to those described previously. The repeated measures MANOVAs were computed separately for clerical and sales confidence judgments.

CHAPTER V

RESULTS

This chapter is divided into four sections. The first section describes preliminary analyses which were conducted to identify possible sources of sample bias and to ascertain whether the order in which judgments were made during the judgment task affected experts' validity estimates. The second section presents data relevant to the primary hypotheses of this research: the effects of job information and validity generalization information on judgments of test validities for two different jobs. The results of statistical analyses which measured the effects of experimental manipulations on confidence interval judgments are presented in the fourth section.

Preliminary Analyses for Possible Sample Bias and Order Effects

Sample Bias

The effects of job information and validity generalization information on judgments of test validities were studied among professionals from the field of Industrial/Organizational Psychology. Subjects in the study, who were all members of Division 14, the Society for Industrial and Organizational Psychology, Inc. (SIOP), of the American Psychological Association (APA) during the 1986-1987 membership year, were randomly assigned to one of four experimental conditions (which will be referred to by number in the statistical tables):

1. congruent job information/no validity generalization information provided,
2. incongruent job information/no validity generalization information provided,
3. congruent job information/validity generalization information provided,
4. incongruent job information/validity generalization information provided.

Membership Category. Table 1 shows the number of surveys sent and the number of respondents by Division 14 membership category. A chi-square test of cell frequencies indicates that response rates in each of the four experimental conditions are independent of Division 14 membership category (^{2} = 4.80).

Educational Background. Because of the nature of the judgment task employed in this research, it is possible that education and/or work experiences would have an impact on experts' validity estimates. Therefore, it was necessary to ensure that these background characteristics were comparable across experimental conditions. Table 2 shows the educational background of the entire expert sample and across the four experimental conditions. Of prime consideration is the level of educational degree earned ("Level of Education") and the field of study ("Type of Degree"). (It was not possible to conduct chi-square tests of response rates by educational level or field of study because of the plethora of empty and low frequency [n<5] cells.)

The majority of respondents (85.1%) earned the highest degree in their field - the Ph.D. or Ed.D. A much smaller number (14.9%) earned the Masters degree or had fulfilled all requirements for the Ph.D. with the exception of the dissertation. The average number of years from the time the degree was earned across the entire sample was 19.47 (s.d.=13.20). This statistic suggests that the sample was comprised of many senior individuals from Division 14. Given the large number of Fellows who were sampled and responded to the survey (relative to their actual membership rate in Division 14), this finding is not surprising. Rather, it further substantiates the use of the term "expert" to describe the subject population. Because 27.2% of the respondents were Associate members of Division 14, and yet only 14.9% of the respondents had earned less than the Ph.D., it appears that division membership status in not an accurate indicator of educational level within this sample.

The most common field of study of respondents was Industrial/Organizational Psychology (I/O). Thirty-nine of the 95 individuals who received the Ph.D. (41.7%), and 11 of the 17 individuals who were educated at the Masters level (64.7%), received their degree in I/O. Many of the respondents (26.3%) received a graduate degree in General Psychology. This degree was primarily granted to those receiving the Ph.D. (29 of 30). This type of degree was more common in graduate programs twenty years ago and more, whereas today, most Ph.D. degrees in Psychology are likely to be based on more specialized fields of study.

A comparative analysis of respondents from within each of the four experimental conditions suggests that they share similar educational backgrounds. As indicated in Table 2, the subsamples within each of the experimental conditions are primarily comprised of Ph.D. psychologists who have majored in either I/O or General Psychology.

To determine whether individuals who had educational degrees outside of the area of I/O Psychology had other educational exposure to the field, subjects were asked to specify any area of specialization within their graduate studies. Only 54 of the 113 respondents (47.8%) who provided data on their educational background also indicated an area of specialization. Of these 54 individuals, 44 did not have a Ph.D. in I/O Psychology. Twenty-three of these 44 respondents (52.3%) reported that they had specialized in I/O, 6 (13.6%) reported specializing in quantitative methods, and 3 (6.8%) reported specializing in organizational performance/behavior. The remaining 12 individuals (27.3%) specialized in counseling/clinical, experimental, or consumer behavior. These data indicate that of the entire sample of 114 respondents, 73 (64.0%) either earned a graduate degree or specialized in I/O Psychology. A large percentage of the remaining individuals had graduate training in complementary fields such as psychometrics or organizational behavior.

It should be noted that the preceding analysis only considered the educational background of respondents to the survey. Since expertise is acquired through educational and professional experiences, some of the respondents with educational degrees outside the area of I/O Psychology may have developed expertise in this field through work experiences. Therefore, it is likely that the 64% statistic cited above is an underestimate of the professional expertise possessed by the expert sample employed in this research.

Work Background. Table 3 shows the work background of the respondent sample by current affiliation and type of job. As indicated in the "total" column, 43 respondents (37.7%) were associated with academic institutions. Over one-fifth of the sample (23.7%) were employed by a private consulting firm or were self-employed. Twenty-one respondents (18.4%) were employed in private industry and eleven (9.6%) held jobs in the public sector.

Each respondent was asked to provide his/her current job title and a brief description of current job responsibilities. As might be expected, quite a few unique job titles resulted from this line of questioning. Any respondent who provided a job title in the area of testing or selection research, management research, personnel research or psychology, selection exam development, organizational research, etc. or who described his/her work as being primarily industrial or organization research, was classified into the area of "applied I/O research". Only academic job titles were not reclassified into this category. With the exception of the 43 academicians, the current work background of 54 respondents (77.1% of the 70 non-academicians) could be described as "applied I/O research". The remainder of the non-academicians were either internal management consultants (n=9), managers (n=4), marketing researchers (n=2) or counselors (n=1).

On the basis of the preceding analysis of the educational and work background of the research sample, it was concluded that the respondents to the survey possessed sufficient expertise within the field of I/O Psychology to be used in the judgment task employed in this research. Based on the diversity of Division 14 members' professional backgrounds, and the fairly homogeneous educational and work experiences of this sample, it is likely that some self-selection took place so that those more qualified to perform the task returned surveys. Evidence of this inference, though not substantial, was observed from the eight surveys that were returned by individuals who "did not feel qualified to participate in the research." These analyses also indicated that there were no systematic differences in the educational or work experiences of respondents within each experimental condition.

Order Effects on Validity Judgments

For this judgment task, judges were required to estimate validities for seven different cognitive abilities for each of four different types of performance measures. These 28 validity judgments were made for two different jobs: clerical and sales. Because little is known about the processes judges used to produce their validity estimates, it was important to investigate whether the order in which these judgments were made affected judges' estimates of test validities.

To identify and control potential order effects, the order in which validities were judged was manipulated in two ways: 1) the clerical job was presented first or second; and 2) cognitive predictors were presented in two different orders (e.g., general mental ability, verbal, quantitative, perceptual speed, memory, spatial/mechanical, and psychomotor versus the reverse order). Judges within each experimental condition were randomly assigned to one of the four different judgment orders that resulted from crossing these two order factors.

Table 4 contains descriptive statistics for different categories of validity judgments for each of the judgment order conditions. Table 5 contains the results of two 2x2 MANOVAs which measured job- and predictor-order main and interaction effects on the sets of 28 validity judgments made for "clerical" and "sales" job titles.

A visual inspection of the average overall validities reported in Table 4 shows them to be fairly homogeneous. As can be seen at the top of Table 5, none of the order manipulations had a significant effect on the linear combination of "clerical" validity judgments. That is, judgments of clerical validities were similar whether clerical validities were judged first or second and no matter in which order predictors were presented. In addition, level of validity when the clerical job was presented first or second was the same no matter in what order the predictors were presented.

When judgments were made for the "sales" job title, predictor order had a significant affect on experts' judgments ($p.<.05$). This effect can be observed in Table 6 which contains average validity judgments by predictor order. There was a tendency for judgments of verbal, quantitative, and spatial/mechanical "sales" validities to be lower for the second predictor order. (Although a similar trend appears to be occurring in the "clerical" validity estimates, the effect was not large enough to be statistically significant.)

Table 7 contains results of two repeated measures MANOVAs based on job- and predictor-order (between-subjects) effects and predictor and criterion (within-subjects) effects. These MANOVAs show similar trends in judgments for both "clerical" and "sales" validity judgments, namely, that predictor order interacted with type of criterion ($p. < .05$). This significant interaction suggests that level of validity across types of criteria is not the same for individuals in the two conditions of predictor order. This effect can be observed in Table 6 where, in general, average "ratings", "rankings", and "work samples" validities are lower in the second order condition and average "production" validities are higher in the second order condition. However, the average difference in level of validity across the criterion types for the two predictor orders is only .02. Therefore, while predictor order did not affect level of validity across the composite vector of "clerical" validities, the manipulation did have a small, but statistically significant effect on the level of "clerical" and "sales" validity across the criterion validity vectors.

While the finding that predictor order affects "sales" (but not "clerical") validity judgments has implications for conclusions about factors affecting judges' validity estimates, this one significant order effect does not render the remainder of the study uninterpretable. Order conditions were randomly assigned within each of the experimental conditions, thus counterbalancing this effect across the factors being studied.

Results of Hypothesis-Testing: Effects of Job Information and Validity Generalization Information on Validity Judgments

The principal focus of this research was on the effects of job and validity generalization information on experts' validity judgments. This section, which reports results of analyses assessing these effects, is divided into three main parts. The first part reviews analyses which were conducted to describe experts' validity judgments. The descriptions include relevant univariate t-tests which compare average individual validities and average judgments across categories of validities where they are appropriate. The second part of this section reports more rigorous analyses of the effects of the experimental manipulation of job information on validity judgments (Hypothesis #1). The analyses consist of univariate and multivariate tests conducted separately for validity judgments made when clerical SKAPs versus sales SKAPs were provided and job titles were held constant. These analyses are contrasted to comparable analyses for validity judgments made when "clerical" versus "sales" job titles were provided and SKAPs were identical. This section also includes results of repeated measures MANOVA analyses which examine both between- and within-subject effects on validity judgments. (The results of the repeated measures MANOVAs are intended to be more descriptive in nature since there were no hypotheses offered regarding the effects of predictor or criterion factors on judgment outcomes.) The third part reports the results of MANOVA and ANOVA analyses which measured the effects of the validity generalization manipulation on validity judgments for "clerical" and "sales" job titles (Hypothesis #2). These analyses are similar to those described above for measuring the job information effect.

Descriptions of Experts' Validity Judgments

Fifty-six judgments of cognitive validities were elicited from each of 114 experts under four different experimental conditions. Each judgment represented a judge's estimate of validity for one of seven different cognitive abilities for predicting each of four different performance measures, for one of two different jobs.

Two types of information were hypothesized to affect experts' validity judgments and were manipulated in this research: job descriptive information and validity generalization information. Table 8 shows the mean estimate of validity for different categories of judgments within and across all experimental conditions. The results in this table indicate some of the trends in experts' judgments that were further investigated and are reported later in this section.

Experts' average validity estimates across the experimental conditions and different types of predictors and criteria ranged from .12 to .42 (x = .28). The variance in judges' estimates was fairly consistent; standard deviations ranged from .07 to .14. Average estimated validity was consistently higher across clerical judgments in comparison to sales judgments (x= .31 vs. .26). Judges were also consistent in the relative magnitude of estimated validities for the seven predictors across all experimental conditions. With only a few exceptions, the magnitude of average estimated validities followed this order: general mental ability (GMA), verbal ability, quantitative ability, perceptual speed, memory, spatial/mechanical ability, and psychomotor ability. Similarly, experts' average criterion validities tended to be highest for the work sample criterion and lowest for ratings and production criteria.

General effect of job information. The effect of "congruent" (i.e., job title matched SKAPs) versus "incongruent" (i.e., job title did not match SKAPs) job information on overall validity judgments (i.e., not disaggregated for clerical and sales job titles or SKAPs) is illustrated by comparing judgments in experimental conditions #1 to #2 and by comparing judgments in experimental conditions #3 to #4 from Table 8. The first comparison shows the sole effect of congruent versus incongruent job information; the latter comparison includes the additional effect of validity generalization information.

The first notable trend is that job information appears to affect clerical and sales validity judgments similarly when validity generalization information is not present. As Table 8 shows, when job information is incongruent, overall validity increases (.26 vs .29 for clerical validities and .23 vs. .29 for sales validities). This increase for clerical validities is not statistically significant ($t = 1.03$, $p.<.30$), but is statistically significant for sales validities ($t = 2.41$, $p.<.02$). Job information does not appear to have a significant affect on clerical or sales judgments when validity generalization information is present in the judgment task. Overall validity slightly decreases for clerical judgments when job information is congruent (.35 vs. .33; $t = .64$, $p.<.53$) and slightly increases for sales judgments when job information is incongruent (.25 vs. .28; $t = .99$, $p.<.33$).

The reader is reminded that when job information was incongruent, the SKAPs provided to judges were actually those that were congruent with the alternate job title/job description. Therefore, if SKAP information was influential when experts were making their validity estimates (assuming that job title has no effect), then one would expect clerical validities in the congruent condition to be comparable to sales validities in the incongruent condition.

Likewise, sales validities estimated in the congruent condition should be very similar to clerical validities estimated in the incongruent condition.

Table 8 shows the effect of job information was not consistent for clerical and sales validity judgments. More specifically, when validity generalization results were not provided, the average overall clerical validity in the congruent condition (#1) was not significantly different from the average overall sales validity in the incongruent condition (#2) (.26 vs. .29; $t = -1.03$, $p.<.31$). However, the average overall sales validity in the congruent condition (#1) was significantly less than the average overall clerical validity in the incongruent condition (.23 vs. .29; $t = -2.62$, $p.<.02$).

This incomparability of clerical and sales judgments when SKAP information was identical was also evidenced in the judgments made when validity generalization results were available. However, the magnitude of the validity difference was greater (.35 vs. .28 for clerical SKAPs and .25 vs. .33 for sales SKAPs). These results are both significant ($t = 3.09$, $p.<.003$ for clerical SKAPs and $t= -3.22$, $p.<.002$ for sales SKAPs).

The preceding analyses provide tentative support for Hypothesis #1 - that judgments of validity will be significantly different for jobs with different job titles even when SKAP information is identical. The intent of these analyses is to provide a description of the overall trends apparent in the data regarding the effects of job information. More rigorous tests of this hypothesis, including univariate and multivariate analysis of variance, were conducted which follow later in this section.

General effect of validity generalization information. Expert judges estimated validities either in the presence or absence of validity generalization information. Two types of comparisons are relevant to describing the overall effect of validity generalization information. First, is important to know whether the presence of this information has an affect on overall validity judgments. Second, it is important to assess whether any affect of this information is consistent with the actual true validity estimates that were provided. (The average overall clerical and sales validities reported in the validity generalization results were .41 and .18, respectively.)

A comparison of average overall validity when validity generalization information is absent versus present indicates that the availability of these empirical summaries serves to marginally increase experts' estimates of cognitive validities (.27 vs. .30). The increase is almost statistically significant ($t = 1.9$, $p < .06$). The effects of validity generalization information on experts' judgments of validities can be seen most clearly in Table 8 by comparing the average validity judgment for the "clerical" (or "sales") job title from experimental condition #1 to that in experimental condition #3. This comparison is most meaningful because job information is congruent, thereby illustrating how validity generalization results affect expert judgments in an unambiguous judgment situation (i.e., where job title, SKAPs and validity generalization results are all congruent). In both of these comparisons, average overall validity increases, however the increase is statistically significant for clerical validities (.26 vs. .35; $t = 3.22$, p.<.002) but not for sales validities (.23 vs. .25; $t = .57$, p.<.57).

When job information is incongruent, the validity generalization results that are available correspond to the job title but not to the SKAPs. Therefore, experts might reason that if the profile of skills and abilities is not illustrative of a clerical (sales) job, then the clerical (sales) validity generalization results do not provide an accurate indication of the level of empirical validity and should not influence validity estimates. In fact, the presence of validity generalization results does not consistently affect the level of validity in clerical and sales jobs when job information was incongruent. There was a trend for average overall clerical validity to increase in the incongruent condition when validity generalization information was provided (.29 vs. .33; $t = -1.78$, p.<.08). However, there was no difference in average overall validity when validity generalization information was provided for the sales job in the incongruent condition (.29 vs. .28; $t = .23$, p.<.82).

It should be noted that the effects of providing validity generalization results were not entirely consistent with the actual validity generalization results that were provided. The average overall clerical and sales validities reported in the validity generalization results were .41 and .18, respectively. The presence of the clerical results increased experts' overall clerical validity judgments in the congruent condition from .26 to .35. However, despite the fact that the average sales validity generalization result was .18, the presence of these results also increased experts' overall sales validity judgments from .23 to .25 in the congruent information condition.

These comparisons, which suggest that the presence of validity generalization results tend to raise experts' overall validity judgments regardless of the magnitude of the actual validity generalization data, provide tentative, partial support for Hypothesis #2. This hypothesis stated that judgments of validity for clerical and sales jobs will be significantly different when validity generalization results are provided than when they are not provided. It should be noted that for sales judgments, although validity generalization information appears to have an affect, this affect is surprisingly not consistent with the data provided.

It would be important to measure how job and validity generalization information affect experts' judgments of different types of predictor and criterion validity judgments and individual predictor/criterion validities. However, the use of t-tests on multiple outcome variables increases the experiment-wise risk, lowering the significance level of statistical tests and inflating the error rate. To overcome this problem, these tentative conclusions regarding the effects of experimental manipulations were more rigorously tested using MANOVA and ANOVA procedures. The results of these analyses are reported in the following paragraphs.

Effects of Job Information on Judged Validity

Table 9 contains descriptive statistics for validity judgments made for the "clerical" job title and Table 10 contains the same information for the "sales" job title. To compare the judgments made while SKAPs were identical and job titles differed (i.e., the judgments comprising the response vectors in the first set of MANOVA analyses described below), one needs to compare across these two tables. For example, the average general mental ability/rating validity for sales SKAPs was .36 in the "congruent" condition - "sales" job title - (column #2 in Table 10), and .42 in the "incongruent" condition - "clerical" job title - (column #3 in Table 9).

To determine whether the trends observed above were significant, a multivariate analysis of variance statistical procedure was applied to the judgment data. The MANOVA model consisted of job information and validity generalization information main effects and interaction effect. Two different types of MANOVAs were conducted. In the first MANOVA model, the vector of dependent variables consisted of the 28 validity judgments that were made when either clerical or sales SKAPs were provided (i.e., the job information condition was that SKAPs were identical but job titles differed). The results of these MANOVA analyses are reported separately for clerical and sales SKAP judgments in Table 11. The second MANOVA model was identical to the model described above with the exception that the vector of responses consisted of validity judgments made for either the "clerical" or "sales" job title (i.e., job titles were identical but SKAPs differed). The results of these analyses are reported in Table 12. Results of univariate ANOVAs, which measured the effects of job information on individual validity judgments, are reported in Tables 13 and 15 for judgments made for identical SKAPs and job titles, respectively.

The reader is reminded that when SKAP information was identical, the difference between experts making validity judgments in the "congruent" condition and experts making validity judgments in the "incongruent" condition was that in the former condition, the job title matched the SKAPs that were provided. (The reader is also reminded that general job descriptions that accompanied the job title were ambiguous, and almost identical in content.) Therefore, the significant main effect for job information (Table 11) means that judges who were provided with identical SKAPS judged validities differently solely as a function of having been provided with different job titles/general job descriptions.

Tables 14 and 16 indicate the within-subjects (predictor and criterion) effects on experts' validity judgments that were made for clerical/sales SKAPs and clerical/sales job titles, respectively.

Job information effects on overall validity judgments for clerical SKAPS. As indicated in Table 11, when the vector of dependent variables was comprised of a linear combination of 28 validity judgments that were made when clerical SKAPS were provided, the job information main effect was significant ($p. < .001$). By comparing column #2 from Table 9 with column #3 from Table 10, the differences in the actual mean validity judgments can be seen. Average judgments appear to be higher for verbal ability validities in the incongruent condition and higher for quantitative, perceptual, spatial/mechanical, and psychomotor validities in the congruent condition. The likelihood ratio of .577, shows that the effect of the (incongruent versus congruent) job titles was fairly substantial. In fact, 42% of the variance in judgments can be attributed to differences in job titles/general job descriptions in this research. It should be noted that we could expect shrinkage in explained variance in future samples.

The interaction effect of job information and validity generalization information was also significant (p.<.009), suggesting that the effect of congruent and incongruent job titles was not the same when validity generalization was present versus absent. Comparisons of vectors from experimental conditions #1 and #3 on Table 9 with vectors from experimental conditions #2 and #4 on Table 10 show that validity judgments for clerical SKAPs tend to be slightly higher when job information is incongruent and no validity generalization information is available. When validity generalization information is available, the judgments tend to be much lower when job information is incongruent than when it is congruent. This effect is difficult to interpret, however, because identical validity generalization information was not provided to judges in the "congruent" and "incongruent" conditions[19]. Judges in the "incongruent" condition received "sales" validity generalization results (although not identified as such) while those in the "congruent" condition received "clerical" validity generalization results. The differences in the

[19] To determine whether this potential confound in fact had an affect on the results observed for clerical SKAPs, additional MANOVA analyses were subsequently conducted using only data from judges in the "no vg" condition. Therefore, only the job title effect was tested. The result of this analysis showed that job title had a non-significant affect on clerical SKAP judgments (p.< .79). Therefore, one would conclude that the significant effects measured in the original MANOVA were due to the confound of VG condition with job title.

levels of their judgments may be reflecting the actual differences evident in the validity generalization results provided for these two job titles. (A more meaningful statistical assessment of this effect is reported in the third part of this section.) As a result of this significant interaction effect, it is difficult to interpret the job information main effect for clerical SKAP judgments.

Table 12 contains results of MANOVA analyses that measured the effect of SKAP information on validity judgments made for the "clerical" job title. Therefore, this analysis measures the effect of incongruent versus congruent SKAP information when the job title is held constant. In conjunction with the previous analyses assessing the effect of incongruent versus congruent job titles, these analyses provide a full description of how judges responded to the job information that was provided in the validity judgment task.

As indicated in Table 12, incongruent versus congruent SKAPs had a highly significant affect on validity estimates for the "clerical" job title ($p. < .0001$). This is illustrated by comparing columns #2 and #3 on Table 9 and, additionally, by referring to the SKAP information that was provided (see Appendix D). This significant effect for job information means that judges who were provided with identical job titles judged validities differently as a function of having been provided with different SKAP information. Specifically, it appears that experts' validity judgments are strongly affected by the <u>relative</u> importance of the SKAP information that was provided such that the magnitudes of their judgments (across the congruent and incongruent conditions) correctly reflected the importance of four of the seven abilities. (For example, the average clerical validity estimates from experts in the incongruent condition were higher than the estimates from experts in the congruent condition for verbal ability. This result correctly reflects the information provided that verbal ability is a substantially more important ability in the incongruent condition than in the congruent condition.) These findings suggest that judges were generally sensitive to the relative importance ratings provided as part of the judgment task materials.However, as previously demonstrated, the actual magnitudes of validities were also strongly affected by the job title that was provided.

Taken together, the results for clerical validities reported in Tables 11 and 12 provide partial support for Hypothesis #1 which predicted an independent effect for job title. First, as indicated in Table 11, when SKAP information was identical, judgments of clerical validities were significantly different when judges were provided with a "clerical" versus a "sales" job title. Second, as indicated in Table 12, when job title was identical, judgments of clerical validities were significantly different when clerical versus sales SKAP information was provided. The support for Hypothesis #1 is partial because although job title had a significant independent affect on clerical validity judgments, judges were also strongly influenced by the SKAP information that was provided.

To identify those mean judgments that contributed to the the significant job information and job information x validity generalization information effects in the analysis of validity judgments for clerical SKAPs, univariate ANOVAs were conducted. The results of these ANOVAs are discussed in detail below.

Job information effects on individual validity judgments for clerical SKAPS. Table 13 contains results of univariate ANOVAs which measured the effects of job information and validity generalization information on judgments made when clerical SKAPs were provided. A review of the univariate results for the clerical SKAP validity judgments in Table 10 suggests why the MANOVA main effect for job information is not interpretable. Although the job information main effect was significant ($p.<.05$) for six (21.4%) of the validity estimates, half of these significant main effects were associated with significant interaction effects. In addition, there does not appear to be any meaningful pattern to those validities that evidence significant main effects for job information. As indicated in Table 13, the significant effects on clerical SKAP judgments are spread across all types of predictors and criteria. Therefore, while in some cases judges' clerical validity judgments were somewhat affected by differences in job title, these effects were not substantial in number nor were they systematic for any particular type of predictor or criterion measure.

Job information effects on overall validity judgments for sales SKAPS. Table 11 contains the results of MANOVAs based on judgments made when sales SKAPS were provided. In this analysis, the main effect for job information is highly significant ($p < .0001$), supporting the hypothesis that validity judgments would be significantly different when SKAP information was identical but job titles differed. By comparing column #3 from Table 9 to column #2 from Table 10, one can see why the null hypothesis was rejected. Clearly, when sales SKAPs were provided, mean validity estimates were higher when judges were provided with the "clerical" job title rather than the "sales" job title. More than half of the variance in the linear combination of sales validity judgments can be attributed to differences in job titles alone, suggesting that judges' validity estimates were strongly influenced by the job title irrespective of the SKAP information that was provided. This effect was in the general direction of higher sales SKAP validity estimates when a "clerical" job title was provided.

The interaction of job information and validity generalization information was also significant when experts' judgments were based on sales SKAPs ($p < .05$). However, because of the strength of the job information main effect, it is likely that job information affected experts' validity judgments independent of the validity generalization effect[20]. More specifically, sales SKAP validity judgments are lower for the "sales" job title than for the "clerical" title whether validity generalization information is or is not present. However, the difference in average validities is smaller when validity generalization information is not present. (This conclusion is illustrated by comparing column #6 on Table 10 to column #7 on Table 9 and by comparing column #8 on Table 10 to column #9 on Table 9.)

Table 12 indicates the effect of congruent and incongruent SKAPs on validity judgments for the "sales" job title. As in the previous analysis for the "clerical" job title, the SKAP information had a highly significant effect on experts' validity estimates ($p.<.0001$), such that their sales validity estimates were influenced by the SKAP information that was provided. As can be observed by comparing columns #2 and #3 in Table 10 (i.e., "sales" job title with clerical and sales SKAPs, respectively), and by referring to the actual SKAP information in Appendix D, the magnitude of experts' validity estimates correctly reflected

[20] As discussed in the MANOVA analysis for clerical SKAP judgments, there is a potential confound of VG condition with job title. This confound was examined by conducting an additional MANOVA analysis using only data from judges in the "no vg" condition. The result of this analysis showed that job title had an almost significant affect on sales SKAP judgments ($p.<.14$). Due to the relatively low power associated with conducting a MANOVA with 28 dependent variables and only 61 subjects, it was decided to investigate the job title effect on individual validity judgments using ANOVA (see footnote #21).

the relative importance of five of the seven types of predictors. (For example, the average validity estimates from experts in the incongruent condition were higher than the validity estimates from experts in the congruent condition for both quantitative ability and perceptual speed. These results correctly reflect the information provided that quantitative ability and perceptual speed are substantially more important abilities in the incongruent condition than in the congruent condition.) These findings suggest that judges were generally sensitive to the relative importance ratings provided as part of the judgment task materials.

Despite the fact that there was a tendency for experts' judgments of sales validities to reflect the relative importance of the SKAP information that was provided, the magnitude of their estimates was also somewhat influenced by the job titles for which they were making their estimates. This is evidenced not only in the inappropriate use of the SKAP information for memory and psychomotor ability, but because of the significant effect of different job titles on sales SKAP judgments. Therefore, it appears that there is an independent effect that can be attributed solely to the job title that was available when validities were being judged.

Taken together, the results for sales validities reported in Tables 11 and 12 provide partial support for Hypothesis #1 which predicted an independent effect for job title. First, as indicated in Table 11, when SKAP information was identical, judgments of validities were significantly different when a "clerical" versus "sales" job title was provided. Second, as indicated in Table 12, when job title was information was held constant, judgments of sales validities were significantly different when clerical versus sales SKAP information was provided. The support for Hypothesis #1 is partial because although job title had a significant independent affect on sales validity judgments, judges were also strongly influenced by the SKAP information that was provided.

These MANOVA analyses for sales SKAP and job title validities were followed-up by 28 univariate ANOVAs which were conducted using the 28 sales SKAP validity judgments. These analyses were conducted to identify those mean judgments that contributed to the significant main effect for job information.

Job information effects on individual validity judgments for sales SKAPS. Table 13 contains results of univariate ANOVAs for each of the validity judgments made when sales SKAPs were provided. Of the 28 ANOVAs, 19 (67.9%) evidenced significant main effects for the job information condition. Significant job x validity generalization effects were associated with only three of these significant main effects[21]. Based on these results,

[21] These ANOVA results were essentially replicated using only judges who were in the "no vg" condition. Thirteen significant ANOVA effects were found which duplicated 13 of the 16 significant ANOVA effects found in the original

one would conclude that the statistically significant main effect for job information that was identified in the previously discussed MANOVA analysis is meaningful.

The 16 interpretable main effects for job information (i.e., where differences in job titles resulted in significantly different average validity estimates) occurred in all of the four criterion validity judgments involving quantitative ability, perceptual speed, memory, and psychomotor ability predictors. These results indicate several systematic trends in judges' responses to "clerical" and "sales" job titles which are described below in detail. (If one compares column #3 in Table 9 to column #2 in Table 10, the actual mean validity judgments that were measured in these effects are indicated.)

For quantitative ability validities, experts in the incongruent condition had higher validity estimates for each of the four types of criteria. The difference in average quantitative validities for judges in the congruent (i.e., "sales" job title) and incongruent (i.e., "clerical" job title) conditions was .12 for rating, .13 for ranking, .10 for work sample, and .11 for production criteria.

When estimating perceptual speed validities, judges in the incongruent condition tended to have higher average estimates than judges in the congruent condition. The differences in their judgments were .09, .07, .11, and .11

---

analyses. Based on these results, it was concluded that the previously described confound was not affecting the results and that job title had a significant affect on sales SKAP validity judgments.

for rating, ranking, work sample, and production validities, respectively.

For validities involving memory as the predictor, judges in the incongruent condition made estimates that were .10 higher for a rating criterion, .09 higher for a ranking criterion, .09 higher for a work sample criterion, and .11 higher for a production criterion.

Judges in the incongruent condition estimated validities for psychomotor ability that were, on the average, .065 higher than estimates made by judges in the congruent condition. Their judgments were .06 higher for rating, .04 higher for ranking, .08 higher for work sample, and .08 higher for production.

In summary, the results of the univariate analyses for sales SKAP validity judgments show a clear and very consistent pattern of higher validities for those judges who were provided with the "clerical" rather than the "sales" job title. These results provide empirical support for Hypothesis #1 and suggest that judges' validity estimates were strongly affected by the title assigned to the job, irrespective of the type of SKAP information that was provided. The tendency to estimate quantitative, perceptual speed, memory and psychomotor predictors higher when a clerical job title was provided is somewhat consistent with implicit beliefs regarding the appropriateness of these cognitive abilities for predicting performance in clerical jobs.

Interaction of job information with within-subject factors. Table 14 contains results from repeated measures MANOVAs which were conducted separately for the validity judgments made for clerical and sales SKAPs. These analyses show the between- and within-subject effects on validity judgments and provide some insight into the analyses that were conducted to measure job information main effects.

First, it should be mentioned that "Type of Predictor" and "Type of Criterion" had significant effects on linear combinations of validity judgments that were based on clerical and sales SKAPs. These main effects for predictors and criteria were highly significant for both sets of SKAP-based judgments ($p < .0001$). As indicated by both Wilks's and the canonical correlation coefficient, judges were highly sensitive to the type of predictor and the type of criterion for which they were judging validity. The variance in the linear combination of validities accounted for by the predictor factor was 77% for clerical SKAPs and 81% for sales SKAPs. The variance in the weighted outcome vector accounted for by the criterion factor was 52% for clerical SKAPs and 48% for sales SKAPs. Despite the fact that some of the interactions involving the predictor and criterion factors were significant, the relative strength of these main effects strongly suggests that experts' validity judgments systematically discriminated on the basis of the type of cognitive ability and the type of performance measure on which the validity coefficient was based.

Also notable is the interaction of type of predictor and type of criterion. This effect was significant for both clerical and sales SKAP validity judgments (p.<.0001 and p.<.0001, respectively). The interpretation of this interaction effect is that the effect of type of predictor on SKAP-based validity judgments is not constant across all types of criteria. Experts' judgments appear to reflect the professional belief that not all predictors are equal at predicting different aspects of job performance. Once again, however, this lower-order interaction effect is contained in statistically significant higher-order interaction effects. Specifically, the predictor x criterion interaction (the effect for individual validities) interacts with the job x vg information interaction when the outcome vector contains clerical SKAP judgments (p.<.05) and with job information when the outcome vector contains sales SKAP judgments (p.<.05). These higher order interactions suggest that experts' mean estimates of validities are not the same across different combinations of predictors and criteria and across different judgment conditions (i.e., job information) or combinations of judgment conditions (i.e., job x validity generalization information).

Effects of Validity Generalization Information on Judged Validity

The effects of validity generalization information on judges' validity estimates were tested using 2x2 factor MANOVA and ANOVA models. Analyses were conducted separately for clerical and sales judgments. These statistical models were identical to those used to measure the effects of the job information manipulation, with the important difference that the outcome vector in each analysis consisted of validity judgments made for either the "clerical" or "sales" job title, irrespective of the SKAP information that was provided. As was previously discussed, it is more meaningful to examine the effects of validity generalization information where the information which was provided "matched" the job title that was also provided. (If the validity generalization information was "matched" to the SKAP information rather than to the job title, we would not be able to ascertain whether the validity generalization information did not have an effect because judges were not influenced by it, or because judges did not consider it relevant to the job for which they were judging validities.)

Expert judges estimated validities either in the presence or absence of validity generalization information. As was previously discussed, two types of comparisons are relevant to describing the overall effect of validity generalization information. First, is important to know whether the presence of this information affects validity judgments. Second, it is important to assess whether any effect of this information is consistent with the actual true validity estimates that were provided. (The average overall clerical and sales validities reported in the validity generalization results were .41 and .18, respectively.)

The results of MANOVAs for "clerical" and "sales" job title validities are presented in Table 12. Table 15 contains ANOVA results for the 28 clerical and 28 sales validity estimates. Table 16 contains results of the repeated measures MANOVAs conducted to measure the interactions of the validity generalization factor with within-subject (predictor and criterion) factors.

Validity generalization information effects on overall validity judgments for the "clerical" job title. When the vector of dependent variables was comprised of the 28 validity judgments made for the "clerical" job title, the validity generalization main effect was highly significant ($p < .001$) (see Table 12). The differences in clerical judgments made with and without validity generalization information can be seen by comparing the average judgments in columns #4 and #5 on Table 9. From this comparison, it is apparent that the presence of clerical validity generalization information increased estimated clerical validities for all but one of the 28 predictor/criterion judgments. (The exception was the judgment involving verbal ability with a production criterion.) The likelihood ratio (=.588) and the canonical correlation (r.=.64) show that the effect of validity generalization information was substantial. Forty-one percent of the variance in the linear combination of the 28 "clerical" job title validities was accounted for by this effect. These results provide clear support for Hypothesis #2.

As can be seen in Table 12, the interaction effect involving validity generalization information and job information was not significant, suggesting that the effect of validity generalization information is the same whether SKAP information matched the job title or did not. Although the linear combination of clerical job validities differed in the congruent and incongruent conditions (F_{calc}=2.61; p.<.0001), this factor did not affect the manner in which validity generalization information was used.

To identify those mean validity judgments which contributed to the rejection of the null hypothesis of "no validity generalization effect", 28 univariate ANOVAs were computed. The results of these analyses are presented in detail below.

<u>Validity generalization effects on individual validity judgments for the "clerical" job title.</u> Table 15 contains results of univariate ANOVAs assessing the effect of validity generalization information on each of the 28 "clerical" job title validity judgments. Because the previous MANOVA found a significant main effect for validity

generalization information, it is important to identify on which of these judgments experts' in the "vg" and "no vg" conditions evidenced significant differences in validity estimates and to determine if these differences are consistent with the validity generalization results that were provided.

As indicated in Table 15, 21 of the 28 ANOVAs (75.0%) contain significant effects for validity generalization information. None of the interaction effects associated with these 21 main effects was significant. The main effects for the validity generalization effect formed a very consistent pattern: six of the seven tests of main effects (85.7%) involving the production criterion were <u>not</u> significant, and only one other validity (i.e., for spatial-mechanical/work sample) was also not significant.

Overall, average validity judgments for the "clerical" job title were higher when judges were provided with validity generalization information. This result is congruent with the fact that actual clerical validity generalization values were generally higher than the average judgments in the "no vg" condition. More specifically, judges in the "vg" condition had higher estimates of validities involving <u>general mental ability</u> than judges who were not provided with this information. The differences in average judgments were .07 for the rating criterion, .13 for the ranking criterion, and .13 for the work sample criterion (see Table 9). Estimates of <u>verbal ability</u> validities showed differences of .03 for rating, .08 for ranking, and .08 for work sample criteria. Judges' estimates of validity for a quantitative predictor differed .05, .08, and .07 for rating, ranking and work sample criteria, respectively. For the perceptual speed predictor, judgments from experts with validity generalization information were higher than judgments from experts without this information by .06 for rating and .10 for ranking and work sample. Average judgments for memory were higher in the "vg" condition by .06 for rating, .07 for ranking, .11 for work sample, and .04 for production. For spatial/mechanical ability, there were significant differences for rating and ranking (.06). Experts' estimates of psychomotor validities were .08, .07, and .11 higher for experts in the "vg" condition for rating, ranking, and work sample criteria.

These findings strongly suggest that in the presence of clerical validity generalization results, judges produce significantly higher average estimates of clerical validities in comparison to when these results are not available. The reader is reminded that the measured effect of validity generalization information is averaged across the two SKAP conditions (i.e., some judges had "clerical" SKAPs and other judges had "sales" SKAPs). Therefore, judges' average estimates of clerical validities were systematically higher in the presence of validity generalization information in spite of the fact that the SKAP information may not have been descriptive of a clerical job. The results--a systematic pattern of higher validity estimates for judges in the "vg" condition--are highly consistent with the clerical validity generalization information provided to judges. For example, for 20 of the 21 judgments (95.2%) for which judgments in the "vg" condition were significantly higher than judgments in the "no vg" condition, the validity generalization results were also higher than the average judgments made by experts who did not have access to this information. These results show judges' clerical validity estimates reflect the presence of (generally higher) validity generalization information.

Validity generalization information effects on overall validity judgments for the "sales" job title. Table 12 reports the results of a MANOVA involving the vector of 28 validity judgments made for the "sales" job title. In this analysis, the main effect for validity generalization information was significant (p.<.05). By comparing columns #4 and #5 on Table 10, one can observe the differences in average validity estimates from the "vg" and "no vg" groups. These differences are not as large as those observed for the "clerical" job title, nor are they as consistent. However, the Wilks' lambda for this effect (.623), and the amount of variance in the response vector which can be attributed to the validity generalization factor (38%), provide further support for Hypothesis #2.

The interaction term involving validity generalization and job information was not significant while the job information main effect was highly significant (p.<.0001). The significant effect for job information shows that judges' validity estimates reflected the congruency and incongruency of job title and SKAP information such that they produced generally higher validity estimates when the clerical SKAPS (versus the sales SKAPs) were provided. (The exceptions to this were for validities involving general mental ability and verbal ability). The lack of a significant interaction effect demonstrates that the effect of validity generalization information on sales validity estimates is the same whether SKAP information matched the job title or did not.

This MANOVA analysis was followed by 28 univariate ANOVAs. Each ANOVA measured the effect of the validity generalization factor on individual validity judgments for the "sales" job title.

Validity generalization information effects on individual validity judgments for the "sales" job title. Results from the 28 ANOVAs involving validities for the "sales" job title are presented in Table 15. Of the 28 validity judgments that could have contributed to the rejection of the multivariate null hypothesis for validity generalization information, only three (10.7%) showed significant main effects for this factor. Specifically, judgments of general mental ability (p.<.05) and verbal ability (p.<.001) with a work sample criterion, and perceptual speed with a production criterion (p.<.05) were significantly different for judges in the "vg" and "no vg" conditions. In the first two cases, average judgments by experts in the "vg" condition were higher; in the third case these experts' judgments were lower.

These results, which are neither as consistent or numerous as the results for the clerical job title, demonstrate the power of MANOVA for detecting group differences across multiple outcome measures. Given the small percentage of significant results for the validity generalization main effect found in these univariate analyses, one would have to make fairly conservative claims about the overall effect of this experimental condition on validity judgments for the "sales" job title.

Interaction of validity generalization information with within-subject factors. Table 16 contains results from repeated measures MANOVAs which were conducted separately for the validity judgments made for "clerical" and "sales" job titles. These analyses show the between- and within-subject effects on experts' validity estimates.

As in the previous repeated measures MANOVA analyses for the job information effect, significant main effects were evidenced for both type of predictor, type of criterion, and the predictor x criterion interaction. For "clerical" validities, the variance in the linear combination of validity judgments that can be attributed to these effects is 77%, 50%, and 50%, respectively. For "sales" validities, the variance in the linear combination of validity judgments that can be attributed to these effects is 83%, 50%, and 49%, respectively. These results suggest that although there is a significant effect for validity judgments (i.e., average validities for different predictors are not the same across different criteria), there is a fairly substantial, independent effect for type of predictor for both the "clerical" and "sales" job title judgments.

An interesting finding from the repeated measures MANOVA is that validity generalization information interacts with types of validity judgments both for the clerical and sales validity judgments ($p. < .05$ and $p. < .001$, respectively). This three-way interaction effect accounts for 26% and 32% of the variance in the linear combinations of clerical and sales validities, respectively. One would interpret this effect to mean, generally, that the effect of validity generalization information (presence or absence) is not the same across different combinations of predictors and criteria. It appears, therefore, that only some of the validity generalization information affected experts' judgments. For the "clerical" job title, results on Table 9 suggest that validity generalization information had the largest effect on almost all judgments involving ranking and work sample criteria. Experts' judgments of work sample and production criteria for general mental ability, verbal ability, quantitative ability, and perceptual speed appear to be most affected by validity generalization information for the "sales" job title (see Table 10).

Effects of Job Information and Validity Generalization Information on Judges' Confidence Interval Judgments

In this section results of analyses which examined the degree to which job and validity generalization information affected the confidence which judges had in their validity estimates are reported. Of primary concern were those analyses which were conducted to test Hypotheses #3 and #4. Hypothesis #3 proposed that the presence of validity generalization information would significantly increase confidence in validity judgments. Hypothesis #4 proposed that when information was "congruent" (i.e., when the job title "matched" the SKAPs that were provided), experts would be more confident in their judgment output.

As might be recalled, experts' confidence judgments were simply the difference between the upper-bound and the lower-bound of the judged 90% confidence interval around each validity judgment. Several different analyses were conducted to examine the effects of the experimental manipulations on judges' confidence judgments. First, descriptive statistics were computed across all experimental factors and conditions for categories of confidence judgments and for the clerical and sales confidence judgments. Second, univariate t-tests were computed to initially examine hypothesized differences in confidence judgments. Third, two MANOVAs were computed, one each for the 28 clerical and 28 sales confidence judgments, to measure the overall effects of the between-subjects factors on expert confidence. Fourth, univariate ANOVAs were computed to measure effects of job and validity generalization information on confidence judgments for individual validities. Lastly, a repeated-measures MANOVA (as described in previous paragraphs) was computed using the vector of all 56 confidence judgments. This analysis, exploratory in nature, was conducted to examine the interactions of between- and within-subject factors.

Descriptions of Experts' Confidence Judgments

Table 17 contains descriptive statistics for experts' confidence judgments across different categories of jobs, predictors, and criteria. As indicated in the table, across all of these categories of validity judgments, judges' average judged confidence intervals ranged from .20 to .24, and the standard deviations of these intervals ranged from .06 to .10. From these estimates alone, one would infer that judges' confidence interval judgments are rather consistent across types of predictors, criteria, jobs, and experimental conditions. As predicted, there appears to be a slight (non-significant) tendency for confidence to be lower (i.e., the confidence interval larger) when job information is incongruent. However, there do not appear to be any differences between judges who were provided with validity generalization information and those who were not in the confidence they had in the accuracy of their validity estimates.

To examine the degree to which task relevant information affected judges' confidence in the accuracy of their validity judgments, average confidence intervals were computed for each unique predictor/criterion combination. These descriptive statistics are reported in Table 18 for judgments made for the "clerical" job title, and in Table 19 for judgments made for the "sales" job title. A review of these tables shows that confidence interval estimates where relatively stable. However, there appears to be a tendency for judges' confidence to be higher when job information was congruent. There does not appear to be an effect for validity generalization information. These observations were tested using MANOVA and are described in detail below.

Job Information and Validity Generalization Information Effects on Confidence Judgments: MANOVA Analyses

Because of the experiment-wise risk (i.e., inflated error rates) associated with conducting t-tests on multiple response variables, a MANOVA model was used to test the multivariate hypotheses. A 2x2 (job x validity generalization) MANOVA model was used for testing job and validity generalization effects on confidence interval judgments. Two MANOVAs were conducted, one each for confidence judgments made for "clerical" and "sales" job titles. The results of these analyses are presented in Table 21 and are described below.

Effect of job information. As indicated in Table 20, the job information main effect was not significant for either the clerical (p.<.82) or sales judgments (p.<.41). For the clerical job, only 20% of the variance in the linear combination of confidence judgments can be attributed to the job information condition. For the sales job, the variance accounted for was 27%. Because of these results, it must be concluded that Hypothesis #3 was not supported. Therefore, the effects of job information on individual confidence judgments will not be examined.

Effect of validity generalization information. Table 20 also contains results of MANOVA tests of the effect of the validity generalization manipulation on experts' confidence interval judgments. This effect is not statistically significant for clerical (p.<.12) or sales judgments (p.<.84). The validity generalization manipulation accounted for only 32% and 20% of the variance in the linear response vectors for clerical and sales confidence judgments, respectively. Therefore, this data show no support for Hypothesis #4.

Between- and Within-Subject Effects on Confidence Interval Judgments

A repeated measures MANOVA was conducted to obtain information regarding the interaction of job and validity generalization factors with within-subject factors. This analysis provides information regarding factors (other than those experimentally manipulated) that may have an affect on the rationally-derived confidence intervals surrounding judged validities. Specifically, it would be important to know whether judges are more or less confident in judgments that involve certain types of jobs, predictors, and criteria.

In this analysis, the within-subject effects included type of job, predictor, and criterion factors. It was decided that because none of the measured main effects were significant for either the clerical or sales job, that the repeated measures MANOVA analysis should be conducted for all 56 confidence judgments simultaneously. This would allow for the detection of potential interactions involving type of job. The results of this analysis are presented in Table 21.

First, it should be noted that the main effect for "type of job" is not significant. Although the linear combinations of "clerical" and "sales" confidence judgments were not significantly different, the single MANOVA model provides a means for uncovering interactions involving "type of job." Second, as in the previous repeated measures MANOVA analyses, both predictor and criterion main effects were significant (p.<.001 and p.<.05, respectively). However, neither the predictor main effect nor the criterion main effect were substantial (R^2=.16 and .08, respectively).

Because the effect for type of predictor is contained in two significant second-order interactions, it is not interpretable. One of these higher-order interactions - predictor x job information x validity generalization information - suggests that average confidence for different types of predictors is not the same across the four experimental conditions. This effect can be observed by comparing columns #6-9 on Table 17. It appears that confidence is highest when job information is congruent and validity generalization information is available and that confidence decreases, as predicted, when job information is incongruent and/or validity generalization information is not available. These results are important because they demonstrate that task-related information affects judges' confidence, however the effect is realized across predictor judgments rather than individual validity judgments. Therefore, while Hypothesis #3 was not supported, some evidence does exist to support the conclusion that confidence in judgments is affected in some way by task-related job information.

The other second-order interaction involving type of predictor - job x predictor x validity generalization information - suggests that average confidence in clerical and sales validity judgments is not the same for different types of predictors when validity generalization information is present versus absent. Although this information is not specifically provided in any one of the statistical tables, it can be gleaned from comparing columns #4 and #5 on Table 18 (clerical) and Table 19 (sales).

This last three-way interaction provides information regarding how validity generalization information affects experts' confidence interval judgments. It appears that validity generalization results have a significant effect on judges' confidence in their validity judgments. However, the effect is not on individual confidence judgments but on average confidence judgments for different types of predictors. Because both "type of job" and "type of predictor" are within-subject factors, it appears that individual judges respond differentially to job and predictor type when making their confidence judgments. As a whole, judges in the "no vg" condition have slightly lower confidence than judges in the "vg" condition. However, this effect was only apparent for some of the types of predictors, primarily quantitative, perceptual, spatial/mechanical, and psychomotor confidence judgments in the "sales" job title. In the "sales" job title, confidence was greater for memory in the "vg" condition and greater for verbal in the "no vg" condition.

Table 1

Survey Responses by Division 14 Membership Category

Membership Category	Experimental Condition 1 N	%[a]	2 N	%	3 N	%	4 N	%	Total N
Fellow	11	(39.3)	12	(36.4)	7	(25.0)	12	(48.0)	42
Member	12	(42.9)	12	(36.4)	11	(39.3)	7	(28.0)	41
Associate	5	(17.9)	9	(27.3)	10	(35.7)	6	(24.0)	31
Total	28		33		28		25		114

[a] Numbers in parentheses indicate the percentage of individuals in the experimental condition from each membership category.

χ^2 = 4.80 (df=6) N.S.

Table 2

Educational Background of Expert Sample

Category	Experimental Condition 1 N	1 %[a]	2 N	2 %	3 N	3 %	4 N	4 %	Total N	Total %
N	28		33		28		25		114	
Level of Education										
Ph.D.	26	(92.9)	26	(78.8)	23	(82.1)	20	(80.0)	95	(83.3)
ABD	0	(0.0)	0	(0.0)	2	(7.1)	0	(0.0)	2	(1.8)
MA/MS/MBA	2	(7.1)	7	(21.2)	3	(10.7)	3	(12.0)	15	(13.1)
Ed.D.	0	(0.0)	0	(0.0)	0	(0.0)	2	(8.0)	2	(1.8)
Type of Degree										
I/O	13	(46.4)	14	(42.4)	15	(53.6)	8	(32.0)	50	(43.9)
General	6	(21.4)	11	(33.3)	7	(25.0)	6	(24.0)	30	(26.3)
Exper.	1	(3.6)	1	(3.0)	2	(7.1)	2	(8.0)	6	(5.3)
Social/Per	1	(3.6)	3	(9.1)	0	(0.0)	2	(8.0)	6	(5.3)
Counseling	2	(7.1)	1	(3.0)	0	(0.0)	3	(12.0)	6	(5.3)
Business	1	(3.6)	0	(0.0)	2	(7.1)	1	(4.0)	4	(3.5)
Quantitat.	2	(7.1)	0	(0.0)	0	(0.0)	0	(0.0)	2	(1.8)
App.Social	1	(3.6)	0	(0.0)	0	(0.0)	1	(4.0)	2	(1.8)
Gen. App.	1	(3.6)	0	(0.0)	1	(3.6)	0	(0.0)	2	(1.8)
Education.	0	(0.0)	2	(6.1)	0	(0.0)	0	(0.0)	2	(1.8)
App. Clin.	0	(0.0)	1	(3.0)	0	(0.0)	0	(0.0)	1	(.9)
Human Fac.	0	(0.0)	0	(0.0)	1	(3.6)	0	(0.0)	1	(.9)
Vocational	0	(0.0)	0	(0.0)	0	(0.0)	1	(4.0)	1	(.9)
Missing	0	(0.0)	0	(0.0)	0	(0.0)	1	(4.0)	1	(.9)
Average Years Post Degree										
$\bar{X}$		16.4		19.2		16.9		23.0		19.5
SD		11.7		11.2		14.5		15.7		13.2

[a] Numbers in parentheses indicate the percentage of individuals in each experimental condition at each level of education.

χ^2 tests could not be used to test the independence of these educational categories from experimental conditions because of the large number of empty cells.

Table 3

Work Background of Expert Sample

Current Affiliation/Job[a]	Experimental Condition 1	2	3	4	Total
N	28	33	28	25	114
Academic (Total)	(10)	(12)	(13)	(8)	(43)
I/O - Full Prof.	6	5	6	6	23
- Assoc. Prof.	1	1	3	0	5
- Asst. Prof.	3	1	1	0	5
O/B - Full Prof.	0	1	0	0	1
- Assoc. Prof.	0	0	1	0	1
- Asst. Prof.	0	0	0	2	2
Other/Not Described	0	4	2	0	6
Private Industry	(7)	(5)	(5)	(4)	(21)
Applied I/O Research	4	3	5	2	14
Internal Consulting	1	0	0	1	2
Management	2	1	0	1	4
Marketing Research	0	1	0	0	1
Public Sector	(4)	(4)	(2)	(1)	(11)
Applied I/O Research	4	4	2	1	11
Military	(2)	(1)	(2)	(2)	(7)
Applied I/O Research	2	1	2	2	7
Private Consulting	(5)	(8)	(6)	(8)	(27)
Applied I/O Research	4	5	6	3	18
Gen. Mgt. Consulting	1	2	0	4	7
Marketing Research	0	1	0	0	1
Counseling	0	0	0	1	1
Public Utility	(0)	(0)	(0)	(1)	(1)
Applied I/O Research	0	0	0	1	1
Research Institute	(0)	(3)	(0)	(0)	(3)
Applied I/O Research	0	3	0	0	3
Missing	(0)	(0)	(0)	(1)	(1)

a χ^2 tests could not be used to test the independence of these educational categories from experimental conditions because of the large number of empty cells.

Table 4

Means and Standard Deviations of Categories of Validity Judgments By Judgment Order

Category	(1) Predictor Order [a] 1	(2) Predictor Order [a] 2	(3) Job Order [b] 1	(4) Job Order [b] 2	(5) Predictor x Job Order 1,1	(6) Predictor x Job Order 1,2	(7) Predictor x Job Order 2,1	(8) Predictor x Job Order 2,2
N	56	58	56	58	25	31	31	27
Overall	.29 (10)[c]	.28 (10)	.29 (11)	.28 (09)	.31 (12)	.27 (08)	.28 (10)	.28 (10)
Type of Job								
Clerical	.31 (10)	.30 (11)	.31 (11)	.30 (10)	.32 (12)	.31 (08)	.30 (11)	.29 (11)
Sales	.26 (11)	.26 (10)	.27 (11)	.25 (10)	.29 (13)	.24 (09)	.26 (10)	.27 (11)
Type of Predictor								
GMA	.38 (11)	.40 (13)	.41 (13)	.37 (11)	.40 (13)	.37 (09)	.41 (13)	.38 (13)
Verbal	.38 (11)	.36 (11)	.38 (13)	.36 (09)	.39 (13)	.37 (09)	.37 (12)	.35 (09)
Quantitative	.36 (11)	.36 (12)	.37 (13)	.35 (10)	.39 (13)	.34 (10)	.36 (13)	.36 (10)
Perceptual	.29 (11)	.29 (12)	.29 (12)	.28 (10)	.31 (12)	.27 (09)	.28 (13)	.29 (10)
Memory	.30 (11)	.29 (11)	.30 (12)	.29 (10)	.32 (14)	.28 (09)	.28 (11)	.30 (11)
Sp/Mech	.16 (13)	.14 (12)	.15 (13)	.15 (12)	.18 (15)	.14 (11)	.13 (10)	.16 (13)
Psychomotor	.14 (13)	.14 (11)	.14 (13)	.14 (11)	.16 (17)	.13 (09)	.12 (09)	.15 (13)
Type of Criterion								
Ratings	.28 (10)	.25 (10)	.27 (11)	.26 (09)	.30 (12)	.26 (08)	.25 (10)	.26 (10)
Rankings	.29 (10)	.27 (12)	.28 (12)	.27 (10)	.30 (13)	.28 (08)	.27 (12)	.27 (11)
Work Sample	.33 (11)	.32 (10)	.33 (11)	.32 (09)	.34 (13)	.31 (10)	.32 (11)	.32 (10)
Production	.26 (10)	.28 (10)	.28 (10)	.26 (10)	.29 (12)	.24 (09)	.28 (09)	.28 (10)

[a] Order 1 = GMA, Verbal, Quantitative, Perceptual Speed, Memory, Spatial/Mechanical, Psychomotor
Order 2 = Psychomotor, Spatial/Mechanical, Memory, Perceptual Speed, Quantitative, Verbal, GMA

[b] Order 1 = Clerical first, sales second
Order 2 = Sales first, clerical second

[c] Standard deviations are in parentheses

Table 5

Effects of Order of Jobs and Order of Predictors on Validity Judgments for "Clerical" and "Sales" Job Titles

Effect	Df Num,Den	Wilks' Λ	F	Canonical R	R^2
		"Clerical" Job Title			
Predictor Order	28,83	.698	1.28	.54	.29
Job Order	28,83	.730	1.10	.52	.27
Predictor x Job	28,83	.771	.88	.48	.23
		"Sales" Job Title			
Predictor Order	28,82[a]	.620	1.79*	.62	.38
Job Order	28,82	.854	.50	.38	.15
Predictor x Job	28,82	.738	1.04	.51	.26

[a] One subject did not provide estimates for four of the "sales" job title validity judgments and was excluded from this analysis.

* p.<.05

Table 6

Means and Standard Deviations of Validity Judgments
For the "Clerical" and "Sales" Job Titles By Predictor Order

Type of Validity	"Clerical" Job Title 1[a]	"Clerical" Job Title 2[b]	"Sales" Job Title 1	"Sales" Job Title 2
N	56	58	56	58
General Mental Ability				
Ratings	.39 (14)	.39 (15)	.36 (14)	.37 (14)
Rankings	.41 (16)	.41 (17)	.36 (14)	.38 (14)
Work Samples	.45 (15)	.44 (16)	.41 (14)	.42 (15)
Productivity	.36 (13)	.38 (15)	.31 (14)	.37 (16)
Verbal				
Ratings	.38 (14)	.35 (15)	.37 (15)	.34 (13)
Rankings	.40 (15)	.37 (17)	.39 (13)	.36 (14)
Work Samples	.41 (15)	.40 (15)	.42 (16)	.38 (12)
Productivity	.32 (15)	.33 (16)	.35 (15)	.35 (13)
Quantitative				
Ratings	.38 (12)	.35 (14)	.30 (18)	.29 (15)
Rankings	.40 (14)	.37 (18)	.32 (17)	.30 (16)
Work Samples	.45 (14)	.44 (14)	.38 (16)	.38 (15)
Productivity	.38 (13)	.39 (13)	.29 (16)	.32 (15)
Perceptual Speed				
Ratings	.31 (14)	.27 (15)	.24 (17)	.23 (17)
Rankings	.31 (13)	.28 (16)	.27 (17)	.26 (18)
Work Samples	.38 (17)	.37 (18)	.27 (18)	.29 (18)
Productivity	.32 (14)	.33 (15)	.22 (20)	.26 (18)
Memory				
Ratings	.32 (13)	.28 (13)	.26 (13)	.24 (12)
Rankings	.31 (13)	.28 (14)	.27 (13)	.26 (13)
Work Samples	.37 (15)	.35 (14)	.31 (14)	.31 (12)
Productivity	.31 (13)	.32 (14)	.25 (15)	.28 (13)
Spatial/Mechanical				
Ratings	.16 (15)	.12 (13)	.14 (14)	.11 (13)
Rankings	.17 (16)	.15 (15)	.15 (14)	.12 (13)
Work Samples	.21 (20)	.20 (15)	.16 (15)	.15 (13)
Productivity	.15 (16)	.16 (14)	.13 (15)	.12 (12)
Psychomotor				
Ratings	.16 (15)	.12 (11)	.10 (12)	.09 (13)
Rankings	.16 (14)	.12 (12)	.10 (12)	.11 (14)
Work Samples	.23 (20)	.21 (17)	.12 (14)	.15 (15)
Productivity	.19 (17)	.17 (13)	.10 (14)	.12 (15)

[a] GMA, Verbal, Quantitative, Perceptual Speed, Memory, Spatial/Mechanical, and Psychomotor

[b] Psychomotor, Spatial/Mechanical, Memory, Perceptual Speed, Quantitative, Verbal, and GMA

Table 7

Order (Between-Subjects) Effects and Within-Subjects Effects On Validity Judgments for "Clerical" and "Sales" Job Titles

Effect	Df Num,Den	Wilks' Λ	F	Canonical R	R^2
		"Clerical" Job Title			
Predictor (TofP)	6,105	.228	59.13***	.88	.77
TofP x Predictor Order (PO)	6,105	.979	.37	.14	.02
TofP x Job Order (JO)	6,105	.956	.80	.21	.04
TofP x PO x JO	6,105	.965	.63	.19	.03
Criterion (TofC)	3,108	.539	30.78***	.68	.46
TofC x PO	3,108	.921	3.10*	.28	.08
TofC x JO	3,108	.990	.35	.10	.01
TofC x PO x JO	3,108	.984	.57	.13	.02
TofP x TofC	18,93	.485	5.49***	.72	.52
TofP x TofC x PO	18,93	.809	1.22	.44	.19
TofP x TofC x JO	18,93	.789	1.31	.45	.20
TofP x TofC x PO x JO	18,93	.847	.94	.39	.15
		"Sales" Job Title			
Predictor (TofP)	6,104[a]	.172	83.47***	.89	.83
TofP x Predictor Order (PO)	6,104	.902	1.89	.31	.10
TofP x Job Order (JO)	6,104	.975	.44	.16	.02
TofP x PO x JO	6,104	.955	.82	.21	.05
Criterion (TofC)	3,107	.539	30.38***	.68	.46
TofC x PO	3,107	.904	3.79*	.31	.10
TofC x JO	3,107	.995	.19	.07	.01
TofC x PO x JO	3,107	.999	.03	.03	.00
TofP x TofC	18,92	.523	4.66***	.69	.48
TofP x TofC x PO	18,92	.785	1.40	.46	.22
TofP x TofC x JO	18,92	.914	.48	.29	.09
TofP x TofC x PO x JO	18,92	.830	1.05	.41	.17

[a] One subject was eliminated from the analysis because four sales SKAPs validity judgments were missing.

* $p < .05$
** $p < .001$
*** $p < .0001$

Table 8

Means and Standard Deviations Across Different Categories of Validity Judgments

Category	(1) Overall	(2) Job Information Congruent	(3) Job Information Incongruent	(4) VG Information Absent	(5) VG Information Present	(6) Experimental Condition 1	(7) Experimental Condition 2	(8) Experimental Condition 3	(9) Experimental Condition 4
N	114	56	58	61	53	28	33	28	25
Overall	.28 (10)[a]	.27 (10)	.30 (09)	.27 (09)	.30 (10)	.25 (09)	.29 (10)	.30 (10)	.31 (09)
Type of Job									
Clerical	.31 (10)	.31 (11)	.31 (10)	.28 (10)	.34 (10)	.26 (10)	.29 (10)	.35 (11)	.33 (08)
Sales	.26 (11)	.24 (10)	.28 (11)	.26 (09)	.26 (12)	.23 (08)	.29 (09)	.25 (12)	.28 (13)
Type of Predictor									
GMA	.39 (12)	.38 (11)	.39 (12)	.37 (13)	.41 (10)	.36 (12)	.38 (14)	.41 (11)	.42 (10)
Verbal	.37 (11)	.35 (11)	.39 (11)	.35 (11)	.39 (11)	.32 (11)	.37 (11)	.38 (10)	.41 (12)
Quantitative	.36 (12)	.34 (10)	.38 (12)	.35 (12)	.37 (11)	.32 (10)	.37 (12)	.36 (10)	.39 (12)
Perceptual	.29 (11)	.27 (11)	.30 (11)	.28 (11)	.30 (11)	.25 (11)	.30 (11)	.29 (10)	.30 (12)
Memory	.30 (11)	.27 (10)	.32 (12)	.28 (11)	.32 (11)	.24 (08)	.31 (12)	.30 (10)	.33 (12)
Sp/Mech	.15 (12)	.15 (13)	.15 (11)	.14 (12)	.16 (12)	.13 (12)	.14 (12)	.17 (14)	.15 (10)
Psychomotor	.14 (12)	.14 (12)	.14 (13)	.12 (11)	.16 (14)	.10 (10)	.14 (11)	.17 (13)	.14 (14)
Type of Criterion									
Ratings	.27 (10)	.25 (10)	.28 (10)	.26 (10)	.28 (10)	.23 (10)	.27 (10)	.27 (10)	.29 (10)
Rankings	.28 (11)	.26 (12)	.29 (10)	.25 (11)	.31 (10)	.23 (11)	.27 (11)	.30 (12)	.31 (09)
Work Sample	.32 (10)	.32 (11)	.33 (10)	.30 (10)	.36 (10)	.28 (10)	.31 (10)	.36 (11)	.35 (09)
Production	.27 (10)	.26 (09)	.28 (11)	.27 (10)	.27 (11)	.25 (07)	.29 (11)	.27 (11)	.27 (11)

[a] Standard deviations are in parentheses

Table 9

Means and Standard Deviations of Validity Judgments
for the "Clerical" Job Title

Type of Validity	(1) Overall	(2) Job Information Congruent	(3) Job Information Incongruent	(4) VG Information Absent	(5) VG Information Present	(6) Experimental Condition 1	(7) Experimental Condition 2	(8) Experimental Condition 3	(9) Experimental Condition 4
N	114	56	58	61	33	28	33	28	25
GMA									
Ratings	.39(15)	.37(15)	.42(14)	.36(15)	.43(13)	.32(15)	.39(14)	.41(14)	.46(12)
Rankings	.41(17)	.39(18)	.43(15)	.35(15)	.48(16)	.32(15)	.38(15)	.47(18)	.49(12)
Work Sample	.44(16)	.43(17)	.45(14)	.38(15)	.51(13)	.35(16)	.41(14)	.51(14)	.52(11)
Production	.37(14)	.36(13)	.39(15)	.36(15)	.39(13)	.34(12)	.37(16)	.37(13)	.40(13)
Verbal									
Ratings	.37(14)	.31(12)	.42(14)	.35(15)	.38(14)	.28(12)	.42(14)	.34(12)	.43(14)
Rankings	.39(16)	.34(16)	.43(14)	.35(17)	.43(14)	.28(16)	.41(16)	.40(15)	.47(12)
Work Sample	.40(15)	.35(16)	.45(13)	.36(15)	.44(14)	.29(14)	.43(14)	.42(15)	.47(12)
Production	.32(15)	.27(13)	.38(15)	.32(16)	.32(15)	.24(13)	.39(15)	.29(14)	.35(15)
Quantitative									
Ratings	.36(13)	.37(13)	.36(13)	.34(14)	.39(11)	.34(13)	.34(15)	.39(12)	.39(11)
Rankings	.39(16)	.39(18)	.38(15)	.33(15)	.45(16)	.32(14)	.34(16)	.46(19)	.43(12)
Work Sample	.45(14)	.47(14)	.43(14)	.42(14)	.49(13)	.42(14)	.41(14)	.51(13)	.46(13)
Production	.39(13)	.40(12)	.37(14)	.38(12)	.39(14)	.39(09)	.37(15)	.40(15)	.38(13)
Perceptual									
Ratings	.29(14)	.33(14)	.25(13)	.26(15)	.32(13)	.31(16)	.22(13)	.35(12)	.29(13)
Rankings	.30(15)	.33(15)	.26(14)	.25(15)	.35(14)	.29(16)	.23(14)	.38(14)	.31(13)
Work Sample	.38(17)	.44(17)	.31(16)	.33(16)	.43(18)	.38(16)	.28(13)	.50(15)	.35(18)
Production	.33(15)	.37(14)	.28(14)	.31(15)	.35(14)	.35(15)	.27(14)	.39(13)	.30(14)
Memory									
Ratings	.30(13)	.28(11)	.32(14)	.27(13)	.33(12)	.25(11)	.29(15)	.31(11)	.35(13)
Rankings	.30(13)	.28(11)	.32(14)	.27(13)	.33(12)	.25(11)	.29(15)	.31(11)	.35(13)
Work Sample	.36(14)	.36(15)	.37(14)	.31(14)	.42(13)	.28(11)	.34(15)	.44(14)	.41(12)
Production	.32(14)	.29(12)	.34(15)	.30(14)	.34(12)	.25(10)	.33(16)	.34(11)	.35(14)
Spatial/Mechanical									
Ratings	.14(14)	.15(15)	.13(13)	.12(13)	.17(15)	.12(14)	.12(12)	.19(16)	.16(13)
Rankings	.16(16)	.17(17)	.15(14)	.13(14)	.19(16)	.12(15)	.13(14)	.21(18)	.17(14)
Work Sample	.20(17)	.22(18)	.18(16)	.17(16)	.24(19)	.18(16)	.17(16)	.27(20)	.20(17)
Production	.15(15)	.17(15)	.14(14)	.15(15)	.16(15)	.15(14)	.16(16)	.19(16)	.13(13)
Psychomotor									
Ratings	.14(13)	.15(14)	.13(13)	.10(11)	.18(15)	.09(11)	.12(10)	.21(15)	.15(15)
Rankings	.14(13)	.15(14)	.13(12)	.11(11)	.18(14)	.10(13)	.11(10)	.20(14)	.15(14)
Work Sample	.22(18)	.24(19)	.20(17)	.17(14)	.28(21)	.17(14)	.17(14)	.32(20)	.23(20)
Production	.18(15)	.19(14)	.17(16)	.16(14)	.20(16)	.15(12)	.17(15)	.22(16)	.17(17)

[a] Standard Deviations are in parentheses

Table 10

Means and Standard Deviations of Validity Judgments for the "Sales" Job Title

Type of Validity	(1) Overall	(2) Job Information Congruent	(3) Job Information Incongruent	(4) VG Information Absent	(5) VG Information Present	(6) Experimental Condition 1	(7) Experimental Condition 2	(8) Experimental Condition 3	(9) Experimental Condition 4
N	114	56	58	61	53	28	33	28	25
GMA									
Ratings	.36(14)[a]	.36(13)	.36(14)	.37(14)	.36(13)	.38(13)	.37(15)	.35(12)	.36(14)
Rankings	.37(14)	.38(14)	.37(14)	.37(14)	.37(14)	.37(13)	.37(15)	.38(15)	.36(13)
Work Samples	.42(15)	.43(14)	.41(15)	.39(16)	.45(12)	.40(15)	.38(17)	.46(13)	.44(11)
Production	.34(16)	.34(15)	.34(17)	.36(15)	.32(16)	.37(12)	.35(18)	.32(17)	.32(16)
Verbal Ability									
Ratings	.36(14)	.35(13)	.35(15)	.36(14)	.35(14)	.38(15)	.35(14)	.34(12)	.35(17)
Rankings	.38(14)	.39(14)	.36(13)	.35(15)	.40(13)	.37(16)	.34(13)	.41(12)	.39(13)
Work Samples	.40(14)	.43(14)	.38(14)	.36(13)	.46(14)	.40(13)	.33(12)	.47(14)	.45(14)
Production	.35(14)	.38(11)	.32(16)	.33(15)	.37(13)	.36(10)	.31(17)	.40(12)	.34(15)
Quantitative Ability									
Ratings	.30(17)	.24(14)	.35(18)	.31(14)	.28(19)	.26(12)	.35(14)	.22(15)	.34(22)
Rankings	.31(16)	.25(15)	.37(16)	.31(14)	.31(18)	.25(13)	.36(13)	.26(17)	.37(18)
Work Samples	.38(15)	.33(16)	.42(14)	.36(15)	.40(16)	.30(15)	.41(13)	.37(16)	.43(15)
Production	.31(16)	.26(13)	.35(17)	.32(14)	.28(18)	.27(10)	.37(15)	.26(16)	.32(19)
Perceptual Speed									
Ratings	.24(17)	.16(13)	.31(18)	.25(16)	.21(19)	.17(12)	.32(15)	.15(14)	.28(21)
Rankings	.26(17)	.19(15)	.33(17)	.25(17)	.27(18)	.16(13)	.33(15)	.22(16)	.33(18)
Work Samples	.28(18)	.20(15)	.36(18)	.31(17)	.25(20)	.20(12)	.40(15)	.19(18)	.31(20)
Production	.24(19)	.17(15)	.31(20)	.28(16)	.19(21)	.18(11)	.37(15)	.16(18)	.24(24)
Memory									
Ratings	.25(12)	.22(10)	.28(13)	.25(12)	.26(13)	.21(10)	.28(12)	.23(11)	.28(15)
Rankings	.26(13)	.23(13)	.29(14)	.25(13)	.28(14)	.21(12)	.29(13)	.26(13)	.30(15)
Work Sample	.31(13)	.28(12)	.33(14)	.29(12)	.33(14)	.26(10)	.32(13)	.31(14)	.35(15)
Production	.26(14)	.23(11)	.30(16)	.28(13)	.25(16)	.23(09)	.31(14)	.22(13)	.27(18)
Spatial/Mechanical									
Ratings	.13(13)	.12(12)	.13(15)	.12(13)	.13(14)	.11(10)	.13(15)	.13(14)	.13(15)
Rankings	.13(13)	.12(13)	.14(14)	.12(13)	.14(14)	.11(11)	.13(14)	.14(14)	.15(14)
Work Samples	.16(14)	.15(15)	.16(14)	.15(14)	.16(14)	.15(15)	.15(14)	.16(14)	.16(14)
Production	.13(13)	.11(13)	.14(14)	.13(12)	.12(15)	.12(12)	.14(12)	.11(14)	.13(16)
Psychomotor									
Ratings	.10(12)	.07(11)	.12(13)	.09(11)	.10(14)	.06(08)	.12(12)	.09(13)	.11(15)
Rankings	.10(13)	.09(12)	.12(14)	.10(12)	.11(14)	.07(10)	.13(12)	.10(14)	.12(15)
Work Samples	.14(15)	.12(14)	.15(16)	.14(15)	.13(15)	.09(12)	.17(16)	.14(15)	.13(16)
Production	.11(15)	.09(12)	.14(17)	.12(14)	.11(16)	.07(09)	.16(16)	.10(15)	.11(18)

[a] Standard Deviations are in parantheses

Table 11

Effects of Job and Validity Generalization Information on Validity Judgments for Clerical and Sales SKAPs

Effect	Df Num,Den	Wilk's Λ	F	Canonical R	R^2
		Clerical SKAPs			
VG Information	28,83	.577	2.17**	.65	.42
Job Information	28,83	.578	2.17**	.65	.42
VG x Job	28,83	.633	1.98**	.63	.40
		Sales SKAPs			
VG Information	28,82[a]	.647	1.60*	.59	.35
Job Information	28,82	.477	3.21***	.72	.52
VG x Job	28,82	.638	1.66**	.60	.36

[a] One subject was eliminated from the analysis because four sales SKAPs validity judgments were missing.

* $p < .05$
** $p < .001$
*** $p < .0001$

Table 12

Effects of Job and Validity Generalization Information on Validity Judgments for "Clerical" and "Sales" Job Titles

Effect	Df Num,Den	Wilk's Λ	F	Canonical R	R^2
		"Clerical" Job Title			
VG Information	28,83	.588	2.07**	.64	.41
Job Information	28,83	.532	2.61***	.68	.47
VG x Job	28,83	.743	1.03	.51	.26
		"Sales" Job Title			
VG Information	28,82[a]	.623	1.77*	.61	.38
Job Information	28,82	.481	3.16***	.72	.52
VG x Job	28,82	.710	1.20	.54	.29

[a] One subject was eliminated from the analysis because four sales SKAPs validity judgments were missing.

* p.<.05
** p.<.001
*** p.<.0001

Table 13

Effects of Job and Validity Generalization Information on Experts' Validity Judgments for Clerical and Sales SKAPs

Validity Predictor	Criterion	Effect	Clerical SKAPs Type IV SS	Clerical SKAPs F	Sales SKAPs Type IV SS	Sales SKAPs F
GMA	Ratings	VG Info	.037	1.79	.013	.80
		Job Info	.000	.01	.079	4.71*
		VG x Job	.065	3.14	.079	4.73*
		error	2.284		1.816	
	Rankings	VG Info	.141	5.79*	.083	4.26*
		Job Info	.024	1.00	.088	4.53*
		VG x Job	.161	6.62**	.075	3.87*
		error	2.674		2.125	
	Work Samples	VG Info	.316	14.23***	.179	9.61**
		Job Info	.016	.71	.026	1.42
		VG x Job	.072	3.26	.024	1.31
		error	2.442		2.031	
	Production	VG Info	.000	.00	.002	.11
		Job Info	.014	.64	.054	2.44
		VG x Job	.025	1.11	.038	1.72
		error	2.464		2.409	
Verbal	Ratings	VG Info	.025	1.36	.009	.47
		Job Info	.056	3.04	.094	5.03*
		VG x Job	.021	1.12	.026	1.39
		error	2.029		2.037	
	Rankings	VG Info	.234	11.50***	.062	3.15
		Job Info	.019	.92	.051	2.57
		VG x Job	.030	1.47	.007	.35
		error	2.237		2.145	
	Work Samples	VG Info	.453	23.96***	.072	4.24*
		Job Info	.029	1.51	.005	.33
		VG x Job	.001	.04	.004	.25
		error	2.082		1.841	
	Production	VG Info	.040	1.83	.000	.00
		Job Info	.105	4.83*	.002	.10
		VG x Job	.001	.06	.042	2.38
		error	2.401		1.930	
Quanti-tative	Ratings	VG Info	.012	.52	.001	.03
		Job Info	.009	.37	.404	22.12***
		VG x Job	.026	1.12	.070	3.85*
		error	2.600		1.989	

Table 13 (continued)

	Rankings	VG Info	.143	5.51*	.059	2.77
		Job Info	.018	.70	.499	23.37***
		VG x Job	.109	4.18*	.055	2.57
		error	2.856		2.327	
	Work Samples	VG Info	.075	4.15*	.088	4.16
		Job Info	.059	3.27	.268	12.67***
		VG x Job	.031	1.73	.000	.01
		error	1.997		2.310	
	Production	VG Info	.031	.62	.000	.01
		Job Info	.083	3.86*	.366	19.39***
		VG x Job	.034	1.60	.005	.24
		error	2.363		2.054	
Perceptual Speed	Ratings	VG Info	.000	.00	.017	1.04
		Job Info	.023	.88	.235	14.15***
		VG x Job	.050	1.91	.056	3.39
		error	2.859		1.810	
	Rankings	VG Info	.061	2.45	.150	7.67**
		Job Info	.001	.03	.160	8.22**
		VG x Job	.066	2.66	.004	.20
		error	2.723		2.128	
	Work Samples	VG Info	.002	.08	.024	1.03
		Job Info	.217	7.80**	.398	16.95***
		VG x Job	.309	11.10***	.044	1.86
		error	3.061		2.556	
	Production	VG Info	.061	2.13	.000	.00
		Job Info	.129	4.54*	.374	18.24***
		VG x Job	.184	6.45**	.023	1.11
		error	3.138		2.237	
Memory	Ratings	VG Info	.033	2.22	.037	2.36
		Job Info	.000	.00	.267	17.22***
		VG x Job	.026	1.75	.009	.57
		error	1.660		1.689	
	Rankings	VG Info	.095	6.13**	.069	3.81*
		Job Info	.007	.47	.233	12.91***
		VG x Job	.062	3.98*	.000	.01
		error	1.699		1.970	
	Work Samples	VG Info	.234	13.46***	.089	5.30*
		Job Info	.013	.73	.233	13.85***
		VG x Job	.131	7.53**	.004	.22
		error	1.916		1.833	
	Production	VG Info	.016	.84	.001	.06
		Job Info	.000	.01	.379	21.29***
		VG x Job	.110	5.91*	.004	.25
		error	2.051		1.938	

Table 13 (continued)

Spatial-Mechanical	Ratings	VG Info	.028	1.26	.022	1.45
		Job Info	.012	.52	.007	.48
		VG x Job	.029	1.33	.005	.31
		error	2.424		1.681	
	Rankings	VG Info	.073	3.18	.030	1.61
		Job Info	.020	.85	.024	1.30
		VG x Job	.037	1.60	.000	.00
		error	2.516		2.027	
	Work Samples	VG Info	.070	2.71	.010	.41
		Job Info	.116	4.49*	.026	1.06
		VG x Job	.045	1.72	.003	.13
		error	2.852		2.670	
	Production	VG Info	.007	.35	.009	.50
		Job Info	.022	1.06	.026	1.35
		VG x Job	.014	.65	.005	.27
		error	2.323		2.076	
Psycho-motor	Ratings	VG Info	.074	4.24*	.023	1.62
		Job Info	.029	1.68	.095	6.67**
		VG x Job	.127	7.25**	.000	.00
		error	1.921		1.560	
	Rankings	VG Info	.061	3.33	.034	2.32
		Job Info	.027	1.44	.059	4.07
		VG x Job	.090	4.88*	.000	.02
		error	2.032		1.589	
	Work Samples	VG Info	.077	2.79	.081	3.32
		Job Info	.242	8.77**	.196	8.06**
		VG x Job	.280	101.13**	.001	.05
		error	3.038		2.649	
	Production	VG Info	.006	.26	.014	.72
		Job Info	.082	3.44	.194	9.80**
		VG x Job	.109	4.58*	.006	.30
		error	2.613		2.169	

*
** p.<.05
***p.<.001
p.<.0001

Table 14

Between- and Within-Subjects Effects on Validity Judgments for Clerical and Sales SKAPs

Effect	Df Num,Den	Wilks' Λ	F	Canonical R	R^2
		Clerical SKAPs			
Predictor (TofP)	6,105	.222	61.24***	.88	.77
TofP x VG Information (VG)	6,105	.926	1.39	.27	.07
TofP x Job Information (Job)	6,105	.790	4.65***	.46	.21
TofP x VG x Job	6,105	.910	1.73	.30	.09
Criterion (TofC)	3,108	.479	39.11***	.72	.52
TofC x VG	3,108	.813	8.30***	.43	.18
TofC x Job	3,108	.926	2.89*	.27	.07
TofC x VG x Job	3,108	.979	.78	.15	.02
TofP x TofC	18,93	.501	5.15***	.71	.50
TofP x TofC x VG	18,93	.657	2.69***	.59	.35
TofP x TofC x Job	18,93	.742	1.80*	.51	.26
TofP x TofC x VG x Job	18,93	.725	1.96*	.52	.27
		Sales SKAPs			
Predictor (TofP)	6,104[a]	.186	75.84***	.90	.81
TofP x VG Information (VG)	6,104	.955	.81	.21	.04
TofP x Job Information (Job)	6,104	.641	9.71***	.60	.36
TofP x VG x Job	6,104	.919	1.52	.28	.08
Criterion (TofC)	3,107	.517	33.29***	.69	.48
TofC x VG	3,107	.841	6.75***	.40	.16
TofC x Job	3,107	.998	.05	.04	.00
TofC x VG x Job	3,107	.954	1.73	.22	.05
TofP x TofC	18,92	.564	3.96***	.66	.44
TofP x TofC x VG	18,92	.722	1.97*	.53	.28
TofP x TofC x Job	18,92	.724	1.95*	.53	.28
TofP x TofC x VG x Job	18,92	.773	1.50	.48	.23

[a] One subject was eliminated from the analysis because four sales SKAPs validity judgments were missing.

* $p < .05$
** $p < .001$
*** $p < .0001$

Table 15

Effects of Job and Validity Generalization Information on Validity Judgments for "Clerical" and "Sales" Job Titles

Validity Predictor/Criterion		Effect	"Clerical" Job Title Type IV SS	F	"Sales" Job Title Type IV SS	F
GMA	Ratings	VG Info	.081	4.29**	.013	.71
		Job Info	.159	8.45**	.002	.09
		VG x Job	.000	.00	.003	.15
		error	2.107		1.993	
	Rankings	VG Info	.428	17.93**	.000	.00
		Job Info	.043	1.81	.006	.31
		VG x Job	.008	.35	.000	.02
		error	2.620		2.176	
	Work Samples	VG Info	.473	23.60**	.078	3.77*
		Job Info	.018	.90	.013	.61
		VG x Job	.012	.59	.000	.01
		error	2.212		2.262	
	Production	VG Info	.019	.96	.042	1.68
		Job Info	.021	1.08	.002	.07
		VG x Job	.000	.00	.002	.06
		error	2..16		2.756	
Verbal	Ratings	VG Info	.03[illegible]	1.71	.014	.73
		Job Info	.354	20.32**	.044	.21
		VG x Job	.011	.63	.018	.92
		error	1.914		2.153	
	Rankings	VG Info	.235	10.81**	.056	3.07
		Job Info	.272	12.50**	.028	1.53
		VG x Job	.024	1.08	.005	.28
		error	2.378		2.004	
	Work Samples	VG Info	.201	10.70**	.237	13.76**
		Job Info	.243	12.94**	.064	3.72
		VG x Job	.060	3.20	.024	1.39
		error	2.047		1.877	
	Production	VG Info	.000	.02	.035	1.81
		Job Info	.330	16.23**	.085	4.36*
		VG x Job	.048	2.35	.001	.03
		error	2.214		2.116	
Quantitative	Ratings	VG Info	.066	4.08*	.022	.85
		Job Info	.001	.04	.295	11.63**
		VG x Job	.001	.06	.009	.36
		error	1.821		2.769	

Table 15 (continued)

	Rankings	VG Info	.330	13.94**	.001	.03
		Job Info	.001	.04	.344	14.57**
		VG x Job	.009	.40	.000	.02
		error	2.610		2.573	
	Work Samples	VG Info	.122	6.79**	.042	1.97
		Job Info	.036	1.97	.201	9.44**
		VG x Job	.005	.25	.012	.56
		error	1.990		2.317	
	Production	VG Info	.004	.22	.031	1.37
		Job Info	.018	1.01	.192	8.36**
		VG x Job	.000	.02	.015	.64
		error	1.914		2.763	
Perceptual Speed	Ratings	VG Info	.072	4.14*	.026	1.05
		Job Info	.198	11.42**	.561	22.84**
		VG x Job	.010	.58	.003	.12
		error	1.993		2.676	
	Rankings	VG Info	.120	10.38**	.025	1.01
		Job Info	.141	7.36**	.524	21.37**
		VG x Job	.000	.00	.028	1.15
		error	2.181		2.670	
	Work Samples	VG Info	.202	8.52**	.079	2.94
		Job Info	.468	19.70**	.671	25.06**
		VG x Job	.007	.31	.052	1.93
		error	2.699		2.919	
	Production	VG Info	.020	1.07	.166	5.61*
		Job Info	.226	11.94**	.490	16.57**
		VG x Job	.000	.02	.070	2.38
		error	2.149		3.225	
Memory	Ratings	VG Info	.090	5.70*	.003	.24
		Job Info	.036	2.31	.02	6.89**
		VG x Job	.000	.01	.001	.10
		error	1.741		1.609	
	Rankings	VG Info	.157	9.53**	.028	1.61
		Job Info	.062	3.76	.098	5.72*
		VG x Job	.021	1.26	.011	.66
		error	1.803		1.866	
	Work Samples	VG Info	.352	20.34**	.032	1.89
		Job Info	.009	.53	.071	4.20*
		VG x Job	.055	3.16	.003	.20
		error	1.893		1.855	
	Production	VG Info	.074	4.34*	.014	.74
		Job Info	.073	4.26*	.124	6.38**
		VG x Job	.030	1.76	.007	.38
		error	1.864		2.125	

Table 15 (continued)

Spatial-Mechanical	Ratings	VG Info	.076	3.94*	.001	.08
		Job Info	.010	.54	.006	.34
		VG x Job	.003	.17	.002	.10
		error	2.090		2.014	
	Rankings	VG Info	.101	4.27*	.015	.82
		Job Info	.007	.29	.009	.53
		VG x Job	.019	.78	.002	.10
		error	2.584		1.960	
	Work Samples	VG Info	.101	3.37	.002	.11
		Job Info	.036	1.21	.000	.01
		VG x Job	.026	.87	.000	.00
		error	3.257		2.264	
	Production	VG Info	.000	.02	.001	.04
		Job Info	.016	.71	.020	1.10
		VG x Job	.036	1.65	.000	.00
		error	2.413		1.986	
Psycho-motor	Ratings	VG Info	.155	9.29**	.001	.08
		Job Info	.006	.35	.048	3.18
		VG x Job	.058	3.48	.014	.90
		error	1.825		1.655	
	Rankings	VG Info	.144	8.81**	.003	.19
		Job Info	.009	.58	.034	2.01
		VG x Job	.032	1.94	.012	.71
		error	1.786		1.835	
	Work Samples	VG Info	.315	10.52**	.000	.00
		Job Info	.049	1.62	.030	1.37
		VG x Job	.059	1.98	.063	2.83
		error	3.269		2.418	
	Production	VG Info	.053	2.34	.001	.03
		Job Info	.007	.29	.059	2.78
		VG x Job	.035	1.55	.050	2.34
		error	2.457		2.326	

* p.<.05
** p.<.001
*** p.<.0001

Table 16

Between- and Within-Subjects Effects on Validity Judgments for "Clerical" and "Sales" Job Titles

Effect	Df Num,Den	Wilks' Λ	F	Canonical R	R^2
		"Clerical" Job Title			
Predictor (TofP)	6,105	.224	60.48***	.88	.77
TofP x VG Information (VG)	6,105	.882	2.34*	.34	.12
TofP x Job Information (Job)	6,105	.646	9.61***	.60	.36
TofP x VG x Job	6,105	.961	.70	.20	.04
Criterion (TofC)	3,108	.493	37.05***	.71	.50
TofC x VG	3,108	.812	8.34**	.43	.18
TofC x Job	3,108	.959	1.55	.20	.04
TofC x VG x Job	3,108	.974	.96	.16	.03
TofP x TofC	18,93	.496	5.25**	.71	.50
TofP x TofC x VG	18,93	.743	1.79*	.51	.26
TofP x TofC x Job	18,93	.831	1.05	.41	.17
TofP x TofC x VG x Job	18,93	.810	1.21	.44	.19
		"Sales" Job Title			
Predictor (TofP)	6,104[a]	.167	86.71***	.91	.83
TofP x VG Information (VG)	6,104	.907	1.77	.30	.09
TofP x Job Information (Job)	6,104	.614	10.89**	.62	.38
TofP x VG x Job	6,104	.931	1.29	.26	.07
Criterion (TofC)	3,107	.500	35.65***	.71	.50
TofC x VG	3,107	.835	7.06**	.41	.17
TofC x Job	3,107	.991	.33	.10	.01
TofC x VG x Job	3,107	.973	.99	.16	.03
TofP x TofC	18,92	.516	4.79***	.70	.49
TofP x TofC x VG	18,92	.677	2.44**	.57	.32
TofP x TofC x Job	18,92	.781	1.43	.46	21
TofP x TofC x VG x Job	18,92	.773	1.50	.48	.23

[a] One subject was eliminated from the analysis because four sales SKAPs validity judgments were missing.

* p.<.05
** p.<.001
*** p.<.0001

Table 17

Means and Standard Deviations of Confidence Interval Judgments Across Different Categories of Validity Judgments

Category	(1) Overall	(2) Job Information Congruent	(3) Job Information Incongruent	(4) VG Information Absent	(5) VG Information Present	(6) Experimental Condition 1	(7) Experimental Condition 2	(8) Experimental Condition 3	(9) Experimental Condition 4
N	114	56	58	61	53	28	33	28	25
Overall	.22 (07) [a]	.21 (06)	.23 (08)	.22 (08)	.22 (07)	.21 (07)	.23 (08)	.20 (06)	.24 (08)
Type of Job									
Clerical	.22 (08)	.21 (07)	.24 (08)	.22 (08)	.22 (08)	.21 (07)	.24 (08)	.21 (06)	.24 (08)
Sales	.22 (07)	.21 (06)	.23 (08)	.22 (08)	.21 (07)	.22 (07)	.23 (09)	.20 (06)	.23 (08)
Type of Predictor									
GMA	.23 (08)	.22 (08)	.25 (09)	.23 (09)	.23 (08)	.22 (09)	.24 (08)	.21 (06)	.25 (09)
Verbal	.22 (09)	.21 (08)	.23 (10)	.21 (08)	.23 (10)	.20 (07)	.22 (09)	.22 (08)	.24 (11)
Quantitative	.22 (08)	.21 (07)	.24 (09)	.22 (08)	.22 (08)	.21 (07)	.23 (09)	.21 (07)	.24 (09)
Perceptual	.22 (08)	.21 (08)	.24 (09)	.23 (09)	.22 (08)	.20 (07)	.24 (09)	.21 (08)	.23 (09)
Memory	.22 (08)	.21 (07)	.24 (08)	.23 (08)	.21 (07)	.22 (08)	.24 (09)	.20 (06)	.24 (07)
Sp/Mech	.22 (09)	.22 (09)	.23 (09)	.23 (09)	.22 (08)	.23 (09)	.23 (09)	.20 (07)	.24 (09)
Psychomotor	.23 (08)	.21 (07)	.24 (09)	.22 (08)	.23 (09)	.21 (08)	.24 (09)	.21 (07)	.24 (10)
Type of Criterion									
Ratings	.23 (08)	.21 (07)	.24 (09)	.23 (08)	.23 (08)	.21 (07)	.24 (09)	.21 (07)	.24 (09)
Rankings	.23 (08)	.21 (07)	.24 (09)	.23 (09)	.22 (08)	.22 (08)	.24 (09)	.20 (06)	.25 (09)
Work Sample	.22 (08)	.20 (07)	.23 (08)	.22 (08)	.22 (08)	.20 (07)	.23 (08)	.20 (07)	.23 (08)
Production	.22 (07)	.22 (07)	.23 (08)	.22 (07)	.22 (07)	.22 (07)	.23 (08)	.22 (07)	.24 (08)

[a] Standard Deviations are in parentheses

Table 18

Means and Standard Deviations of Confidence Interval Judgments for the "Clerical" Job Title

Type of Validity	(1) Overall	(2) Job Information Congruent	(3) Job Information Incongruent	(4) VG Information Absent	(5) VG Information Present	(6) Experimental Condition 1	(7) Experimental Condition 2	(8) Experimental Condition 3	(9) Experimental Condition 4
N	114	56	58	61	53	28	33	28	25
GMA									
Ratings	.24(13)[a]	.23(15)	.25(10)	.24(15)	.24(10)	.24(09)	.24(10)	.23(10)	.26(10)
Rankings	.23(09)	.21(07)	.25(09)	.24(09)	.23(08)	.23(08)	.24(09)	.20(06)	.27(09)
Work Samples	.22(09)	.20(07)	.24(10)	.23(09)	.21(08)	.21(08)	.24(10)	.20(07)	.24(09)
Production	.23(08)	.22(08)	.24(09)	.23(08)	.23(09)	.22(08)	.24(08)	.22(07)	.25(10)
Verbal Ability									
Ratings	.22(10)	.21(09)	.24(11)	.21(09)	.24(11)	.19(07)	.23(10)	.22(11)	.25(12)
Rankings	.22(10)	.20(08)	.24(11)	.21(09)	.23(10)	.20(08)	.23(10)	.21(08)	.25(12)
Work Samples	.21(09)	.20(08)	.22(10)	.21(09)	.22(09)	.19(07)	.22(10)	.21(09)	.24(10)
Production	.22(09)	.22(09)	.23(09)	.21(09)	.23(09)	.20(08)	.22(09)	.23(09)	.23(10)
Quantiative Ability									
Ratings	.22(09)	.21(07)	.24(10)	.22(09)	.23(08)	.20(07)	.24(10)	.22(08)	.25(09)
Rankings	.22(09)	.21(08)	.24(10)	.23(10)	.22(08)	.21(09)	.24(10)	.20(07)	.24(09)
Work Samples	.22(09)	.20(08)	.23(09)	.21(09)	.22(09)	.20(07)	.23(09)	.20(08)	.24(09)
Production	.23(08)	.22(08)	.23(09)	.23(08)	.22(09)	.22(07)	.23(08)	.21(08)	.24(10)
Perceptual Speed									
Ratings	.22(09)	.21(08)	.24(09)	.22(09)	.22(09)	.20(07)	.25(10)	.21(09)	.23(09)
Rankings	.22(09)	.20(08)	.24(10)	.23(10)	.22(08)	.21(09)	.24(10)	.20(08)	.24(09)
Work Samples	.22(09)	.20(08)	.24(10)	.23(10)	.21(09)	.20(08)	.25(10)	.20(08)	.22(10)
Production	.23(09)	.22(09)	.24(09)	.23(09)	.22(09)	.21(08)	.24(10)	.20(10)	.23(08)
Memory									
Ratings	.23(09)	.21(07)	.25(10)	.24(09)	.21(09)	.22(08)	.25(10)	.19(07)	.24(10)
Rankings	.23(08)	.21(08)	.24(09)	.23(10)	.22(07)	.22(09)	.25(10)	.20(06)	.24(07)
Work Samples	.22(08)	.20(08)	.23(08)	.22(08)	.21(07)	.22(08)	.23(09)	.19(08)	.23(06)
Production	.22(08)	.21(07)	.24(09)	.23(08)	.22(08)	.21(08)	.24(09)	.20(07)	.24(08)
Spatial/Mechanical									
Ratings	.23(10)	.22(09)	.24(10)	.23(10)	.22(09)	.23(10)	.24(11)	.21(09)	.23(09)
Rankings	.23(10)	.22(10)	.24(10)	.24(10)	.22(09)	.24(11)	.24(10)	.20(08)	.24(10)
Work Sample	.22(08)	.21(07)	.23(09)	.21(08)	.22(08)	.21(08)	.22(08)	.20(07)	.24(10)
Production	.22(08)	.22(08)	.23(09)	.23(09)	.22(08)	.23(08)	.22(09)	.21(07)	.24(09)
Psychomotor									
Ratings	.23(10)	.21(08)	.25(12)	.23(11)	.23(10)	.21(08)	.26(12)	.21(08)	.25(12)
Rankings	.23(10)	.21(08)	.25(11)	.23(09)	.23(10)	.21(09)	.24(10)	.21(06)	.25(12)
Work Samples	.22(09)	.21(08)	.23(09)	.21(08)	.23(09)	.20(08)	.23(08)	.21(07)	.24(11)
Production	.22(08)	.22(08)	.23(08)	.22(08)	.23(08)	.21(08)	.23(08)	.22(08)	.24(09)

[a] Standard Deviations are in parentheses

Table 19

Means and Standard Deviations of Confidence Interval Judgments for the "Sales" Job Title

Type of Validity	(1) Overall	(2) Job Information	(3)	(4) VG Information	(5)	(6) Experimental Condition	(7)	(8)	(9)
		Congruent	Incongruent	Absent	Present	1	2	3	4
N	114	56	58	61	53	28	33	28	25
GMA									
Ratings	.23(09)[a]	.23(09)	.23(09)	.23(09)	.23(10)	.22(09)	.24(09)	.24(10)	.23(10)
Rankings	.24(09)	.23(09)	.24(09)	.24(09)	.23(09)	.23(09)	.24(10)	.22(08)	.25(09)
Work Samples	.22(09)	.21(08)	.23(09)	.22(07)	.22(10)	.21(07)	.23(08)	.21(09)	.23(11)
Production	.23(09)	.22(09)	.24(10)	.22(08)	.24(10)	.22(09)	.23(07)	.23(07)	.26(12)
Verbal Ability									
Ratings	.22(10)	.22(09)	.23(11)	.22(09)	.22(11)	.21(08)	.23(10)	.22(09)	.23(12)
Rankings	.22(09)	.21(08)	.23(09)	.22(09)	.22(09)	.21(09)	.23(09)	.21(07)	.23(10)
Work Samples	.21(09)	.20(07)	.23(10)	.21(09)	.21(08)	.20)07)	.23(10)	.19(07)	.23(09)
Production	.23(10)	.22(08)	.24(11)	.23(10)	.23(10)	.22(08)	.24(11)	.22(08)	.24(12)
Quantitative Ability									
Ratings	.22(08)	.21(08)	.23(09)	.23(09)	.21(08)	.22(08)	.24(09)	.19(07)	.22(08)
Rankings	.22(09)	.21(09)	.22(08)	.23(09)	.21(08)	.23(09)	.23(09)	.20(09)	.22(08)
Work Samples	.21(08)	.20(08)	.22(09)	.21(09)	.21(08)	.21(08)	.22(09)	.20(08)	.22(09)
Production	.22(09)	.21(09)	.23(08)	.22(09)	.21(08)	.22(10)	.23(09)	.20(08)	.23(08)
Perceptual Ability									
Ratings	.22(09)	.20(08	.23(09)	.23(10)	.21(07)	.22(09)	.24(11)	.19(07)	.23(08)
Rankings	.22(09)	.20(08)	.23(10)	.23(10)	.21(08)	.21(09)	.23(11)	.19(07)	.23(09)
Work Samples	.21(09)	.20(07)	.23(10)	.22(09)	.21(08)	.21(08)	.23(10)	.19(07)	.23(09)
Production	.22(09)	.21(08)	.23(10)	.23(10)	.22(08)	.23(08)	.24(11)	.20(08)	.23(07)
Memory									
Ratings	.23(08)	.21(07)	.24(09)	.23(09)	.22(08)	.22(07)	.24(10)	.20(07)	.25(09)
Rankings	.23(08)	.21(07)	.24(09)	.23(09)	.22(08)	.23(08)	.23(10)	.20(07)	.25(08)
Work Samples	.22(08)	.21(07)	.24(09)	.23(09)	.22(08)	.23(08)	.23(10)	.20(07)	.23(08)
Production	.22(08)	.21(08)	.23(09)	.23(08)	.22(08)	.23(08)	.23(09)	.20(08)	.24(09)
Spatial/Mechanical									
Ratings	.21(09)	.20(08)	.22(10)	.21(10)	.21(08)	.21(08)	.22(11)	.19(08)	.22(08)
Rankings	.21(09)	.20(08)	.22(10)	.22(11)	.20(07)	.22(10)	.22(11)	.18(05)	.23(09)
Work Samples	.21(08)	.20(07)	.22(08)	.21(09)	.20(07)	.21(09)	.21(09)	.19(06)	.22(07)
Production	.21(09)	.20(08)	.23(10)	.22(11)	.20(07)	.22(09)	.22(12)	.18(06)	.23(08)
Psychomotor									
Ratings	.22(10)	.19(07)	.25(11)	.23(11)	.21(07)	.20(08)	.26(13)	.18(06)	.23(08)
Rankings	.21(09)	.19(07)	.23(10)	.23(10)	.20(07)	.21(09)	.24(12)	.18(06)	.22(08)
Work Samples	.21(08)	.19(07)	.23(09)	.22(10)	.20(07)	.20(08)	.23(10)	.18(05)	.22(08)
Production	.21(09)	.19(07)	.23(10)	.23(10)	20(07)	.21(08)	.24(11)	.18(05)	.22(08)

[a] Standard Deviations are in parentheses

Table 20

Effects of Job Information and Validity Generalization Information on Confidence Judgments for "Clerical" and "Sales" Job Titles

Effect	Df Num,Den	Wilks'	F	Canonical R	R^2
		"Clerical" Job Title			
VG Information	28,83	.680	1.40	.57	.32
Job Information	28,83	.802	.73	.44	.20
VG x Job Information	28,83	.704	1.25	.54	.30
		"Sales" Job Title			
VG Information	28,82[a]	.803	.72	.44	.20
Job Information	28,82	.734	1.06	.51	.27
VG x Job Information	28,82	.786	.80	.46	.21

[a] One subject was eliminated from the analysis because four sales SKAPs validity judgments were missing.

* $p < .05$
** $p < .001$
*** $p < .0001$

Table 21

Between- and Within-Subject Effects Across all Confidence Interval Judgments

Effect	Df Num,Den	Wilks' Λ	F	Canonical R	R^2
Type of Job (TofJ)	1,109	.973	3.02	.16	.03
TofJ x VG Information (VG)	1,109	.966	3.87*	.19	.04
TofJ x Job Information (Job)	1,109	.993	.82	.09	.01
TofJ x VG x Job	1,109	.994	.65	.08	.01
Predictor (TofP)	6,104	.844	3.21**	.40	.16
TofP x VG	6,104	.904	1.83	.31	.10
TofP x Job	6,104	.924	1.42	.28	.08
TofP x VG x Job	6,104	.869	2.60*	.36	.13
Criterion (TofC)	3,107	.914	3.37*	.29	.08
TofC x VG	3,107	.991	.31	.09	.01
TofC x Job	3,107	.981	.70	.14	.02
TofC x VG x Job	3,107	.964	1.34	.19	.04
TofJ x TofP	6,104	.906	1.79	.31	.10
TofJ x TofP x VG	6,104	.873	2.52*	.36	.13
TofJ x TofP x Job	6,104	.949	.93	.23	.05
TofJ x TofP x VG x Job	6,104	.942	1.07	.24	.06
TofJ x TofC	3,107	.981	.71	.14	.02
TofJ x TofC x VG	3,107	.993	.27	.09	.01
TofJ x TofC x Job	3,107	.938	2.37	.25	.06
TofJ x TofC x VG x Job	3,107	.978	.81	.15	.02
TofP x TofC	18,92	.848	.92	.39	.15
TofP x TofC x VG	18,92	.787	1.38	.46	.21
TofP x TofC x Job	18,92	.871	.76	.36	.13
TofP x TofC x VG x Job	18,92	.904	.54	.31	.10
TofJ x TofP x TofC	18,92	.898	.58	.32	.10
TofJ x TofP x TofC x VG	18,92	.816	1.15	.43	.18
TofJ x TofP x TofC x Job	18,92	.855	.87	.38	.14
TofJ x TofP x TofC x VG x Job	18,92	.887	.65	.34	.12

* Pr.<.05
** Pr.<.001

CHAPTER VI

Summary, Discussion, and Implications

Professionals in the personnel selection field who are trusted with the responsibility of test development and validation must scrutinize the techniques they use to establish the value of selection processes. When a new technique is introduced, such as the rational estimation methodology suggested by Schmidt et al. (1983), it is necessary to determine if it is an appropriate means for establishing test validity.

The value of rationally estimated criterion-related validities is based on the accuracy of judges' estimates and on the meaningfulness of inferences that such estimates allow. The meaning of experts' rational judgments can be understood by studying the processes judges use to derive their estimates and by identifying the task-related conditions that affect their output. While research has begun to explore the issue of judgment accuracy (Schmidt et al., 1983; Hirsh et al., 1986), to date virtually nothing is known about the expert validity judgment process.

The present research examined the Schmidt et al. (1983) methodology by studying the effects of task-related information on experts' judgments of test validities.

Specifically, the research focused on examining the effects of relevant job information (i.e., job titles and SKAPs) and validity generalization results on validity estimates for two different jobs. Judges' confidence in each validity estimate was also measured.

This chapter contains a summary and discussion of the study's findings. In the concluding section of the chapter, implications of the research findings for the study and practice of Industrial and Organizational Psychology will be presented.

Summary and Discussion

The role of task-related information was studied among four groups of experts representing varying conditions of job and validity generalization information. A total of 28 validity judgments (i.e., 7 predictors x 4 criteria) for two different jobs were elicited from 114 Fellows, Members, and Associates of the Society for Industrial and Organizational Psychology, Inc. For each judgment, judges also estimated the 90% confidence interval around their estimate.

Hypotheses were advanced regarding the effects of job and validity generalization information on validity and confidence judgments. Hypothesis #1 predicted a main effect for job title such that even though SKAP information was identical, judgments would differ significantly reflecting differences in the job titles that were provided.

Hypothesis #2 predicted that judges who were not provided with validity generalization results would have significantly different validity judgments than judges who were provided with this information. Hypothesis #3 suggested that judges' confidence in their validity estimates, as evidenced by the size of the confidence interval around each judged validity, would be greater when validity generalization results were provided than when they were not provided. Similarly, Hypothesis #4 predicted that when job information was congruent (i.e., job title and SKAPs matched), judges would be more confident in their validity estimates than when the job information was incongruent.

Effects of Job Information on Experts' Validity Estimates

Effect of job title. MANOVA and ANOVA analyses which were conducted to test Hypothesis #1 - the main effect for job title - showed different results for validity estimates made for clerical versus sales SKAPs. While not confirmed for the clerical SKAP judgments, Hypothesis #1 was supported for the sales SKAP judgments. For both sets of SKAP judgments, there were significant (multivariate) main and interaction effects involving job information and validity generalization information. In all cases, these effects accounted for a fairly large percentage of the variance in validity estimates (i.e., from 35% to 52%).

ANOVA analyses, which measured the effects of job title on individual validity judgments, clarified the results found in the MANOVA analyses. Congruent/incongruent job title had a significant, independent main effect on only three (10.7%) of the 28 clerical SKAP judgments. However, the effect of job title on sales SKAP judgments was significant for 16 of the 28 (57.1%) validity estimates. The significant ANOVA job title effects for the 16 sales SKAP judgments showed a consistent pattern for quantitative, perceptual speed, memory, and psychomotor validities such that estimates were higher when the "clerical" job title, rather than the "sales" job title, was provided. Because the SKAP information was held constant, these results show clearly that judgments of validity for sales SKAPs were significantly affected by the type of job title that was provided.

Effect of SKAPs. Although job title clearly affected validity judgments, it should be noted that experts' validity judgments were also significantly affected by the SKAP information that was provided, irrespective of job title. When job titles were held constant and only SKAP information varied, variance accounted for by SKAP information was 41% in judgments made for the "clerical" job title and 38% for the "sales" job title. In addition, ANOVA analyses of judgments for "clerical" and "sales" job titles showed significant differences for 9 (32.1%) and 10 (35.7%) of the validity judgments, respectively. The results from

MANOVA and ANOVA analyses of "clerical" and "sales" job title judgments show that judges did use the SKAP information that was provided. This is further substantiated by comparing average judged validities in the congruent and incongruent conditions to the actual SKAP importance ratings. Levels of average judged validities for particular predictors reflected the relative importance ratings assigned to the corresponding SKAPs.

For the "clerical" job title, experts' judgments were consistent with the SKAP information for 4 of the 7 types of predictors. Validity estimates were appropriately higher for verbal ability, perceptual speed, memory, and spatial/mechanical when clerical SKAPs were provided and lower when sales SKAPs were provided. Judgments of general mental ability, quantitative and psychomotor were not consistent with the differences between the clerical and sales SKAP information.

For the "sales" job title, experts' judgments were consistent with the SKAP information for 5 of the 7 types of predictors. Based on the SKAP information that was provided, experts' over-estimated the magnitudes of memory and psychomotor validities.

<u>Conclusions about the effects of job information</u>. Taken together, analyses for clerical and sales SKAPs and "clerical" and "sales" job titles suggest that although SKAP information affected judges' validity estimates, that job titles, alone, accounted for important differences in the

magnitudes of validity estimates when sales SKAP information was provided. Therefore, Hypothesis #1 was supported when sales SKAPs, but not when clerical SKAPs were provided. The average differences between sales SKAP validity judgments made for the clerical and sales job titles were .045 for general mental ability, .033 for verbal ability, .115 for quantitative, .095 for perceptual speed, .098 for memory, .025 for spatial/mechanical, and .065 for psychomotor. All of these differences reflect higher validities when the "clerical" job title was provided. (It should be noted that significant ANOVA effects for sales SKAP judgments were found for all criterion judgments involving quantitative ability, perceptual speed, memory, and psychomotor predictors.)

Based on the rationale for Hypothesis #1, it is concluded that job titles were more salient determinants of validity judgments when sales SKAP information was provided than when clerical SKAP information was provided. More specifically, it may be (in the case of the clerical SKAP judgments) that the incongruent job title (i.e., the "sales" job title) was less salient than the clerical SKAP information. This would have occurred if the clerical SKAP information easily connoted a clerical type of job (i.e., high in quantitative, perceptual speed, and memory). Therefore, under both job title conditions, when clerical SKAPs were provided judges were basically estimating validities for a clerical type of job. The job title main effect was realized for the sales SKAPs possibly because the sales SKAP information was not as readily associated with a "sales" job. If the sales SKAP information was somewhat ambiguous, the job title may have been a more salient cue upon which judges needed to rely. The effects described above for clerical and sales SKAPs were dependent on the specific types of information and jobs used in this judgment task. Therefore, the appropriateness of generalizing the job information effect to other judgment situations needs further investigation. Nonetheless, the fact that an effect was observed for sales SKAPs suggests that the rational estimation procedure cannot blithely be generalized.

This analysis of the job information effect suggests that either job titles or SKAP information may be more concrete, depending on characteristics of both and how they relate to each other. For example, if the SKAP information easily connotes a particular type of job and the job title is in conflict with this stereotype, then a judge may be forced to reconcile this incongruency. The SKAP information may be used more easily to derive an "image" of the job for which validities are being judged because it unambiguously provides specific information regarding the profile of skills and abilities required in the job. More generally, it might be suggested that the "concreteness" of information may change if, in combination with other information, the information becomes more ambiguous, uninterpretable, or less well-defined.

Effect of Validity Generalization Information on Experts' Validity Estimates

Validity generalization information was hypothesized to have an effect on experts' validity estimates because of its availability (Tversky & Kahneman, 1973) and the value attributed to "outcome" knowledge (Fischhoff, 1975). This main effect was highly statistically significant for validity judgments made for both the "clerical" and "sales" job titles, and accounted for 47% and 38% of the variance in experts' estimates, respectively. When validities were being judged for the "clerical" job title, the presence of validity generalization results was associated with higher estimates for 21 of the 28 judged validities (85.7%). The average differences for judges in the "vg" and "no vg" conditions (where main effects were evidenced) were .11 for general mental ability, .06 for verbal, .07 for quantitative, .09 for perceptual speed, .07 for memory, .06 for spatial/mechanical, and .09 for psychomotor. Clearly, when clerical validity generalization results were present, experts' validity estimates were significantly increased, whether clerical or sales SKAPs were provided.

The presence of validity generalization information did not have as large an effect on "sales" job title judgments as on judgments for the "clerical" job. Validity estimates made for the "sales" job title in the presence of validity generalization results were only significantly different from estimates made in the absence of this information for three (10.7%) of the judges' estimates. However, there was a trend towards judges using the validity generalization information that was provided. This trend can be observed by comparing average validity judgments in the "vg" and "no vg" conditions to the actual sales validity generalization results that were provided. Of the 28 validity judgments, 18 judgments were consistent with the validity generalization results (i.e., validity estimates from judges provided with validity generalization information differed from estimates from judges who did not have this information in the same direction as the provided value). (For example, if the "no vg" group is considered the "control" group: examine the average estimated validity for verbal ability with a work sample criterion. The validity generalization value is .50, the "no vg" group average estimate is .36, and the "vg" group average estimate is .46.) Where "mismatches" existed, there was a strong tendency for judges who had validity generalization information to judge validities higher, even though the provided validity generalization values were lower.

Conclusions regarding the effect of validity generalization information. Taken together, results for the "clerical" and "sales" job titles suggest that validity generalization results have an inconsistent affect on experts' validity estimates. If one assumes on the basis of the significant main effects evidenced in the MANOVA analyses that validity generalization information was salient, then it is necessary to consider alternative explanations for its different affect on validity estimates for the two job titles. It may be, for example, that judges' anchors are not the validity generalization results as was suggested in the Problem Statement but, rather, some generalized validity value such as .30. Validity generalization information may then be used to alter this anchor (Lichtenstein & Slovic, 1971). However, if the validity generalization values do not differ substantially from the original "anchor", then one would expect results in the presence <u>and</u> absence of this information to be fairly similar. According to Tversky and Kahneman (1972), this would occur because judges tend to make inadequate adjustments to "anchor" values.

This explanation is partially supported by the observation made above that the majority (64.3%) of the average "sales" validity judgments made in the presence of sales validity generalization results decrease/increase above the corresponding average judgments in the "no vg" condition in the same direction as the validity generalization value. Therefore, it appears that experts are being influenced by the validity generalization results, but the influence may not be substantial enough to result in statistically significant differences in validity estimates for the "vg" and "no vg" groups. (On the other hand, "clerical" validity judgments made in the presence of clerical validity generalization results are consistent with these results 85.7% of the time. As can be seen by comparing the validity generalization information that was provided to judges, the clerical validity generalization values provided are generally greater in magnitude than the sales validity generalization values.)

A second explanation assumes that validity generalization results may not have been as influential as originally predicted. Although the MANOVA analyses and the previous paragraph both support the conclusion that validity generalization results affected experts' judgments somewhat for the "sales" job title, the effect may not have been as large as necessary to result in statistically significant differences between "vg" and "no vg" groups. This could have occurred, for example, if the validity generalization results for the "sales" job title were less credible or acceptable to the experts. Given the professional attention that has been paid to clerical validity generalization studies, and the paucity of studies for sales-like jobs, it would not be surprising if judges did not place as much value on the validity generalization results that were presented for the latter type of job. Therefore, the sales validity generalization results may have been somewhat discounted relative to the clerical validity generalization results.

<u>Summary of results for Hypotheses #1 and #2.</u> Although the two hypotheses regarding the effects of task-related information were not consistently supported for validity judgments made for "clerical" and "sales" job titles or SKAPs, it does appear that experts were fairly susceptible to the experimental manipulations employed in this research. Differences in the average validity estimates from judges in different experimental conditions were not trivial in many cases (i.e., sometimes as high as .19). This suggests that even if experts' average observed estimates are somewhat accurate, we may need to question whether this accuracy may be due in part to characteristics of the information provided in the judgment task environment.

It may also be necessary to scrutinize other characteristics of the validity judgment task environment (e.g., how estimates are collected, what types of judgments are required, etc.), before it can be concluded that expert accuracy is a meaningful phenomenon. In this research, many of the experts' validity judgments were not only subject to the job and validity generalization information that was provided, but to the order in which judgments were made for different predictors. This order effect (which was evidenced as a main effect for "sales" but not "clerical" validity judgments, and as an interaction effect involving type of criterion for both job titles) demonstrates that many often unanticipated factors associated with the judgment task may affect expert output.

Effects of Job Information and Validity Generalization Results on Experts' Confidence Interval Judgments

Researchers have found that individuals tend to be very confident in the output from complex judgment tasks despite the fact that most of these judgments are very difficult to make (Goldberg, 1959). In addition, factors in the judgment situation have been found to increase judges' confidence even when these factors were not related to judgment accuracy (Oskamp, 1965). With respect to expert judgments of test validities, when validity generalization information was available, or job information was congruent, experts were expected to be more confident (Hypotheses 3 and 4, respectively). These predictions were taken directly from research which has found that when judges have more information on which to base their judgments, or when different types of information are consistent, confidence is increased (Oskamp, 1965; Kahneman & Tversky, 1973; Einhorn & Hogarth, 1978). Although it was hypothesized that these two information conditions would increase the level of confidence that judges had in their output, the data showed a general lack of support for both Hypothesis #3 and #4.

Despite the lack of statistically significant main effects for job information and validity generalization results, the MANOVA analyses show that these effects accounted for a non-trivial amount of variance in linear combinations of "clerical" and "sales" job title judgments. For the job information main effect, the variance accounted for in "clerical" and "sales" judgments was 20% and 27%, respectively. The variance accounted for by the validity generalization main effect was 32% and 20%, respectively.

The repeated measures MANOVA, which measured the effects on confidence interval judgments of predictor, criterion, and job factors along with job and validity generalization information, did evidence some significant interaction effects. Type of predictor and type of criterion were associated with different sized confidence intervals for individual judges. In addition, a significant higher-order interaction involving type of job, type of predictor, and validity generalization information indicated that the effect of different jobs on individual judges' average confidence judgments for different predictors was not the same for judges in the "no vg" and "vg" conditions. This interaction was expected as it reflects, in part, the different levels of the VG information to which judges in the "vg" condition were exposed.

Conclusions regarding the effects on confidence judgments. It may be that expert confidence is somewhat sensitive to inconsistent task-relevant information as was suggested by Kahneman and Tversky (1973). The inconsistency between job title and SKAP information may have been salient to judges because the amount of task-related information with which they were provided was limited. The incongruency would be expected to be most salient where job titles connote a strong stereotype which amplifies the conflict between the job title and SKAP information.

This inference receives some support from t-test analyses (unreported) which showed that confidence judgments for the "clerical" job title were more affected by the incongruent job information than were similar judgments for the "sales" job title. Apparently, the inconsistency between job title and SKAPs was more salient for the "clerical" job for which the more widely held stereotype probably exists. (However, it should be noted that despite the fact that 50% of the t-tests comparing "clerical" confidence judgments from congruent and incongruent conditions were significant, the average difference in estimated validities in the 14 significant t-test comparisons was only .036.)

Because the validity generalization results were manipulated between subjects, the absence versus presence of this information would not have been salient to judges. However, because validity generalization results had a significant affect on experts' validity judgments, signifying that experts did rely somewhat on this information when it was available, it is curious that expert confidence was not increased when this influential source of validity estimates was available.

Based on the lack of significant multivariate effects of job information and validity generalization results on confidence judgments, and the relatively small average difference where significant univariate effects were observed, it appears that forces are operating to stabilize expert confidence. One possible explanation is that Industrial and Organizational Psychologists may have fairly homogenous perceptions regarding the range of true validities. The effect of this perception may be so stable that judges' average confidence interval (which is .22 across the four experimental conditions and 28 validity judgments) may be unaffected by characteristics of the task-related information. An alternative explanation is that the expert judges were simply not effective at translating their subjective confidence into actual confidence intervals.

Implications

If Industrial and Organizational Psychologists are to use rationally-estimated validities appropriately when making inferences about the value of a selection procedures, it is necessary that they understand the judgment conditions that affect the estimates made by experts. Because of the potential complexity of the validity judgment task, one must assume that a multitude of different factors may possibly affect the output of judges. This research, which examined how task-related information affected experts' judgments of test validities, was only able to focus on a limited, but important segment of the judgment task environment.

Data regarding how experts use information that is provided to them in the judgment task can expand our understanding of the validity judgment process by delineating which aspects of the task environment are critical to the estimates produced by groups of experts. On a practical level, if we can identify those types of information that have the strongest influence on judges' output, we can structure rational estimation procedures to ensure that rational estimates of validity are both meaningful and interpretable. In addition, if research indicates that experts' validity judgments are overly sensitive to irrelevant information, or plagued by inappropriate biases, then information will exist to determine whether the output from rational estimation procedures is at all valuable. On a more theoretical level, by defining the types of task-related information that are most critical to judges' output, we can add to our understanding of complex judgment processes. Therefore, the results of this study of job and validity generalization effects on experts' validity judgments have implications for both the study and the practice of Industrial and Organizational Psychology.

The Use of Rational Validity Estimates

Experts' validity estimates. Hypothesis #1 and Hypothesis #2, both of which received partial support, demonstrated that under some conditions, experts' judgments are likely to be unduly influenced by inappropriate or irrelevant information. In this judgment situation, job titles and validity generalization information produced significant differences in validity estimates when all other information in the judgment task was held constant. (The reader is reminded that (1) judges' validity estimates should not have differed when SKAPs were identical but job titles differed; and (2) that the validity generalization information was not relevant for those judges who were in the incongruent condition; i.e., when the SKAPs did not describe the same type of job as was indicated by the validity generalization information.) These effects were not consistent across different sets of SKAPs and different job titles, suggesting that judgment conditions may interact with types of judgments to further complicate our ability to understand and predict the factors that affect experts' validity estimates.

Given the results obtained in the research and summarized above, it is clear that even in the face of findings that suggest that experts', under some conditions, may accurately estimate validities, the assumption that these estimates carry the same meaning as empirically-derived estimates is specious. As Einhorn (1974) has suggested, expertise is as much a function of the accuracy of output as it is of the meaningfulness of process. If judges are not using information accurately, or are unreasonably influenced by irrelevant information, then no matter how accurate their output, there is good reason to question its appropriateness.

On a practical level, the type, amount, and combination of information provided to judges are of critical importance because they determine which information is most salient and, therefore, on which information judges' validity estimates are based. The results of this research have shown that in the absence of the misleading job titles and/or validity generalization results, judges' estimates would probably have been more appropriate because they would have come closer to reflecting the SKAPs that were identified as critical to job performance. However, the sensitivity of judges to the SKAP information may have existed because only a limited amount of information was provided in the judgment task.

Type of job may affects how information is used when judging validities and what average estimated validities result from the process. The effects of job information and validity generalization results were not the same for the two different jobs, suggesting that existing beliefs about a job (i.e., stereotypes), or merely degree of familiarity with a job, may affect how information that is presented in the judgment task is used. Judges, even those with expertise in the field of personnel testing are unlikely to use provided information optimally if their own job-specific beliefs about levels of validity, importance of SKAPs, etc., are most salient.

Experts' confidence interval estimates. An additional concern with the meaningfulness of judges' output concerns their ability to identify the degree of error which might exist in their point estimates of true validity. From this research it appears that the 90% confidence interval around judged validities is substantially unaffected by differences in task-related information or type of job, even though these factors affected average validity judgments. Clearly, for experts' validity estimates to be meaningful substitutes for empirical validities, it would be important for judges to be able to recognize when their own estimates are likely to be more or less errorful, as this information can be easily known about empirically-derived validities.

Implications of Study Results for Research on Complex Judgments Processes

Although this research differed substantively from other psychological studies of complex judgment processes, it attempted to apply and study conclusions drawn from these studies regarding the factors likely to affect experts' judgments. Several findings from this research have implications for the further study of complex judgment processes.

Anchoring and adjustment. The research results partially support the conclusion that judges were using an anchoring and adjustment process (Lichtenstein & Slovic, 1971). From the statistical results of the validity generalization effect, it appears that at least in some situations, validity generalization information was used to make adjustments to some initial value. (The reader may recall that judges' estimates in the presence of validity generalization results generally fell between the estimates made without this information and the actual validity generalization results provided.) However, little is known about how the initial value was identified, or if if this process is consistent across different judgment conditions. Based on the overall lower level of estimated true validities in comparison to validity generalization results, one might infer that judges are using some preconceived, generalized value as their "baseline" estimate. This inference is based on Lichtenstein and Slovic's (1971) explanation of the "anchoring and adjustment" bias which suggests that judgments will not differ substantially from initial values because adjustments to these values are minimal. This inference should be studied further because it has important implications for the anchoring and adjustment phenomenon and the use of rational estimation methods.

The salience of "concrete" information. It was hypothesized that the job title would be more salient because it would be more "concrete" than the SKAP information (Nisbett & Ross, 1980). However, this justification was not entirely upheld by the research results. In this research, it appeared that one set of SKAPs clearly connoted a specific type of job (i.e., clerical) such that when an incongruent job title was associated with the SKAPs, the job title did not have a significant affect on validity judgments. In this situation, the SKAPs may have been more concrete than the job title because they presented a less ambiguous image of the job. Therefore, it appears that information may be more or less concrete as a function of other information which is available. The Nisbett and Ross (1980) suggestion, that some information may be more salient because it is more concrete, may need to be expanded to include the notion that "concreteness" is not a static property of information but, rather, a function of the context within which the information is perceived and interpreted.

Factors affecting the salience of validity generalization information. Validity generalization results were expected to have a significant effect on validity estimates because they provided a ready source of validity values (Tversky & Kahneman, 1973) and because they constitute valued validity information (Fischhoff, 1975). The research indicated that experts were generally strongly influenced by validity generalization results. However, the effects were not consistently powerful across different jobs. Although this information was equally available for both types of jobs, it was not equally valued. Based on these empirical results, it was hypothesized that clerical validity generalization results were considered to have more value than sales validity generalization results. (Primarily because the former have received a large amount of attention in professional journals.) If this inference is accurate, it would suggest that Fischhoff's (1975) research on the effects of outcome knowledge might need to be expanded to include the notion of "valued" outcome knowledge.

Factors affecting confidence judgments. Experts' average confidence intervals were expected to be affected by the amount of information (Oskamp, 1965) and the type of information (Kahneman & Tversky, 1973) that was provided in the judgment task. However, hypotheses concerning these effects were not upheld. It may be that for these factors to affect the amount of confidence that judges have in their output, there can not be any other more powerful, overriding source that determines the degree of error that judges perceive may be associated with their validity estimates. In this research situation, it appears that experts are rather homogenous regarding their perceptions of the range of true validities, and this range was independent of the information or process that was used to derive the true validity estimates. Studies of judgment confidence have not examined how factors within the judgment situation affect judged confidence when preconceived notions of the range of judgments exist. This is an important area of research in rational estimation procedures because confidence in point estimates of validities should determine, in part, how these estimates are used. If these confidence estimates are confounded by other factors, then important information is missing.

APPENDIX A

The Graduate School and University Center
of The City University of New York

Ph.D. Subprogram in Industrial & Organizational Psychology / Box 512
Bernard M. Baruch College 17 Lexington Avenue New York N.Y. 10010
212 725 3074 3080

April 29, 1987

Dear Colleague:

Recently the suggestion was made that experts might be used to rationally estimate test validities in situations where empirical validation research is infeasible (Schmidt, Hunter, Croll & McKenzie, 1983; Hirsh, Schmidt & Hunter, 1986). This suggestion has important implications for the practice of industrial/organizational psychology. However, additional research is needed before conclusions can be drawn about the value of this new validation strategy.

A research study has been designed that will expand upon the research cited above by investigating judgments of "true" (rather than observed) validities and judgments of validities involving several different types of criteria. Because Division 14 is the likely source of experts for organizations seeking to substantiate test use based on rational estimates of employment validities, it is the logical source of subjects for research that concerns rational estimation procedures. As a Division 14 member, your assistance in this research effort would be greatly appreciated. Approximately 45 minutes of your time is needed to complete the enclosed judgment task.

Enclosed in this envelope are the materials and information you will need to estimate the true validities of different cognitive ability tests and measures of job performance for two different jobs. The judgment task has been designed to be similar to the way in which industrial and organizational psychologists might be asked to apply this methodology in practice. Your responses will be kept anonymous and confidential and will be used for research purposes only. Therefore, please do not identify yourself on the judgment materials.

Thank you in advance for your assistance and the generous contribution of your time and expertise. I will be happy to send you a summary of the results; simply indicate your interest on the postcard provided. Again, thank you.

Sincerely,

Karen Weinberg

GENERAL INSTRUCTIONS AND PROCEDURES

The task that you are about to perform simulates a judgment task concerning the value of various employment tests. The judgment materials accurately represent the types of materials available to psychologists who might actually be used to judge employment validities. You are asked to review the materials and approach the validity judgment task as if you were actually providing a real organization with your judgments of the true validity of employment tests for two different jobs.

To ensure that all judges approach the task in a uniform manner you have been provided with this set of instructions. It may be helpful for you to first browse through and familiarize yourself with the materials enclosed in this packet (except the folder marked JOB #2 - this should only be opened when you have finished judging validities for JOB #1). Then, after reading these instructions, you can review the materials more thoroughly and proceed with the judgment task.

General Description of the Judgment Task

The main focus of this research is on obtaining "expert" judgments of the true validity of several cognitive predictors and four types of criteria for two different jobs - a clerical and a sales job. You have been provided with different types of information that may be useful in making these judgments - i.e., organizational job descriptions of the jobs in question, job analysis results (skill/ability ratings), examples of tests which measure potentially relevant cognitive abilities, etc.

In addition to the judgments of **true validity**, you are also asked to indicate your **degree of confidence in each of your true validity judgments** by placing a confidence interval around each judged validity. Finally, to clarify the types of information needed to make such judgments, you are asked to indicate **how important each type of information is** when estimating true validities. (How these judgments will be made will be discussed in greater detail below.)

Procedures for Making Your Judgments

Before you begin the judgment task you should place the following materials within reach:

(1) **Reference Sheet #1** ("Descriptions of Predictors" and "Descriptions of Criterion Measures");
(2) **Reference Sheet #2** ("The Effects of Criterion Unreliability and Range Restriction on Observed Validity"); and
(3) **Folders marked "JOB #1" and "JOB #2"** (These numbers indicate the order in which you are to make your judgments.)

Then, please follow these steps when making your judgments for the two jobs by starting with JOB #1:

STEP #1: Read carefully the job descriptive information for JOB #1 and then any other information from the reference sheets that is relevant to the first validity judgment you are making. (This is probably a better strategy than reviewing all of the information first and then making all of the judgments.)

STEP #2: Make your judgment of **true validity** for the first predictor-criterion pair by placing a vertical line through the value that corresponds to your estimate on the scale on the right. **Your judgment of the true validity for any predictor-criterion pair should correspond to the unrestricted, unattenuated, population validity that you would expect in that particular job.**

STEP #3: Make your **confidence judgment** for the first predictor-criterion pair by placing brackets to the left and right of your true validity estimate to indicate the lower-bound ([ ) and upper-bound ( ]) of your **90% confidence interval. The 90% confidence interval is the range of validity values within which you are 90% certain the true** value actually falls.

STEP #4: Make and record your judgments of true validity and your confidence judgments for each of the remaining predictor-criterion pairs for JOB #1.

STEP #5: Follow the same steps (STEPS 1 - 4) to make your judgments for JOB #2.

STEP #6: After completing all of the true validity and confidence judgments for JOB #1 and JOB #2, then make your **Importance Ratings** for the various types of information you used in making your judgments. The sheet for making these judgments is included in the folder for JOB #2. The directions for making these judgments are at the top of the page labeled "**Importance Ratings**".

STEP #7: Complete the **Demographic Questions** and fill out the postcard.

STEP #8: Please check that you have made all of the judgments, that you have recorded your responses accurately, and that you have not left any of the demographic questions blank. **Then, mail the two Judgment Records and the Importance Ratings and Demographic Questions** in the stamped, addressed envelope provided. (It is not necessary to return any of the other judgment materials.) **The postcard is to be mailed separately.**

THANK YOU AGAIN FOR YOUR ASSISTANCE

APPENDIX B

REFERENCE SHEET #1

Descriptions Of Predictors

Type of Test	Definition	Examples	
General Cognitive Ability	The ability to learn a wide variety of tasks by understanding verbal and numerical concepts and principles and by being able to use such information effectively to perceive, understand, and solve problems. General Ability tests usually involve combinations of items or subtests that measure verbal, math, spatial, and reasoning abilities.		Wonderlic Personnel Test Personnel Classification Test Adaptability Test General Aptitude Test Battery Otis Self-Administering Tests of Mental Ability
Verbal Ability	The ability to use and understand spoken and written English in an effective manner. This ability is measured using a variety of different tests that measure grammar, spelling, vocabulary, and reading comprehension.	Vocabulary Grammar Reading Comprehension Spelling	Short Employment Tests - Verbal General Clerical Test (Part I) PSI Basic Skills Test - Reading Comprehension Comprehensive Test of Basic Skills (Test 4)
Quantitative Ability	The ability to understand simple mathematical concepts and principles, to perform simple mathematical operations quickly and accurately, and to apply such knowledge effectively to problems requiring quantitative solutions. This ability is measured using a variety of different tests such as arithmetic computation, arithmetic problems, and graphs/tables.	Arithmetic Computation Arithmetic Problems Graphs & Tables	Flanagan Industrial Tests - Arithmetic Tests of Clerical Ability - Arithmetic I Flanagan Industrial Tests - Scales
Perceptual Ability	The ability to perceive visual information (e.g., numbers, letters, pictures, figures) quickly and accurately and to perform simple tasks requiring perceptual comparison, evaluation, or scanning. This ability is measured using a variety of different tests such as number, name, or figure comparisons, coding/filing, or visual scanning. Tests of perceptual ability are generally highly speeded.	Number Compar. Name Compar. Figure Compar. Visual Scanning Coding/Filing	Minnesota Clerical Test Hay Aptitude Tests Battery - Name Finding Flanagan Industrial Tests - Inspection Employee Aptitude Survey - Visual Pursuit Flanagan Aptitude Classification Test - Coding

Type of Test	Definition
Memory	The ability to learn and ... or information presented ... or meaningless (e.g., ... including coding and ... simply to recall inform... information to complete ...
Spatial/Mechanical Ability	The ability to perceive, ... objects and to perceive ... physical/mechanical pr... measured using a vari... mechanical comprehen... than tests of mechanical ...
Psychomotor Ability	The ability to move or ... coordinate such move... requiring eye-hand co... may use paper/pencil ... requires examinees to ... objects and, possibly, ha...

Descriptions Of Criterion Measures

Supervisory Ratings:	Supervisory ratings of job performance are considered to be global measures of job performance because they represent job-related behaviors that occur over relatively long periods of time and because they reflect overall job performance. Supervisory ratings are generally summaries of multiple ratings of performance where the individual ratings may reflect general dimensions (e.g., "quality of work" or "job knowledge") or job-specific dimensions (e.g., "meets deadlines for submission of invoices" or "accurate at balancing cash drawer") of performance.
Production (Quality/Quantity):	Production criteria are objective measures of job proficiency which are available in jobs in which job performance results in an observable product or outcome. Measures of production frequently take into consideration both the amount produced or the rate of production (quantity) and the degree to which output is free of errors, conforms to standards, or was produced without waste (quality).
Supervisory Rankings:	Supervisory rankings ... compare the job perfo... rankings across multiple ...
Work Samples:	Work samples are ... the same behaviors, ... clerical tasks such as ... involve simulations of ... "production amount") ... "organizational skill", "...

REFERENCE SHEET #1

Descriptions Of Predictors

...amples		Type of Test	Definition	Examples	
	Wonderlic Personnel Test Personnel Classification Test Adaptability Test General Aptitude Test Battery Otis Self-Administering Tests of Mental Ability	Memory	The ability to learn and recall visual stimuli (e.g., numbers, letters, pictures, or figures) or information presented verbally. The information may be meaningful (e.g., directions) or meaningless (e.g., codes). This ability is measured using a variety of different tests including coding and following/applying directions. The tests may require examinees simply to recall information that was presented visually or orally, or to use such information to complete other tasks.	Coding Directions	Flanagan Aptitude Classification Tests - Memory Personnel Tests for Industry - Oral Directions
...cabulary ...rammar ...eading Comprehension ...pelling	Short Employment Tests - Verbal General Clerical Test (Part II) PSI Basic Skills Test - Reading Comprehension Comprehensive Test of Basic Skills (Test #[illegible])	Spatial/ Mechanical Ability	The ability to perceive, visualize, or mentally manipulate two- and three-dimensional objects and to perceive and understand the relationships between objects so that physical/mechanical principles are understood and applied accurately. This ability is measured using a variety of different tests including spatial relations, location, and mechanical comprehension. Tests of spatial ability are generally more highly speeded than tests of mechanical comprehension.	Spatial Relations Mechanical Comprehension Location	SRA Mechanical Aptitudes - Space Relations Flanagan Industrial Tests - Assembly MacQuarrie Test for Mechanical Ability - Location
...rithmetic Computation ...rithmetic Problems ...raphs & Tables	Flanagan Industrial Tests - Arithmetic Tests of Clerical Ability - Arithmetic II Flanagan Industrial Tests - Scales	Psychomotor Ability	The ability to move one's fingers, wrists, hands and arms rapidly and precisely and to coordinate such movements with one's eyes. This ability is measured using tasks requiring eye-hand coordination or finger/hand dexterity. Tests of psychomotor ability may use paper/pencil or may be performance-based. The latter type generally requires examinees to perform realistic tasks that involve the manipulation of small objects and, possibly, hand tools.	Coordination Finger/Hand Dexterity	Flanagan Aptitude Classification Tests - Coordination Purdue Pegboard
...umber Compar. ...ame Compar. ...igure Compar. ...isual Scanning ...oding/Filing	Minnesota Clerical Test Hay Aptitude Tests Battery - Name Finding Flanagan Industrial Tests - Inspection Employee Aptitude Survey - Visual Pursuit Flanagan Aptitude Classification Test - Coding				

Descriptions Of Criterion Measures

...erformance because they represent job-related ...all job performance. Supervisory ratings are ...fied general dimensions (e.g., "quality of work" ...ained" or "accurate at balancing cash drawer")	Supervisory Rankings:	Supervisory rankings are global measures of relative job performance. In general, supervisors are required to subjectively evaluate and compare the job performance of individuals within the same work group. As with supervisory ratings, the rankings may be summaries of rankings across multiple dimensions of job performance or may consist of a single overall ranking.
...which job performance results in an observable ...e amount produced or the rate of production ...d without waste (quality).	Work Samples:	Work samples are measures of job or task proficiency. They are usually developed directly from job functions and are designed to require the same behaviors, knowledge, and abilities that are required in the performance of the job. Clerical work samples may include samples of clerical tasks such as entering written orders into a computer, filing, checking the accuracy of invoices, etc. Sales work samples frequently involve simulations of customer interactions. Performance on work samples may be measured objectively (e.g., "number of errors", "production amount") or may consist of subjective ratings (using ratings scales or checklists) on dimensions of relevant performance (e.g., "organizational skill", "persuasiveness", etc.).

REFERENCE SHEET #2

THE EFFECTS OF CRITERION UNRELIABILITY AND RANGE RESTRICTION ON OBSERVED VALIDITY

Observed validities that most industrial/organizational psychologists are familiar with generally underestimate the actual relationship between test and job performance. This underestimation is due primarily to the effects of two different factors on the measurement of test validity. How these factors - criterion unreliability and predictor range restriction - affect observed validities is demonstrated below.

If true validity is . . . . . .60 .50 .40 .30 .20

and criterion reliability is equal to .70, then

the observed validity is . . . .50 .42 .33 .25 .17

if, in addition, the ratio of observed to true dispersion (i.e. range restriction) is equal to .60, then

the observed validity is . . . .33 .27 .21 .15 .10

The formulas to correct oberved validities for each of these effects so as to obtain estimates of true validity follow. When both corrections are necessary, the observed validity is corrected *first* for range restriction and then for unreliability.

To correct an observed validity for criterion unreliability:

$$\text{true validity} = \frac{r}{r_{xx}^{1/2}}$$

where r = observed validity
r_{xx} = criterion reliability (e.g., .70)

To correct an observed validity for range restriction:

$$\text{true validity} = \sqrt{[1 + k_{RRi}(1/r^2 - 1)]^{-1}}$$

where r = observed validity
k_{RRi} = ratio of test score SD in the selected group to test score SD in the unselected group.

APPENDIX C

JUDGMENT RECORD

Job Title: Product Sales Assistant

JUDGMENT RECORD

Job Title Product Sales Assistant

JUDGMENT RECORD

Job Title: Account Clerical Assistant

JUDGMENT RECORD

Job Title: Account Clerical Assistant

IMPORTANCE RATINGS

Instructions: Below is a list of the different types of information which were included in the judgment materials and which you may have used when making your true validity and confidence judgments. After reviewing this list, please add to the list (on the blank spaces provided) a brief description of any general type of additional information which you used (e.g., first-hand research experience, published studies, etc.). Then, rank order all of the types of information based on how important (i.e., influential) they were when you were making your validity judgments. Place the ranks you are assigning in the spaces provided on the left. (A lower number indicates greater importance. Please, no ties.)

Rank	Type of Information
____	Job Title/Job Family
____	Description of duties/responsibilities
____	Skill and ability ratings
____	Predictor and criterion descriptions
____	Validity generalization results
____	____________________
____	____________________
____	____________________
____	____________________

DEMOGRAPHIC QUESTIONS

EDUCATIONAL BACKGROUND

1. Type of degree (circle highest degree attained):

   Ph.D. Psy.D. Ed.D. MA/MBA/MS

2. Year degree completed: ____________________

3. Title of this degree (e.g., Ph.D. in Applied Social Psychology):

   ____

4. Area of specialization (if applicable (e.g., Human Factors))

   ____

RELEVANT EXPERIENCE

5. Primary work location (circle one which applies):

   Academia Private Industry Public Sector
   Private Consulting Military Other ____________

6. Brief description of this position (e.g., associate professor - teaching/research in organizational psychology; external consultant - training programs, etc.):

   ____
   ____

7. Years in current position: ____________________

8. Total years of experience in work described in #6 above:

   ____

9. Brief description of other relevant work experience (including years) within the field of Industrial/Organizational Psychology:

   ____
   ____
   ____

IMPORTANCE RATINGS

Instructions: Below is a list of the different types of information which were included in the judgment materials and which you may have used when making your true validity and confidence judgments. After reviewing this list, please add to the list (on the blank spaces provided) a brief description of any general type of additional information which you used (e.g., first-hand research experience, published studies, etc.). Then, rank order all of the types of information based on how important (i.e., influential) they were when you were making your validity judgments. Place the ranks you are assigning in the spaces provided on the left. (A lower number indicates greater importance. Please, no ties.)

Rank	Type of Information
____	Job Title/Job Family
____	Description of duties/responsibilities
____	Skill and ability ratings
____	Predictor and criterion descriptions
____	____________________
____	____________________
____	____________________
____	____________________
____	____________________

DEMOGRAPHIC QUESTIONS

EDUCATIONAL BACKGROUND

1. Type of degree (circle highest degree attained):

   Ph.D. Psy.D. Ed.D. MA/MBA/MS

2. Year degree completed: ____________________

3. Title of this degree (e.g., Ph.D. in Applied Social Psychology):

   ____

4. Area of specialization (if applicable [e.g., Human Factors])

   ____

RELEVANT EXPERIENCE

5. Primary work location (circle one which applies):

   Academia Private Industry Public Sector
   Private Consulting Military Other ____________________

6. Brief description of this position (e.g., associate professor - teaching/research in organizational psychology; external consultant - training programs, etc.):

   ____
   ____

7. Years in current position: ____________________

8. Total years of experience in work described in #6 above:

   ____

9. Brief description of other relevant work experience (including years) within the field of Industrial/Organizational Psychology:

   ____
   ____
   ____

APPENDIX D

JOB INFORMATION AND
JOB ANALYSIS RESULTS

Job Title: **Account Clerical Assistant** (Incongruent)

Duties and Responsibilities: The Account Clerical Assistant job is located in the Account Marketing group of a mid-size company. An individual in this job is responsible for handling customers' orders that come in over the phone or through the mail. This includes transferring customer orders onto order forms and entering the orders into a computerized database, determining the proper codes and prices of products being ordered based on information found in product catalogues, assigning order completion dates based on availability of products and requested date of shipment, determining cost and method of delivery based on size of shipment, and preparing customer correspondence.

Skill and Ability Importance Ratings: An empirical job analysis involving the full population (N = 237) of incumbents in the Account Clerical Assistant job and their supervisors was conducted. Raters were asked to rate the importance of each of the skills/abilities on a 5-point Likert-type scale (1="not at all important for successful job performance"; 5="extremely important for successf..l job performance").

The numerical results provided below are the percentage of respondents rating the skill/ability either 4 ("important") or 5 ("extremely important").

Skill/Ability	% Rating Important
Oral Communication — The ability to speak fluently and to understand spoken English words and sentences in an effective manner.	97%
Written Communication — The ability to understand written English and to write English so that others will understand.	61%
Number Facility — The ability to perform simple arithmetic operations quickly, accurately, and appropriately in response to problems requiring quantitative solutions.	66%
Reasoning — The ability to use information appropriately to understand a problem, develop a logical answer to a problem, or design a course of action to follow.	72%
Perceptual Speed — The ability to compare and to note similarities/differences among objects or symbols (e.g., words, codes, number amounts, etc.) quickly and accurately.	38%
Mechanical Comprehension — The ability to understand physical/mechanical principles and how to apply these to actual objects, tools, and machines.	8%
Memory — The ability to remember pertinent information such as rules or procedures, names, descriptions, amounts, etc. when needed.	90%
Selective Attention — The ability to concentrate on a task and maintain effective performance although distractions are present in the task environment.	33%
Fine Motor Coordination — The ability to manipulate small objects by moving one's fingers, wrists, and hands precisely in response to visual stimuli.	16%
Wrist-Finger Speed — The ability to move fingers, wrists, and hands as rapidly as possible where accuracy is less important than speed of movement.	14%
Detail Orientation — Skill at carrying-out tasks thoroughly and accurately by attending to all of the small features of the task.	31%
Persuasiveness — Skill at influencing the behavior and/or attitudes of others by providing them with convincing information.	93%
Social Sensitivity — Skill at interpreting the motives, feelings, and behavior of others and responding appropriately to achieve a desired effect.	89%

JOB INFORMATION AND
JOB ANALYSIS RESULTS

Job Title: Account Clerical Assistant (Congruent)

Duties and Responsibilities: The Account Clerical Assistant job is located in the Account Marketing group of a mid-size company. An individual in this job is responsible for handling customers' orders that come in over the phone or through the mail. This includes transferring customer orders onto order forms and entering the orders into a computerized database, determining the proper codes and prices of products being ordered based on information found in product catalogues, assigning order completion dates based on availability of products and requested date of shipment, determining cost and method of delivery based on size of shipment, and preparing customer correspondence.

Skill and Ability Importance Ratings: An empirical job analysis involving the full population (N = 237) of incumbents in the Account Clerical Assistant job and their supervisors was conducted. Raters were asked to rate the importance of each of the skills/abilities on a 5-point Likert-type scale (1="not at all important for successful job performance"; 5="extremely important for successful job performance").

The numerical results provided below are the percentage of respondents rating the skill/ability either 4 ("important") or 5 ("extremely important").

Skill/Ability	% Rating Important
Oral Communication — The ability to speak fluently and to understand spoken English words and sentences in an effective manner.	37%
Written Communication — The ability to understand written English and to write English so that others will understand.	52%
Number Facility — The ability to perform simple arithmetic operations quickly, accurately, and appropriately in response to problems requiring quantitative solutions.	92%
Reasoning — The ability to use information appropriately to understand a problem, develop a logical answer to a problem, or design a course of action to follow.	55%
Perceptual Speed — The ability to compare and to note similarities/differences among objects or symbols (e.g., words, codes, number amounts, etc.) quickly and accurately.	91%
Mechanical Comprehension — The ability to understand physical/mechanical principles and how to apply these to actual objects, tools, and machines.	9%
Memory — The ability to remember pertinent information such as rules or procedures, names, descriptions, amounts, etc. when needed.	69%
Selective Attention — The ability to concentrate on a task and maintain effective performance although distractions are present in the task environment.	94%
Fine Motor Coordination — The ability to manipulate small objects by moving one's fingers, wrists, and hands precisely in response to visual stimuli.	14%
Wrist-Finger Speed — The ability to move fingers, wrists, and hands as rapidly as possible where accuracy is less important than speed of movement.	19%
Detail Orientation — Skill at carrying-out tasks thoroughly and accurately by attending to all of the small features of the task.	96%
Persuasiveness — Skill at influencing the behavior and/or attitudes of others by providing them with convincing information.	17%
Social Sensitivity — Skill at interpreting the motives, feelings, and behavior of others and responding appropriately to achieve a desired effect.	20%

JOB INFORMATION AND
JOB ANALYSIS RESULTS

Job Title: **Product Sales Assistant** (Incongruent)

Duties and Responsibilities: A Product Sales Assistant is located in the Product Marketing group of a mid-size company and may be rotated to one of the company's regional stores as needed. An individual in this position is responsible for receiving and taking customer orders over the phone or in person, providing descriptions of the features of products to customers by reading product manuals and giving customers product pamphlets, using reference materials to obtain relevant product information, writing correspondence to customers concerning their orders, and obtaining payment and credit information.

Skill and Ability Importance Ratings: An empirical job analysis involving the full population (N = 219) of incumbents in the Product Sales Assistant job and their supervisors was conducted. Raters were asked to rate the importance of each of the skills/abilities on a 5-point Likert-type scale (1="not at all important for successful job performance"; 5="extremely important for successful job performance").

The numerical results provided below are the percentage of respondents rating the skill/ability either 4 ("important") or 5 ("extremely important").

Skill/Ability	% Rating Important
Oral Communication — The ability to speak fluently and to understand spoken English words and sentences in an effective manner.	27%
Written Communication — The ability to understand written English and to write English so that others will understand.	92%
Number Facility — The ability to perform simple arithmetic operations quickly, accurately, and appropriately in response to problems requiring quantitative solutions.	92%
Reasoning — The ability to use information appropriately to understand a problem, develop a logical answer to a problem, or design a course of action to follow.	95%
Perceptual Speed — The ability to compare and to note similarities/differences among objects or symbols (e.g., words, codes, number amounts, etc.) quickly and accurately.	89%
Mechanical Comprehension — The ability to understand physical/mechanical principles and how to apply these to actual objects, tools, and machines.	9%
Memory — The ability to remember pertinent information such as rules or procedures, names, descriptions, amounts, etc. when needed.	49%
Selective Attention — The ability to concentrate on a task and maintain effective performance although distractions are present in the task environment.	96%
Fine Motor Coordination — The ability to manipulate small objects by moving one's fingers, wrists, and hands precisely in response to visual stimuli.	14%
Wrist-Finger Speed — The ability to move fingers, wrists, and hands as rapidly as possible where accuracy is less important than speed of movement.	19%
Detail Orientation — Skill at carrying-out tasks thoroughly and accurately by attending to all of the small features of the task.	84%
Persuasiveness — Skill at influencing the behavior and/or attitudes of others by providing them with convincing information.	17%
Social Sensitivity — Skill at interpreting the motives, feelings, and behavior of others and responding appropriately to achieve a desired effect.	30%

JOB INFORMATION AND JOB ANALYSIS RESULTS

Job Title: Product Sales Assistant (Congruent)

Duties and Responsibilities: A Product Sales Assistant is located in the Product Marketing group of a mid-size company and may be rotated to one of the company's regional stores as needed. An individual in this position is responsible for receiving and taking customer orders over the phone or in person, providing descriptions of the features of products to customers by reading product manuals and giving customers product pamphlets, using reference materials to obtain relevant product information, writing correspondence to customers concerning their orders, and obtaining payment and credit information.

Skill and Ability Importance Ratings: An empirical job analysis involving the full population (N = 219) of incumbents in the Product Sales Assistant job and their supervisors was conducted. Raters were asked to rate the importance of each of the skills/abilities on a 5-point Likert-type scale (1="not at all important for successful job performance"; 5="extremely important for successful job performance").

The numerical results provided below are the percentage of respondents rating the skill/ability either 4 ("important") or 5 ("extremely important").

Skill/Ability	% Rating Important
Oral Communication — The ability to speak fluently and to understand spoken English words and sentences in an effective manner.	97%
Written Communication — The ability to understand written English and to write English so that others will understand.	61%
Number Facility — The ability to perform simple arithmetic operations quickly, accurately, and appropriately in response to problems requiring quantitative solutions.	86%
Reasoning — The ability to use information appropriately to understand a problem, develop a logical answer to a problem, or design a course of action to follow.	73%
Perceptual Speed — The ability to compare and to note similarities/differences among objects or symbols (e.g., words, codes, number amounts, etc.) quickly and accurately.	20%
Mechanical Comprehension — The ability to understand physical/mechanical principles and how to apply these to actual objects, tools, and machines.	8%
Memory — The ability to remember pertinent information such as rules or procedures, names, descriptions, amounts, etc. when needed.	59%
Selective Attention — The ability to concentrate on a task and maintain effective performance although distractions are present in the task environment.	22%
Fine Motor Coordination — The ability to manipulate small objects by moving one's fingers, wrists, and hands precisely in response to visual stimuli.	16%
Wrist-Finger Speed — The ability to move fingers, wrists, and hands as rapidly as possible where accuracy is less important than speed of movement.	14%
Detail Orientation — Skill at carrying out tasks thoroughly and accurately by attending to all of the small features of the task.	21%
Persuasiveness — Skill at influencing the behavior and/or attitudes of others by providing them with convincing information.	93%
Social Sensitivity — Skill at interpreting the motives, feelings, and behavior of others and responding appropriately to achieve a desired effect.	89%

APPENDIX E

VALIDITY GENERALIZATION RESULTS

FOR CLERICAL JOBS[a]

	Type of Criterion			
Type of Predictor	Rating	Ranking	Work Sample	Production Quality & Quantity[b]
General Mental Ability	.44	.66	.60	.31
Verbal Ability	.32	.52	.50	.22
Quantitative Ability	.40	.64	.55	.35
Perceptual Speed	.38	.45	.61	.37
Memory	.32	.35	.53	.34
Spatial/ Mechanical Ability	.24	.37	.42	.17
Psychomotor Ability	.27	.23	.54	.31

[a] Results reported are the mean estimated true validities.
[b] This is a composite of production quality and quantity.

These results were obtained from:
Nathan, B.R., & Alexander, R.A. (1985). The predictability and substitutability of criteria: A meta-analytic investigation. Unpublished manuscript.

VALIDITY GENERALIZATION RESULTS

FOR SALES JOBS[a]

Type of Predictor	Type of Criterion: Rating	Ranking	Work Sample	Production Quality & Quantity[b]
General Mental Ability	.22	.28	.43	.17
Verbal Ability	.14	.36	.50	.35
Quantitative Ability	.09	.18	.36	.13
Perceptual Speed	.05	.32	.09	.00
Memory	.18	.22	.28	.14
Spatial/ Mechanical Ability	.08	.13	.13	.03
Psychomotor Ability	.05	.08	.10	.01

[a] Results reported are the mean estimated true validities.
[b] This is a composite of production quality and quantity.

These results were summarized from all relevant published validity studies in the Journal of Applied Psychology and Personnel Psychology from 1918 to present. A complete bibliography is available upon request.

REFERENCES

Alpert, M., & Raiffa, H. (1968). A progress report on the training of probability assessors. Unpublished manuscript, Harvard University, Boston, MA.

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 508-600). Washington, DC: American Council on Education.

Becker, B.E. & Cardy R.L. (1986). Influence of halo on appraisal effectiveness: A conceptual and empirical reconsideration. Journal of Applied Psychology, 71, 662-671.

Bernardin, H.J., Morgan, B.B., & Winne, P.S. (1980). The design of a personnel evaluation system for police officers. JSAS Catalog of Selected Documents in Psychology, 10, 1-280.

Bhaskar, R., & Simon, H.A. (1977). Problem solving in semantically rich domains: An example from engineering thermodynamics. Cognitive Science, 1, 193-215.

Blaz, F., & Ghiselli, E.E. (1972). The mixed standard scale: A new rating system. Personnel Psychology, 25, 185-199.

Borgida, E., & Nisbett, R.E. (1977). The differential impact of abstract vs. concrete information on decisions. Journal of Applied Psychology, 7, 258-271.

Brandt, L.W. (1968). A note on the use of judges. The Journal of Psychology, 68, 261-266.

Bray, D.W., & Grant, D.L. (1966). The assessment center in the measurement of potential for business management. Psychological Monographs, 80 (17, Whole No. 625).

Brehmer, B. (1970). Inference behavior in a situation where the cues are not reliably perceived. Organizational Behavior and Human Performance, 5, 330-347.

Bruner, J.S., Goodnow, J.J., & Austin, G.A. (1976). A study of thinking. New York: Wiley.

Carroll, J.S., & Payne, J.W. (1976). The psychology of the parole decision process: A joint application of attribution theory and information-processing psychology. In J.S. Carroll & J.W. Payne (Eds.), Cognition and social behavior (pp. 13-32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Christensen-Szalanski, J.J., Beck, D.E., Christensen-Szalanski, C.M., & Koeopsel, T.D. (1983). Effects of expertise and experience on risk judgments. Journal of Applied Psychology, 68, 278-284.

Churchill, G.A., Jr., Ford, N.M., Hartley, S.W., & Waller, O.C., Jr. (1985). The determinants of salesperson performance: A meta-analysis. Journal of Marketing Research, 22, 103-118.

Cook, R.L., & Stewart, T.R. (1975). A comparison of seven methods for obtaining subjective descriptions of judgmental policy. Organizational Behavior and Human Performance, 13, 31-45.

Cornelius, E.T., DeNisi, A.S., & Blencoe, A.G. (1984). Expert and naive raters using the PAQ: Does it matter? Personnel Psychology, 37, 453-464.

Cornelius, E.T., & Lyness, K.S. (1980). A comparison of holistic and decomposed judgment strategies in job analyses by job incumbents. Journal of Applied Psychology, 65, 155-163.

Crocker, J. (1981). Judgment of covariation by social perceivers. Psychological Bulletin, 90, 272-292.

Crocker, J. (1982). Biased questions in judgment of covariation studies. Personality and Social Psychology Bulletin, 8, 214-220.

Cronbach, L.J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." Psychological Bulletin, 52, 177-193.

Cronbach, L.J., & Gleser, G.C. (1965). Psychological tests and personnel decisions. Urbana, IL: University of Illinois Press.

deGroot, A.D. (1966). Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.), Problem solving: research, method & teaching (pp.19-50). New York: Wiley.

Ebbesen, E.B., & Konecni, V.J. (1975). Decision making and information integration in the courts: The setting of bail. Journal of Personality and Social Psychology, 32, 805-821.

Ebbesen, E.B., & Konecni, V.J. (1980). On the external validity of decision-making research: What do we know about decisions in the real world? In T.S. Wallsten

(Ed.), Cognitive processes in choice and decision behavior (pp. 21-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), Formal representation and human judgment (pp. 507-514). New York: John Wiley & Sons.

Edwards, W. (1983). Human cognitive capabilities, representativeness, and ground rules for research. In P. Humphreys, O. Svenson, & A. Vari (Eds.), Analyzing and aiding decision processes (pp.79-91). New York: North-Holland Publishing Co.

Einhorn, H.J. (1971). Use of non-linear, non-compensatory models as a function of task and amount of information. Organizational Behavior and Human Performance, 6, 1-27.

Einhorn, H.J. (1972). Expert measurement and mechanical combination. Organizational Behavior and Human Performance, 7, 86-106.

Einhorn, H.J. (1974). Expert judgment: Some necessary conditions and an example. Journal of Applied Psychology, 59, 562-571.

Einhorn, H.J., & Hogarth, R.M. (1978). Confidence in judgment: Persistence of the illusion of validity. Psychological Review, 85, 395-416.

Einhorn, H.J., & Hogarth, R.M. (1981). Behavioral decision theory: Processes of judgment and choice. In M.R. Rosenweig & L.W. Porter (Eds.), Annual review of psychology (Vol. 32) (pp. 53-88). Palo Alto, CA: Annual Reviews Inc.

Faust, D. (1982). A needed component in prescriptions for science - empirical knowledge of human cognitive limitations. Knowledge-Creation Diffusion Utilization, 3, 555-570.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Fine, S.A., & Wiley, W.W. (1971). An introduction to functional job analysis. Washington, DC: W.E. Upjohn Institute.

Fischhoff, B. (1975). Hindsight = foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1, 288-299.

Fischhoff, B., & Byeth, R. (1975). "I knew it would happen": Remembered probabilities of once-future things. Organizational Behavior and Human Performance, 13, 1-16.

Fischhoff, B., Slovic, P., & Lichenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. Journal of Experimental Psychology: Human Perception and Performance, 3, 552-564.

Fried, Y., & Ferris, G.R. (1986). The dimensionality of job characteristics: Some neglected issues. Journal of Applied Psychology, 71, 419-426.

Goldberg, L.R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender Gestalt test. Journal of Consulting Psychology, 23, 25-33.

Goldberg, L.R. (1965). Diagnosticians versus diagnostic signs: The diagnosis of psychosis versus neurosis from the MMPI. Psychological Monographs, 79, (9, whole no. 602).

Goldberg, L.R. (1968). Simple models or simple processes? Some research on clinical judgments. American Psychologist, 23, 483-496.

Goldberg, L.R. (1971). Five models of clinical judgment: An empirical comparison between linear and nonlinear representations of the human inference process. Organizational Behavior and Human Performance, 6, 458-479.

Grinnell, M., Keeley, S.M., & Doherty, M.E. (1971). Bayesian predicitons oe faculty judgments of graduate school success. Organizational Behavior and Human Performance, 6, 379-387.

Gruppen, L.D., & Wolf, F.M. (1986). Expertise, efficiency, and the construct validity of patient management problems. Journal of Medical Education, 60, 878-880.

Guilford, J.P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Guilford, J.P., & Hoepfner, R. (1971). The analysis of intelligence. New York: McGraw-Hill.

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

Harris, R.J. (1975). A primer of multivariate statistics. New York: Academic Press.

Hirsh, H.R., Schmidt, F.L., & Hunter, J.E. (1986). Estimation of employment validities by less experienced judges. Personnel Psychology, 39, 337-344.

Hogan, J.C., & Fleishman, E.A. (1979). An index of the physical effort required in human task performance. Journal of Applied Psychology, 64, 197-204.

Hogarth, R.M. (1974). Process tracing in clinical judgment. Behavioral Science, 19, 298-313.

Hogarth, R.M. (1980). Judgment and choice: The psychology of decision. New York: John Wiley & Sons.

Jenkins, H.M., & Ward, W.C. (1965). Judgment of contingency between responses and outcomes. Psychological Monographs: General and Applied, 79, 1-17.

John, E.R., & Miller, J.G. (1957). The acquisition and application of information in the problem solving process. Behavioral Science, 2, 291-300.

Johnson, D.M. (1955). The psychology of thought and judgment. New York: Harper.

Johnson, P.E., Hassebrock, F., Duran, A.S., & Moller, J.H. (1982). Multimethod study of clinical judgment. Organizational Behavior and Human Performance, 30, 201-230.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.

Kane, J.S. (1984). Performance distribution assessment: A new breed of appraisal methodology . In H.J. Bernardin & R.W. Beatty (Eds.), Perfomance appraisal: Assessing human behavior at work (pp.325-347). Boston, MA: Kent Publishing.

Kaplan, M.F., & Kemmerick, G.D. (1974). Juror judgment as information integration: Combining evidential and nonevidential information. Journal of Personality and Social Psychology, 30, 493-499.

Kassin, S.M. (1979a). Base rates and prediction: The role of sample size. Personality and Social Psychology Bulletin, 5, 210-213.

Kassin, S.M. (1979b). Consensus information, prediction, and causal attribution: A review of the literature and issues. Journal of Personality and Social Psychology, 37, 1966-1981.

Lent, R.H., Aurbach, H.A., & Levin, L.S. (1971). Research design and validity assessment. Personnel Psychology, 24, 247-274.

Lichtenstein, S.C., & Fischhoff, B. (1976). Do those who know more also know more about how much they know? (Oregon Research Institute Research Bulletin No. 16). Eugene, OR: Oregon Research Institute.

Lichtenstein, S.C., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. Journal of Experimental Psychology, 89, 46-55.

Lindblom, C.E. (1964). The science of muddling through. In W.J. Gore & J.W. Dyson (Eds.), The making of decisions (pp. 155-169). New York: Free Press.

Mahoney, M.J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research, 1, 161-175.

Mahoney, M.J., & DeMonbreun, B.J. (1977). Psychology of the scientist: An analysis of problem solving bias. Cognitive Therapy and Research, 1, 229-238.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Miller, G.A. (1956). The magical number seven, plus or minus two. Psychological Review, 63, 81-97.

Mynalt, L.R., Doherty, M.E., & Tweney, R.D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. Quarterly Journal of Experimental Psychology, 29, 85-95.

Nathan, R.A. & Alexander, R.A. (no date). The predictability of criteria: A meta-analytic investigation. Unpublished manuscript.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Neisser, U. (1964). Visual search. Scientific American, 210, 94-102.

Neisser, U. (1967). Cognitive psychology. New York: Appleton-Century-Crofts.

Nisbett, R.E., & Ross, L.D. (1980). Human inference: Strategies and shortcomings of social judgment. Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

Northrop, L.C. (1976). The definition and measurement of judgment (Technical Memorandum 76-1). Washington, DC: Personnel Research and Development Center, U.S. Civil Service Commission-Bureau of Policies and Standards.

Olson, C.L. (1976). Some apparent violations of the representativeness heuristic in human judgment. Journal of Experimental Psychology: Human Perception and Performance, 2, 599-608.

Oskamp, S. (1965). Overconfidence in case-study judgments. Journal of Consulting Psychology, 29, 261-265.

Parry, M.E. (1968). Ability of psychologists to estimate validities of personnel tests. Personnel Psychology, 21, 139-147.

Payne, J.W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. Organizational Behavior and Human Performance, 16, 366-387.

Payne, J.W., & Braunstein, M.L. (1971). Preferences among gambles with equal underlying distributions. Journal of Experimental Psychology, 87, 13-18.

Payne, J.W., Braunstein, M.L., & Carroll, J.S. (1978). Exploring pre-decisional behavior: An alternative approach to decision behavior. Organizational Behavior and Human Performance, 22, 17-44.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 87, 1-28.

Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Peterson, C.R., & Beach, L.R. (1967). Man as an intuitive statistician. Psychological Bulletin, 68, 29-46.

Phelps, R.H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? Organizational Behavior and Human Performance, 21, 209-219.

Pitz, G.F., & Sachs, N.J. (1984). Judgment and decision: Theory and application. In M.R. Rosenweig & L.W. Porter (Eds.), Annual review of psychology (Vol. 35) (pp. 139-163). Palo Alto, CA: Annual Reviews Inc.

Primoff, E.S. (1974). How to prepare and conduct job element examinations. Washington, DC: U.S. Government Printing Office.

Reilly, R.R., Zink, D.L., & Israelski, E.W. (1984). Comparison of direct and indirect methods for setting minimum passing scores. Applied Psychological Measurement, 8, 421-429.

Roose, J.E., & Doherty, M.E. (1976). Judgment theory applied to selection of life insurance salesman. Organizational Behavior and Human Performance, 16, 231-249.

Roose, J.E., & Doherty, M.E. (1978). A social judgment theoretic approach to sex discrimination in faculty salaries. Organizational Behavior and Human Performance, 22, 193-215.

Sandoval, J., & Whelan-Miille, M.P. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48, 249-253.

SAS Institute Inc. (1985). SAS User's Guide: Statistics (5th ed.). Cary, NC: SAS Institute Inc.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. Psychological Bulletin, 66, 178-200.

Scarpello, V., & Campbell, J.P. (1983). Job satisfaction: Are all the parts there? Personnel Psychology, 36, 577-600.

Schein, V.E. (1978). Sex role stereotyping, ability, and performance: Prior research and new directions. Personnel Psychology, 31, 259-268.

Schmidt, F.L., & Hunter, J.E. (1977). Development of a general solution to the problem of valdity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F.L., Hunter, J.E., Croll, P.R., & McKenzie, R.C. (1983). Estimation of employment test validities by expert judgment. Journal of Applied Psychology, 68, 590-601.

Schultz, D. (1981). A history of modern psychology (3rd. ed.). New York: Academic Press.

Shanteau, J.C., & Anderson, N.H. (1972). Integration theory applied to judgments of the value of information. Journal of Experimental Psychology, 92, 266-275.

Simon, H.A. (1955). A behavioral model of rational choice. Quarterly Journal of Economics, 69, 99-118.

Simon, H.A. (1956). Rational choice and the structure of the environment. Psychological Review, 63, 129-138.

Simon, H.A. (1979). Information processing models of cognition. In M.R. Rosenweig & L.W. Porter (Eds.), Annual review of psychology (Vol. 30) (pp. 363-396). Palo Alto, CA: Annual Reviews Inc.

Slovic, P. (1969). Analyzing the expert judge: A descriptive study of a stockbrocker's decision processes. Journal of Applied Psychology, 53, 255-263.

Slovic, P. (1972a). From Shakespeare to Simon: Speculations and some evidence - about man's ability to process information (Oregon Research Institute Research Bulletin No. 12). Eugene, OR: Decision Reseaerch.

Slovic, P. (1972b). Limitations of the mind of man: Implications for decision making in the nuclear age. In H.J. Ofway (Ed.), Risk vs. benefit: Solution or dream (Report LA 48600000-MS). Los Alamos, NM: Los Alamos Scientific Lab.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavior decision theory. In M.R. Rosenweig and L.W. Porter (Eds.), Annual review of psychology (Vol. 28) (pp. 1-39). Palo Alto, CA: Annual Reviews Inc.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Response mode, framing, and information-processing effects in risk assessment. In R. Hogarth (Ed.), New directions for methodology of social and behavioral science: Question framing and response consistency (pp. 21-36). San Francisco: Jossey Bass.

Slovic, P., & Lichtenstein, S. (1968). Importance of variance preferences in gambling decisions. Journal of Experimental Psychology, 78, 546-554.

Slovic, P., & Lichtenstein, S (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 6, 649-744.

Slovic, P., & MacPhillamy, D.J. (1974). Dimensional commensurability and cue utilization in comparative judgment. Organizational Behavior and Human Performance, 11, 172-194.

Smedslund, J. (1963). The concept of correlation in adults. Scandinavian Journal of Psychology, 4, 165-173.

Smith, J.E. & Hakel, M.D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. Personnel Psychology, 32, 677-692.

Soelberg, P.O. (1967). Unprogrammed decision making. Industrial Management Review, 8, 19-29.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Svenson, O. (1985). Cognitive strategies in a complex judgment task: Analyses of concurrent verbal reports and judgments of cumulated risk over different exposure times. Organizational Behavior and Human Decision Processes, 36, 1-15.

The Society for Industrial and Organizational Psychology, Inc. The Industrial-Organizational Psychologist, 24, 75.

Tittle, C.K. (1982). Use of judgmental methods in item bias studies. In R.A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore, MD: The Johns Hopkins University Press.

Trattner, M., & Fine, S.A., & Kubis, J.R. (1955). A comparison of worker requirement ratings made by reading job descriptions and by direct job observation. Personnel Psychology, 8, 183-194.

Treiman, D. (1979). Job evaluation: An analytic review. An interim report to the Equal Employment Opportunity Commission. Committee on Occupational Classification and Analysis, National Research Council. Washington, D.C.: National Academy of Science.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 6, 105-110.

Tversky, A., & Kahneman, D. (1972). Anchoring and calibration in the assessment of uncertain quantities (Oregon Research Institute Bulletin No. 12). Eugene, OR: Decision Research.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458.

Upshaw, H.S. (1975). Judgment and decision processes in the formation and change of social attitudes. In M.F. Kaplan and S. Schwartz (Eds.), Human judgment and decision processes (pp. 201-228). New York: Academic Press.

U.S. Equal Employment Opportunity Commission. (1978). Guidelines on employee selection procedures. Federal Register, 35, 12333-12336.

Vernon, P.E. (1950). The structure of human abilities. London: Methuen and Co.

Wallsten, T.S. (1980). Processes and models to describe choice and inference behavior. In T.S. Wallsten (Ed.), Cognitive processes in choice and decision behavior (pp. 215-237). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ward, W.C., & Jenkins, H.M. (1965). The display of information and the judgment of contingency. Canadian Journal of Psychology, 19, 231-241.

Wason, P.C., & Johnson-Laird, P.N. (1972). Psychology of reasoning: Structure and content. London: Batsford.

Wiggins, N., & Hoffman, P.J. (1968). Three models of clinical judgment. Journal of Abnormal Psychology, 73, 70-77.