A META-ANALYSIS ON THE VALIDITY OF PERSONALITY TESTS USED FOR PERSONNEL SELECTION.

A thesis presented in partial fulfillment
of the requirements for the degree of

Master of Science in Psychology

at Massey University

Eugene Ng

1990

ABSTRACT

For decades, personality tests have been commonly used as one of the tools for personnel selection. However, through the use of various validity generalization techniques, researchers have claimed that they have very poor validity in predicting job performance. These claims were made by Guion and Gottier (1965), Ghiselli (1973), and Schmitt, Gooding, Noe and Kirsch (1984). Each of these researchers used different statistical techniques and methodologies to reach their conclusions. The latest study by Schmitt et al. (1984), used a statistical validity generalization technique called meta-analysis. Based on data collected from only two journal publications they claimed that personality tests had a validity of .15. The present study tested the conclusions of the Schmitt et al. (1984) study, by re-analysing the same data using a more accurate meta-analysis technique and by incorporating a larger data base. In addition to this, any new data from 1952 up to 1990 was included in an overall analysis to find out the current validity of personality tests. A flexible coding technique which interacted with a computerised data base allowed any combination of data to be separately analysed. This made it possible to discover which types of personality tests worked best in differing situations such as different sample types and criterion measures. Results of the Schmitt et al. (1984) re-analysis showed that by correcting coefficients for unreliability, the overall validity was significantly higher than the Schmitt et al. (1984) result. A separate analysis revealed that vocational tests had the highest validity of the six personality test types. The sample-types with the highest validities were Supervisory and Skilled workers. The best criterion-types were in the "Other" category whereby measures were developed specifically for the type of job. The overall analysis incorporating 38 years of research showed that personality tests had a validity of .22. This was significantly higher than the figure quoted by Schmitt et al. (1984). Results showed that personality tests in their present state are generally poor predictors of job performance, however when they are modified to become more job specific, their validity improves. It is suggested that in the future, personality tests should be specifically designed for the purpose of personnel selection and for specific jobs.

ACKNOWLEDGEMENTS

I wish to thank my supervisor Dr Mike Smith for his encouragement, faith and advice throughout this study. Many thanks also to my friends in the Psychology Department who were always full of encouragement and support when things got tough. Finally my most sincere thanks and appreciation to two of my closest friends Eddie and Margaret who had to put up with my moans and groans whilst also fulfilling the role as my parents.

TABLE OF CONTENTS

LIST OF APPENDICES

APPENDIX

LIST OF TABLES

TABLE

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.01 The Origin of Personnel Psychology

The development of industrial psychology began in the World War I era which saw the start of the use of psychological testing in industrial settings. Much of this development was concerned with the concept of looking at worker individual differences in an effort to understand personnel problems. In the following years, industrial psychology began establishing itself as a personnel-oriented field, however, its use in industry was still limited.

It was during the second World War that industrial psychology saw more widespread use and application due to the increased sophistication of warfare and the urgent need for rapid mobilization, (Howell & Dipboye, 1982). Throughout this period, personnel training, selection, and placement were regarded as some of the most important focal areas in applied psychology. These applications formed the basis of modern personnel psychology.

Personnel Psychology is defined as a general label for that aspect of industrial and organizational psychology concerned with: a) the selecting, supervising and evaluating of personnel, and b) a variety of job related factors such as morale, personal satisfaction, management-worker relations, counselling and so forth, (Reber, 1985).

In almost every area of personnel psychology there is a monetary utility applied directly to either the training of personnel or their on-the-job performance. Failure in either of these areas costs the organization a substantial amount of money in the form of direct decreases in production or lost opportunity costs.

1.02 Psychological Testing in Personnel Selection

The objective of any personnel selection procedure is to select high performing employees that will maximize benefits to the organization. In the past this procedure has generally been a hit and miss affair. One of the most popular ways to increase the "hit-rate" of selecting high performing employees is to use psychological tests that can accurately predict the performance of personnel in particular situations. The logic for developing and using psychological tests for personnel decisions is a simple one. The major assumption is that different individuals have different probabilities of success in different jobs. These probabilities are largely dependent on the individual characteristics and abilities that are unique to each person. The application of a psychological test to identify these individual characteristics for selection purposes, utilizes the assumption that people with certain characteristics will have a greater probability of success in a particular job than people without those characteristics.

Clark Hull in 1928, defined a psychological test as : *The measurement of some phase of a carefully chose sample of an individual's behaviour.* This definition is broad enough to identify personnel selection devices other than the standard paper-and-pencil test (Landy, 1989, p54).

Tests that can be administered as personnel selection devices can be categorised into three main groups according to their content (McCormick & Tiffin, 1974). The first group includes tests of basic human abilities, such as cognitive abilities and psychomotor skills. These are used to determine whether individuals possess the capacity to learn a given job if they are given adequate training.

The second major group includes tests of achievement that measure job-specific abilities such as typing or operating machinery. Certain tests in this group are also referred to as work-sample tests because they provide a sample of a working situation and assess the incumbent's performance in dealing with the tasks of that situation.

An example of a work-sample test would be the In-Basket Test whereby incumbents are presented with a series of "practical tasks" such as answering phones. The incumbents are then assessed on how they deal with the phone call and what the outcome was. The incumbents would then move on to the next task to be dealt with.

The third major group of tests are personality and interest tests. These are designed to measure personality characteristics or patterns of interests of individuals on the assumption that these characteristics and interests may be related to job performance.

1.03 Types of Personality Tests

The use of personality tests in personnel selection represents a marked contrast to the other types of tests. Instead of selecting people who meet or exceed a predictor cut-off on a single variable, interest and personality tests have often utilized a multivariate procedure in which a set of interest or personality dimensions scores are compared with a norm, (Muchinsky, 1986).

The number of available personality tests runs into several hundred. They generally fall into three main categories- pencil and paper personality questionnaires, interest inventories and projective techniques. The pencil and paper personality questionnaires are in fact inventories which contain a series of statements or questions relating to behaviours, attitudes and feelings. The responses are then tallied up and scored by comparing them with existing data and norms.

Some of the more common pencil and paper tests used for selection purposes are the Minnesota Multiphasic Personality Inventory (MMPI), the California Psychological Inventory (CPI), and the 16 Personality Factor Questionnaire. The California Psychological Inventory for example is essentially derived from the MMPI and is very similar except that where the MMPI was designed for "abnormal people", the CPI was designed for a "Normal population".

The CPI provides scores on eighteen components over four classes, as follows:

Class I:	Measures of poise, ascendency, and self assurance.
Class II:	Measures of socialization, maturity and responsibility.
Class III:	Measures of achievement potential and intellectual efficiency.
Class IV:	Measures of intellectual and interest modes.

Interest inventories tap the strength of a person's interests in such things as hobbies, recreation, leisure time activities, sport and jobs. This is done in the hope that certain people who are more interested in certain things may perform the job better than other people. Examples of the more common interest inventories are the Strong Vocational Interest Blank (SVIB), the Kuder Preference Record and the Vocational Preference Inventory. The SVIB for example, was primarily developed for use in vocational guidance counselling and determines whether or not the subject's pattern of interests agrees with the interest patterns of people in each of other various occupations. Scoring is based on grades ranging from A to C. The higher the score for any "occupation", the closer the person's interests are related to people who are successful in that occupation.

The final group of personality tests are the projective techniques. Some of the more common projective measures used are the Rorschach Inkblot Technique and the Holtzman Inkblot Technique. The use of these tests involves the subject being presented with an intentionally ambiguous stimulus such as an ink-blot picture, and being asked to give his or her interpretation of what they "see" in the picture.

Scoring is subjective in the form of an expert who interprets the individual's replies and patterns of responses and makes an assessment of the individual. These tests are unreliable and are used more commonly in a clinical setting rather than the occupational setting.

All the personality tests mentioned above have been used in industrial psychology and personnel selection for decades and by innumerable organizations. The actual utility of information gained using these and other personality measures in personnel selection will be systematically reviewed in this study.

1.04 Test Validity and Reliability

The process of reviewing the utility of the tests used to select personnel involves assessing their validity and reliability in the workplace. The basic premise is to determine the strength of the relationship between test performance and job performance. For this purpose a criterion-related validity coefficient can be derived from test performance measures and measures taken on-the-job.

There are two types of criterion-related validity; concurrent validity and predictive validity. Concurrent validity is an estimate of the accuracy in measuring current performance. It is obtained by correlating performance on the chosen criterion with current test results. Predictive validity is an estimate of the measurement's accuracy in measuring future performance or behaviour. It is derived by correlating performance on an established test with performance on a chosen criterion (Conrad & Maul, 1981). In the case of personnel selection, predictive validity provides the most relevant information about future job performance, whereas concurrent validity would be more appropriate for comparing the performance of current workers to establish the effects of an independent variable such as job experience on current job performance.

1.05 Sources of Validity Information

In a study by Monahan and Muchinsky (1983), 394 personnel selection studies published in *Personnel Psychology* between 1950 and 1979 were analysed according to selected aspects of the research design. This gave an indication of the research developments and trends that have occurred over the past three decades in the area of personnel selection.

Perhaps an indication of the amount of research published in the area of personnel selection can be given by the vast number of validity studies reported between 1953 and 1960. There are a number of reasons for this trend. One is that between these years, *Personnel Psychology* incorporated the *Validity Information Exchange* (VIE) which was a section where researchers could convey their validation findings in a brief report format (see figure.1)

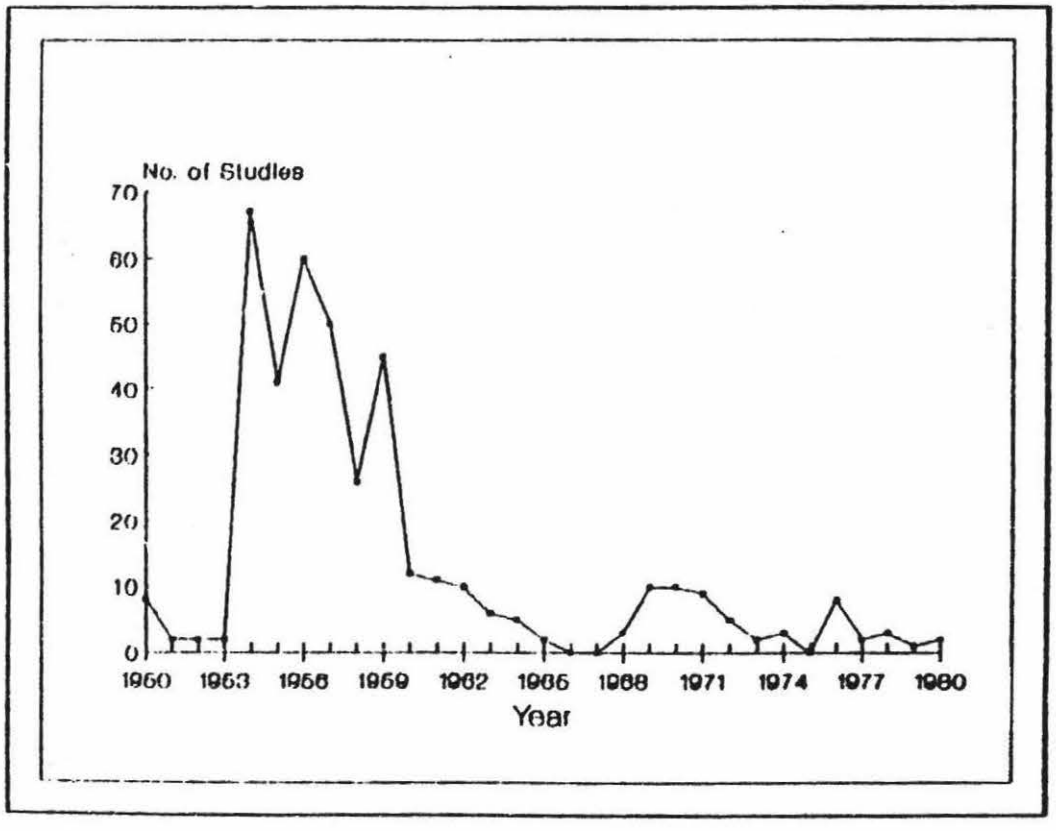


Figure 1. Number of validity studies published in Personnel Psychology between 1950 and 1980. (after Monahan & Muchinsky, 1983).

The VIE remained a major source of validity reports until 1960, after which a steady decline of reports led to the discontinuance of the VIE in 1965. Apart from the demise of the VIE, there are other reasons attributable to the decline in research reports. One of those reasons may have been the increasing social and legal pressures brought against psychological testing during that period. There also has been a change in the research interests of industrial psychologists as more social, and environmental factors were incorporated into the discipline.

The 1980's saw a renewed interest in personnel selection research with the use of more accurate research tools such as Meta-Analyses and the development of fast microcomputers. Personnel selection research has been characterized by a heavy reliance on subjective criterion measures and small sample sizes (Monahan & Muchinsky, 1983). With the emphasis now on validity and reliability of research designs, larger sample sizes are being used along with a trend toward more objective criteria.

In order to find out more about the fidelity of the use of personality tests in the applied setting, some form of cumulative study would need to be performed. A cumulative study would provide a global account in one form or another, of research findings relevant to a particular area. Such cumulative studies have been performed in the past in the form of narrative reviews and meta-analyses.

1.06 Early Reviews of Personality Tests

The first major review of the validity of personality tests
used for personnel selection was carried out by Ghiselli and Barthol (1953). Their results showed a wide range of reported validities ranging from very low and negative, to moderately high. A more comprehensive narrative review was carried out a decade later by Robert Guion and Richard Gottier (1965).

In the decade between those two studies there had been such a proliferation of new personality measures that Dunnette (1962) was moved to urge a moratorium on construction of new tests until those already available were better utilized.

The main conclusion of Guion and Gottier's (1965) narrative review was that "there is no generalizable evidence that personality measures can be recommended as good or practical tools for employee selection. The best that can be said is that in *some* situations, for *some* purposes, *some* personality measures can offer helpful predictions (Guion & Gottier, 1965).

In the Guion and Gottier (1965) study, the definition of a personality measure was broadened to include measures of interest. Their reason for doing this was that interest and temperament measures are used for a common role: the measurement of variables presumed to be associated with motivation. They mention that the distinction between interest variables and variables measured in the more conventionally named personality tests is one which is very difficult to make either at an empirical or at a theoretical level.

1.07 Traditional Methods Used for Cumulating Results.

The narrative review technique used by Guion and Gottier (1965) can best be described as a "qualitative literature review" where the reviewer collects the results of a number of studies and takes each one at face value in an effort to integrate the entire collection in an overall conclusion. This technique has several limitations in terms of the generalizability of the conclusions particularly if there are a large number of studies being reviewed. In this case, the studies will usually not be comparable in design, measures, criteria etc., and findings will typically vary across studies, (Hunter & Schmidt, 1990). The ability for the human mind to be able to consciously account for the variances and hence the associated fluctuations in the quality of each study becomes increasingly limited when the number of studies increases.

Glass (1976) mentions that "*one can no better grasp the results of hundreds of studies in the traditional narrative review than one can grasp the sense of 500 test scores without the aids for organizing, depicting and interpreting data.*" (Glass 1976, p3).

Cooper and Rosenthal (1980) have shown that even when the number of studies reviewed is as small as seven, reviewers who use narrative-discursive methods and reviewers who use quantitative methods, usually reach different conclusions.

One method developed to overcome the problem of large numbers of studies is the traditional vote counting method. This approach involves classifying each study into one of three categories, either significantly positive, significantly negative or undecided. The number of studies falling into each category is then tallied and whichever category has the greatest number of votes is the winner. This method assumes that all studies have the same characteristics and completely ignores differing sample sizes.

Hedges and Olkin (1980) have pointed out that if there is a true effect, then in any set of studies in which mean statistical power is less than about .50, the probability of a false conclusion using the vote counting method increases as the number of studies increases! Therefore, the traditional vote counting method is fatally flawed both statistically and logically and does not really overcome the problem of integrating large numbers of studies.

Cook and Leviton (1980) reviewed the traditional methods of cumulating research findings and compared them to meta-analytic techniques. They state that counting the incidence of significant results is overly conservative because results which are in the right direction but fail to reach the chosen significance level, will be counted as negative results.

The narrative review and the vote counting method are both considered as qualitative literature reviews. Qualitative methods serve to cumulate and assess each piece of research separately. However, each piece of research has its own irregularities and variances. In order to be able to integrate results from different studies, the irregularities and variances must be controlled.

1.08 Artifacts and Sources of Error

To describe the many forms of error and variance in research results, Hunter and Schmidt (1990) use the term "study artifacts". Sampling error is one of the main sources of artifactual variance. This is because study validity varies randomly with differing population types and sizes. Larger sample sizes are more reliable and therefore should be weighted accordingly.

Error of measurement is another source of artifact variance. It occurs in both the dependent and independent variables. In the case of the dependent variable, validity will be systematically lower than true validity to the extent that job performance is measured with random error. This problem stems from the use of poor measurement techniques and data recording errors. In the case of the independent variable, the validity for a test will be lower than true validity, to the extent that the selection rules assume that employees have a lower variation in the predictor than applicants.

The third artifact is range variation, which occurs because correlations from different studies will have different variances on the independent variable. In order to compare these correlations the variances must be controlled. When this occurs, study validity will be systematically lower than true validity to the extent that selection policy causes incumbents to have a lower variation in the predictor than is true of applicants.

The fourth artifact is the result of errors such as inconsistent coding, exceeding parameter rules, interpretation and typographical errors. These errors are probably the most commonly made and are known simply as reporting errors.

The fifth artifact is deviation from perfect construct validity. This occurs in the dependent variable and is related to the degree to which the actual criteria matches the ultimate criteria. Criterion deficiency or contamination affects the accuracy of the reported validity coefficient. This problem highlights the importance of proper criterion development.

Another source of error occurs through the use of dichotomized dependent and independent variables. Often, variables associated with interviewing or job performance measures are dichotomized into "accept" and "reject" categories, or "more than" and "less than" responses. These responses are inaccurate and are difficult to measure, especially when an attempt is made to produce a quantitative summary.

Variance due to extraneous factors is the final artifact. This stems from individual variations in each incumbent's experience or background that may have an affect on the outcome at the time of the assessment.

Failure to account for any of these artifacts leads to errors in the estimation of statistical significance and variance across studies. Although controlling for every single artifact may not yet be feasible, the objective would be to control for as many as possible.

1.09 Meta-Analysis vs Traditional Methods

One statistical technique which has been used over the last two decades is a method of validity generalization known as Meta-Analysis. Meta-analytic techniques, although controversial, are very powerful tools in modern research. They have the ability to overcome the earlier problems of large numbers of studies, subjective ratings and account for nearly all the artifacts mentioned above. Compared to the narrative review and the traditional vote-counting method, meta-analysis appears to be more systematic and objective in reaching a conclusion.

Meta analysis is first a qualitative literature review. Each piece of research is considered in turn, then the articles are combined in a quantitative review and the results are synthesised into a final conclusion. Meta-analysis is not a single technique, but rather a flexible set of techniques that can be adapted to the question at hand, provided that enough information is available in the reports of research (Cook & Leviton, 1980).

Early applications devised before the 1960's were of three main types. The first generally resulted in an estimate of the average correlation found in all the studies summarized. The second resulted in a correlation between some characteristic of the studies and the correlation found in the studies. The third type simply correlated data obtained from each study with other data or characteristics obtained from within the studies.

From the 1960s onwards, meta-analytic techniques were developed in more detail by people such as Robert Rosenthal and Gene Glass who coined the term "meta-analysis" in his two papers in 1976. Since the late 70s there have been literally hundreds of published and unpublished meta-analyses, (Rosenthal, 1984).

1.10 The Hunter and Schmidt Methodology

Over the past decade, there have been major methodological advancements made in the area of meta-analysis. Two researchers who were the major forces behind these changes were John Hunter, Professor of Psychology at Michigan State University and Frank Schmidt, Professor of Human Resources at the University of Iowa.

In association with Gregg B. Jackson, Hunter and Schmidt published a book on meta-analysis emphasising the importance of correcting results for artifacts such as unreliability and error of measurement on both the dependent and independent variables (see Hunter, Schmidt & Jackson, 1982).

The method used by Hunter and Schmidt for conducting a meta-analysis has been employed in the majority of recent of meta-analytic studies. Fisher and Gitelson concisely state that: *"the Schmidt-Hunter meta-analysis is based on the idea that much of the variation in results across samples or studies is due to statistical artifacts and methodological problems rather than to truly substantive differences in underlying population correlations"*, (Fisher & Gitelson, 1983, p320).

The Hunter and Schmidt method is conducted in three stages. The first involves calculating the mean correlation across studies weighted according to the sample sizes used in each study. The purpose for this procedure is that correlations from studies with large sample sizes are more reliable and are more likely to represent the true population value than are correlations from studies with small sample sizes.

The second stage involves calculating the total variance of sample correlations around the sample weighted mean. From this, the variance attributable to artifacts is then able to be calculated (stage three), and subtracted from the total variance. Theoretically when the variance attributable to artifacts is subtracted from the total variance, the remaining unexplained variance is often very small, indicating that apparently inconsistent results across studies are not truly inconsistent, but occur only because of statistical artifacts.

When correcting for artifacts, personnel selection is a special case according to Hunter and Schmidt (1982, 1990), because the predictor is used in an imperfect form. The cause of this stems from the fact that job performance data is only available on incumbents who are currently working. The incumbent population is different from the applicant population because poor workers are usually fired or quit voluntarily. Therefore, the relevant population correlation used to assess the practical impact of the test should be corrected for error of measurement in the dependent variable (criterion), but not in the independent variable (predictor).

1.11 Meta-Analysis in Personnel Selection.

In the past two decades there have been three major meta-analyses carried out in the area of personnel selection. The first one was by Ghiselli (1973) and the other two were simultaneously published by Hunter and Hunter and Schmitt, Gooding, Noe and Kirsch in 1984. The methodology used to carry out these analyses varied somewhat resulting in inconsistent conclusions among the three studies. The Ghiselli (1973) study employed the same technique used in his (1966) study and is simply an updated version of this previous study. The Ghiselli (1966) meta-analysis corrected only for sampling error which, as the author himself points out, means that the reported weighted coefficients are under-estimations of the true validity.

The most recent study was by Schmitt, Gooding, Noe, and Kirsch (1984), who conducted a meta-analysis on predictors of job performance. They reported a low overall validity coefficient of .15 for personality tests including measures of interest. However, their results have been criticised by many as being inaccurate because only sampling error was accounted for and major artifacts such as error of measurement on both the dependent and independent variables were not corrected in the analysis.

Schmitt et al. (1984) claim that, after sampling error variance was corrected, much unexplained variance remained and that their findings are inconsistent with earlier validity generalization work. In a review article by McDaniel, Hirsh, Schmidt, Raju and Hunter (1986), it was stated that the Schmitt et al. (1984) claims were not correct and were misleading. McDaniel et al. (1986) stated that the Schmitt et al. (1984) results were the product of using larger average sample sizes from a narrow scope of research articles. The main reason for the larger sample sizes was that the Schmitt et al. (1984) data base consisted only of published studies. In fact only studies published in two American journals namely, *Personnel Psychology* and *the Journal of Applied Psychology* were included. All were also published after ordinary small-sample validity studies had ceased to be considered worthy of publication by these journals.

The Schmitt et al. (1984) study was not the only one to restrict itself to these two journals. The narrative review by Guion and Gottier (1965) also used studies published only in *Personnel Psychology* and *Journal of Applied Psychology.* The authors of these studies however, do state their reasons for limiting their data base parameters. For example, the journal *Personnel Psychology* was chosen by Monahan et al. (1983) because over the years the journal has published articles by researchers from several nationalities, as well as from academic, industrial and governmental research sectors. It is one of the leading journals in the area of industrial and organizational psychology and the journal has been in existence long enough (since 1948) for a longitudinal analysis of published research to be made.

They also chose the *Journal of Occupational Psychology* because one of the most prevalent topics is personnel psychology with some 31% of its articles having been published in this area. It is significant however, that there is seemingly little consistency in the journals used to conduct meta-analyses.

Admittedly the effect of incorporating a small number of articles that have not been published or have been published in other journals will probably have no effect on the global conclusion of a meta-analytic study. It may however have an effect on the qualitative accuracy of the final true validity coefficient. Researchers such as Rosenthal (1979) and Arkin, Cooper, and Kolditz (1980) stress the importance of a thorough search of the literature.

In a paper by Rosenthal (1979), focusing on the area of selective publications, it is mentioned that for any given area, one cannot tell how many studies have been conducted but never reported. He outlines what is now known as the "file drawer problem", which stems from the fact that most journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results.

Given this type of selective publication, the credibility and generalizability of cumulative research methods using only published journal articles, in particular the traditional vote-counting method, would be questionable to say the least.

As more reports are being published, simultaneously more unpublished reports are cumulated. This is occurring to the extent that we can no longer afford to ignore them if a true generalizable global analysis is to be performed.

All published and unpublished reports should be examined, as editorial policies in journals may bias published reports towards confirming replications or statistically significant studies. The difficulty faced here is the problem of how to obtain unpublished reports. The only solution is to conduct a thorough search and to inform people about what type of information is being sought. With the advent of electronic mail, facsimiles and modems, the literature search is much easier to carry out today than it was a decade ago, as communication to most universities and data sources can be achieved relatively quickly and inexpensively.

1.12 Personality Test Validity

In 1984, Hunter and Hunter applied their meta-analytic formulas to past meta-analyses and major comparative reviews that had already been done such as Ghiselli (1966, 1973), Vineberg and Joyner (1982) and Reilly and Chao (1982).

During this time, the Schmitt et al. (1984) study was being carried out and so missed being included in the Hunter and Hunter (1984) study. However, in a study by Hunter and Hirsh (1987), the meta-analysis by Schmitt et al. (1984) was re-calculated using Hunter and Hunter's formulas for correcting error of measurement and attenuation. Their comment on the Schmitt et al. (1984) study, was that they made no correction for unreliability or range restriction hence validity was not accurately reported because the proper correction for these artifacts was not made.

Schmitt et al. (1984) thought they had found inconsistencies between their results and those of Hunter and Hunter (1984). However, Hunter and Hirsh's (1987) re-calculations of the Schmitt et al. (1984) data showed that those inconsistencies were due to the absence of corrections for error of measurement or attenuation.

Hunter and Hirsh (1987) recalculated the data used in the Schmidt et al. (1984) study using their own formulas and corrected for error of measurement in performance ratings by using a constant reliability coefficient of 0.60 for personality tests. The results of this recalculation increased the overall validity coefficients of personality tests used for predicting performance ratings by 31 percent and the validity for training criteria was increased by 13 percent. The validity for status change and promotion was increased by only 3 percent.

The average validity for these three criteria in the Hunter and Hirsh (1987) study was .19. This was a 36 percent increase over the overall validity coefficient of .15 quoted by Schmitt et al. (1984). Given these results it appears that the overall validity coefficient of .15 quoted for personality tests by Schmitt et al. (1984), was a substantial underestimate of the true validity.

Table 1

Ghiselli (1973), Schmitt et al. (1984), and Hunter and Hirsh (1987) validities for personality as a function of sample type and overall means.

Sample Type	***Ghiselli(1973)***	***SNGK (1984)***	***H & H (1987)***
Professional	.05	.16	-
Supervisor	.34	.17	-
Clerical	.20	-.02	-
Skilled Labour	.22	-	-
Unskilled Labour	.16	.06	-
Sales	.32	-	-
Overall	.22	.15	.24

Table one shows a comparison between the three major statistical analyses carried out on personality tests. The Ghiselli (1965) results are the mean unweighted correlation coefficients which were calculated in that study. The Schmitt et al. (1984) results are sample weighted correlation coefficients and the Hunter and Hirsh (1987) results are mean sample weighted correlation coefficients corrected for unreliability. The overall statistics given at the bottom of table one were calculated using the same method as were used in each study.

The results shown in table one are descriptive rather than comparative. Any form of comparison between the three studies would need to consider that there will be large inconsistencies in coding criteria, categorization of sample types and most of all, the methodology used. The differences between the Ghiselli (1965) results and the Schmitt et al. (1984) results are quite substantial. When all three overall coefficients are looked at, it appears that simply averaging the correlations (Ghiselli's study) gives an inflated result, applying sample weighting (SNGK's study) gives an underestimated result, and sample weighting and correcting for unreliability boosts that underestimate by a considerable amount. The last two statements are probably true, but the first statement is probably not.

Unweigh'ed mean correlations are dependent on the magnitude and frequencies of correlations. For example, if there are two studies, with one having a validity of .60 and a sample size of 50 000 and the other having a validity of .02 and a sample size of 50, the mean validity for both studies will be .31 even though the first study had a sample size 1000 times larger than the other. However if they were weighted according to sample size the mean validity would be .599 ! Therefore, Ghiselli's results were dependent on the magnitudes and frequencies of the validities reported in his data.

1.13 The Current Situation.

At present there are still conflicting results and conclusions about the validity of personality tests used for personnel selection. Validities ranging from .15 to .65 have been quoted by a number of researchers. (Schmitt et al., 1984, Ghiselli, 1973, Hunter & Hirsh, 1984). The latest published comment is from an article in *The Psychologist* (March 1990), where one writer claims that a validity of .15 for personality tests *" is nonsense and professionally damaging nonsense at that!"*

Muchinsky (1986), after reviewing the results of Guion and Gottier (1965), Ghiselli (1973) and Schmitt et al. (1984), concluded that, for the most part, the reported validity coefficients for interest and personality tests have been unimpressive. He also stated that in recent years there seems to have been a revival of interest in personality testing.

Bernardin and Bownas (1984) have advocated that we re-examine the role of personality measures in personnel selection. When a universally acceptable procedure for validity generalization studies can be agreed upon, perhaps then the results and conclusions will be less conflicting. The Hunter and Schmidt meta-analytic technique appears to be a procedure which gives the best possible estimate of true validity.

1.14 The Benefits of Computerized Meta-Analysis

One of the objectives of the present study is to highlight the benefits of meta-analysis over traditional procedures. The main purpose of the present study is to carry out a meta-analysis of the validity of personality tests used for selection. The meta-analytic techniques of Hunter and Schmidt (1990) will be employed to perform the data analysis. The entire analysis will be carried out on an I.B.M. PC using a meta analysis program written in GW-Basic version 3.2.

The ability to computerize the meta-analytic procedure has several major advantages over the conventional hand calculated meta-analyses and other validity generalization procedures. The greatest advantage is the ability to store and manipulate large volumes of data. The flexibility of computerization allows the user to instantly analyse any number of combinations of data. This flexibility is limited only by the amount of detail the researcher chooses to use in the data coding stage. Any new studies can be entered into the data file instantly so that the outcomes of the analysis can be updated as soon as the new information is entered.

Any advancements and future modifications that may need to be made to the formulas used, can quickly be made without having to rewrite the entire program. The potential benefits and utility of a meta-analytic statistical software package are enormous as it develops into an even more powerful statistical tool that can be quickly and inexpensively applied to almost any research area.

1.15 Statement of the Hypotheses

The present study has three major domains, the first concerns the limited use of only two "selected" journals by Schmitt et al. (1984) in their 1984 meta-analysis. Because a meta-analysis is supposed to be a cumulative account across all findings, any limitation of data collection defeats the purpose of the analysis. The present study will re-calculate the Schmitt et al. (1984) data using the Hunter and Schmidt method which corrects for unreliability. In addition to the Schmitt et al. (1984) data, will be any relevant data found in other journals. Furthermore, the data search will be extended to cover the years between 1982 and 1990, a period which no previous cumulative research has covered.

The second major area looks at the different factors and criteria involved in the selection and coding of data used to perform a meta-analysis on personality tests for personnel selection. The data will be coded so as to allow different study characteristics to be tested. The effects of selective data collection can then be

demonstrated by excluding data with characteristics such as studies using only professional subjects or using only performance as a criterion.

The final major area is concerned with the claims made about the predictive validity of personality tests by Guion and Gottier (1965), Ghiselli (1973), Schmitt et al. (1984) and Hunter and Hirsh (1987). The first three authors conclude that personality tests have very poor predictive validity. Hunter and Hirsh (1987) do not however, claim that personality tests have a high predictive validity, they maintain that the validity of .15 presented by Schmitt et al. (1984) is too low an estimate and that their higher estimates using corrections for criterion and predictor unreliabilities are more accurate.

The present study attempts to provide support for the Hunter and Hirsh (1987) conclusions and challenges the Schmitt et al. (1984) conclusions by hypothesizing that by using the Hunter and Schmidt (1990) meta-analytic technique and by extending the scope of journals used, an overall validity coefficient for personality tests will be significantly higher than the Schmitt et al. (1984) overall result of .15.

If the differences are significant in this study then it may be assumed that:

A) Results and conclusions can vary greatly according to the decision rules used to categorise the data.

B) Results and conclusions can vary greatly by the method used to calculate mean validity coefficients.

C) Results and conclusions can vary greatly according to criteria used to select study data.

If these three points prove to be valid then the implications are that studies using meta-analytic techniques need to be in a more compatible format. Thorough literature reviews and standardized coding and data selection processes should be used if conclusions are to be compared or contrasted with other studies.

CHAPTER TWO

METHOD

The first stage of the meta-analysis was to define the parameters of the present study. The previous studies by Schmitt et al. (1984) and Guion and Gottier (1965) were used as guidelines. These parameters were:

1. All articles used were published between 1952 and 1990. The present analysis covered 38 years of personality research.

2. The classification of personality measure has been broadened to include measures of interest in the same way as the definition used by Guion and Gottier (1965). This means that the targeted studies are those focussing on personality measures. Studies incorporating interest measures, however, were also included if they had already been classified as personality measures in past validity generalization studies.

3. Studies that were included in the present analysis had to conform to the specified "Study Characteristics" mentioned below in order to meet the statistical and methodological requirements of the study.

2.01 Data Collection

The data collection phase was performed in three steps. The first step was to obtain the reference lists for both the Guion and Gottier (1965) review and the Schmitt et al. (1984) meta-analysis. The Guion and Gottier study provided references from the *Journal of Applied* Psychology and *Personnel Psychology* for studies covering the twelve year period between 1952 and 1963. The Schmitt et al. (1984) study covered research from the same two journals between the years of 1964 and 1982.

These two studies combined, provided references from two major journals in the area of personnel selection over a thirty year time span (1952-1982). Using these references as a starting point, the primary task was to cover the years from 1982-1990 in *Personnel Psychology* and the *Journal of Applied Psychology*, and to search for articles in any other journals that may have had relevant data for the period 1952-1990, a 38 year period.

The second step in the data collection phase was to perform a thorough search of the literature on personality tests used in occupational and related settings. Both a manual reference and a computerized search of on-line data bases were conducted (Psych-info) using the DIALOG Information Retrieval Service. The on-line search was conducted using combinations of over 60 selected key words and was limited to articles published after 1965. The result of the search yielded 97 references, however only half of those were related to personality testing in an occupational setting.

Following the on-line search, a manual search of relevant journals was conducted to ensure full coverage of the literature. All volumes from *seven* major occupational and general psychological journals were searched. Those journals were:

Journal of Applied Psychology
Journal of Occupational Psychology
Personnel Psychology
Psychological Bulletin
Organizational Behaviour and Human Decision Making
International Review of Applied Psychology
Journal of Personality

The third stage of the data collection phase was the selection and recording of the articles. For an article to be included, it had to meet certain study criteria to ensure that the statistics reported in each study and the statistical requirements of the meta-analytic formulas were compatible.

The study criteria required were:

1. Only studies reporting Pearson *r* correlation coefficients could be used. Those correlations had to be between a predictor and a criterion and not between two predictors.

2. All studies had to be applied to or carried out in an occupational setting. Studies using personality tests in a clinical, educational or any other setting in which job success was not a variable, were not included.

3. Only "recognized" personality measures were included in the study. No personality measures in the form of derived constructs or in any other form were to be included. In order to be considered "recognized", the personality measure had to appear in the "The Ninth Mental Measurements Yearbook" (Mitchell, 1985).

2.02 Data Coding

The data coding phase was one of the most important stages of the analysis. It determined the flexibility and manipulative characteristics of the data as well as the ability to obtain results from different groups. Much care was taken to code the data in such a way as to leave as many options open as possible for combining the different characteristics of each study.

Data coding consisted of 12 variables which were:

1. Study number
2. Year
3. Design
4. Sample type
5. Criterion type
6. Reliability type
7. Inclusion or exclusion
8. Test type
9. Cross validation
10. Validity coefficient
11. Sample size
12. Criterion reliability coefficient

Study Number

A three digit number was coded for each article according to the same method used in the Schmitt et al. (1984) study. The articles used in the present study have study numbers that correspond to the study numbers assigned to each article in the Schmitt et al. (1984) data. The purpose of this was to enable comparisons to be made between the Schmitt et al. (1984) data, the Guion and Gottier (1965) data and the data collected in the present study. It also allowed each article to be linked to the journal from which it was published.

Year

The year of publication for each validity coefficient was coded as a two digit number.

Design

The experimental design type was coded into predictive and concurrent designs, using single digit codes. Several studies incorporated a combination of both predictive and concurrent designs in their research.

Sample Type

The sample type categories used were the same as those used in the Schmitt et al. (1984) meta-analysis. The studies were coded into seven major categories. The categories were:

1 - *Professional* and semi-professional eg. Nurses, Policemen and Firemen.
2 - *Supervisor* eg. Managerial staff and Foremen.
3 - *Clerical* eg. Typists and receptionists.
4 - *Skilled labour* eg. Complex machine operators.
5 - *Unskilled* labour eg. Process workers and orderlies.
6 - *Sales* eg. All salespeople.
7 - *Other* eg. any other occupations that did not fall into the existing coding categories.

Reliability Type

Five forms of reliability measurement were recorded. They were: inter-rater reliability, internal consistency, test-retest, parallel forms, and split-half reliability.

Criterion Type

The types of criterion that were used in each article were the same as those used in the Schmitt et al. (1984) study. Nine categories were used.
They were:

1 - Performance ratings
2 - Turnover
3 - Achievement/grade
4 - Production
5 - Status change
6 - Wages
7 - Work sample
8 - Other
9 - Composite

Cross Validation

Each validity coefficient was coded for whether it had been cross validated.

Inclusion/Exclusion

One of the underlying problems in meta-analysis is an inconsistency in the quality of research published in journals. Because meta-analysis depends largely on sample sizes, a poorly designed study with a large sample size will have a larger impact on the final result than would a well designed study with a smaller sample size, thereby contaminating the final results and conclusions.

Another case for exclusion are studies using obscure personality tests which have been poorly constructed. The results of these studies should not be compared alongside studies using reputable comprehensively validated tests.

The inclusion/exclusion category enabled a distinction to be made between articles which were definitely allowed into the analysis, definitely not allowed into the present analysis but were included in other analyses, and articles where there was some doubt about their inclusion into the present study. The function of this category was to lower the chances of allowing Type I and Type II errors to be committed due to the inclusion of poorly designed studies in the meta-analysis.

The coding consisted of three categories:

0 -In
1 -Out
2 -Undecided

Test Type

The final coding category was an identification of which type of personality test was used for each reported validity coefficient. Six sub categories were derived using the categories used by Mitchell (1985), they were:

0 - Nonprojective character and personality tests
1 - Measures of interests
2 - Projective tests
3 - General vocational tests
4 - Leadership and managerial behavioural tests
5 - Miscellaneous tests

Nonprojective character and personality tests included the popular and traditional pencil and paper tests such as the Minnesota Multiphasic Inventory, California Psychological Inventory and the 16 Personality Factor Questionnaire.

Measures of interests included tests such as the Strong-Campbell Interest Inventory, the Kuder Occupational Interest Survey, and the Self Directed search.

The projective tests category involved traditional and controversial projective tests such as the Rorschach Inkblot Test, the Holtzman Inkblot technique, and the Thematic Apperception Test.

Vocational tests assessed the general career interests of subjects. Tests such as the Vocational Preference Inventory and the Vocational Planning Inventory were included.

The leadership and managerial behavioural tests were a more specific set of tests which assessed managerial and leadership qualities. Examples of such tests are the Leader Behaviour Description Questionnaire (LBDQ) and the Leadership Opinion Questionnaire (LOQ).

Miscellaneous tests consisted of tests that were custom-made for special purposes such as army recruiting. These tests are composite tests and may have been made up of separate scales taken from other personality tests such as the CPI or MMPI.

The data were recorded on data recording sheets designed for this study. An example of the data recording sheets and summary data coding sheet which was used as a quick reference when coding the data into the computer, is provided in Appendix A.

2.03 Construction of the Meta-Analysis Program

The data was analyzed using the Hunter and Schmidt meta-analytic procedures (Hunter, Schmidt & Jackson, 1982; Hunter & Schmidt, 1990). One of the main objectives in the design of the program was to enable it to be used on any I.B.M. PC, XT, AT or compatible computer system thereby increasing its utility.

The meta-analysis program was written in Microsoft GW-Basic and incorporates the Hunter and Schmidt meta-analytic formulas. It consists of two main units: the program body which houses the formulas and data processing commands, and the data body which contains all the data organized into rows and columns. Each row represents one record containing a single validity coefficient and its related coding data. The data body can be manipulated by deleting rows according to any coding criteria. The program will then run with the remaining data giving a full validity generalization result on any subset of the main data.

Each subset of data can then be individually processed to see how it fits in to the overall picture. The sample weighted mean of these subsets should be equal to the overall validity coefficient. This method is analogous to constructing a jigsaw puzzle using both a top-down and a bottom-up approach. The entire picture can be looked at from the start and taken apart piece by piece to see how it all fits in. Alternatively, the pieces can be broken down into almost any shape or size (coding flexibility), and linked to build the completed picture.

For each criterion-type and sample-type group, individual files were created to allow them to be analysed individually. This made it possible to find out the validity of each test-type when using different criteria and sample-types.

2.04 Data Analysis Procedures

Study data is entered into a data file located at the end of the meta-analysis program. The program then reads in the data which is organized into five fields, four of which contain statistical data and one which contains coding data. Data is processed individually, record by record then cumulated and the sample weighted mean of the correlations is calculated. This sample weighted mean is then corrected for predictor and criterion unreliability. By choice, the program can run either with or without correcting for predictor or criterion unreliability.

For comparative purppses, 24 separate sub-programs were constructed enabling the results to show how each of the six test types performed according to given sample types and criterion types.

2.05 Reported Validity Coefficients

After reading in the raw data from a data file, the program then calculated the mean unweighted correlation coefficient then corrected that correlation for sampling error using the reported sample sizes. Following that, it corrected the mean sample weighted correlation coefficient for unreliability in both the criterion and the predictor variables. At this stage the program could be split to run through correcting only for unreliability in the criterion and not in the predictor or vice versa.

Because the present study deals with the special case of personnel selection, corrections for predictor unreliability were not made, following the suggestion by Hunter and Schmidt (1990). However, to cover both sides of this issue and to give the clearest picture of how this would affect the final validity coefficient, two final results were reported. One was where predictor unreliability is corrected for and the other was where it is not.

2.06 Variance in Correlations

Variance in correlations was calculated at three levels. The first to be calculated was the variance in the uncorrected correlations. The second level was the variance in the correlations after correction for sampling error and the third level was the variance of the final validity coefficients after all corrections had been made.

2.07 Other statistics

The program also calculates the standard deviations of the correlations before and after corrections for unreliability as well as the amount of variance due to sampling error reported in percentages. A chi-square test of significance was performed and 95% confidence intervals are also calculated firstly after correction for sampling error and again after correction for unreliability.

2.08 Criterion and Predictor Reliability Coefficients

When a study is being corrected for attenuation and unreliability, two of the variables that affect the magnitude of the final validity coefficient are the reported criterion and predictor reliability coefficients. Because not all studies reported this information, there was the problem of how to deal with missing data.

There were three ways to overcome this difficulty. The first was simply not to correct for unreliability, another was to use substituted reliability estimates and the third was to use artifact distributions to make up for missing data. Hunter and Hirsh (1987), used the second method by substituting their own estimated criterion reliability coefficients for data that Schmitt et al. used in their 1984 meta-analysis. For example, the average correlations for performance criteria were corrected for range restriction using a reliability of 0.60 and the correlation for general mental ability was corrected for range restriction using a reliability of 0.67.

The data in the present study was corrected for unreliability using original and substituted correlation coefficients both separately and combined. Studies which do not report criterion reliability coefficients were assigned the same substituted coefficients as used by Hunter and Hirsh (1987). The results were also calculated using only the estimated reliability coefficients in order to be able to make a direct comparison of the conclusions of the Hunter and Hirsh (1987) study with the present study. The estimated reliability coefficients are shown in table 2.

Table 2.

Estimated criterion reliability coefficients.

Criteria	***Reliability Coefficient***
Performance Ratings	.60
Training	.81
All others	.60

The reliability for performance ratings and training criteria were taken from Pearlman, Schmidt and Hunter (1980). Hunter and Hirsh (1987) also used the same estimates in their meta-analysis. The reliability of .60 for all other criteria was used because Schmidt and Hunter (1977) have pointed out that this is probably a conservative estimate of average criterion reliability, therefore the use of this figure would still lead to a slightly underestimated validity coefficient but would improve its accuracy. Pearlman, Schmidt and Hunter (1980) stated that these coefficients are somewhat conservative leading to under-correction of the mean validity of a distribution.

2.09 The Dangers of Type I and Type II Errors

A Type I error in a study of this nature would mean that the researcher has claimed that a particular study has a high predictive validity when it doesn't. A Type II error would mean that the researcher claims that the study has a low predictive validity when it may in fact have a higher validity. Widespread use of an unknowingly invalid personnel selection test can be a costly mistake. Adopting conservative criterion reliability coefficients reduces the risk of incurring a Type I error which in the context of this type of study, is more serious than a Type II error.

When correcting for predictor unreliability, the coefficients derived from studies which report them were used to correct those studies individually. Studies which do not report predictor reliability coefficients will be coded with a zero in which case the program did not correct for unreliability.

One of the purposes of providing such a variety of results under different conditions, is to find out the effects of differences in the magnitude and inclusion of reliability coefficients on the final results. There may be a significant difference in correlation coefficients corrected under differing conditions.

2.10 Test of Statistical Significance

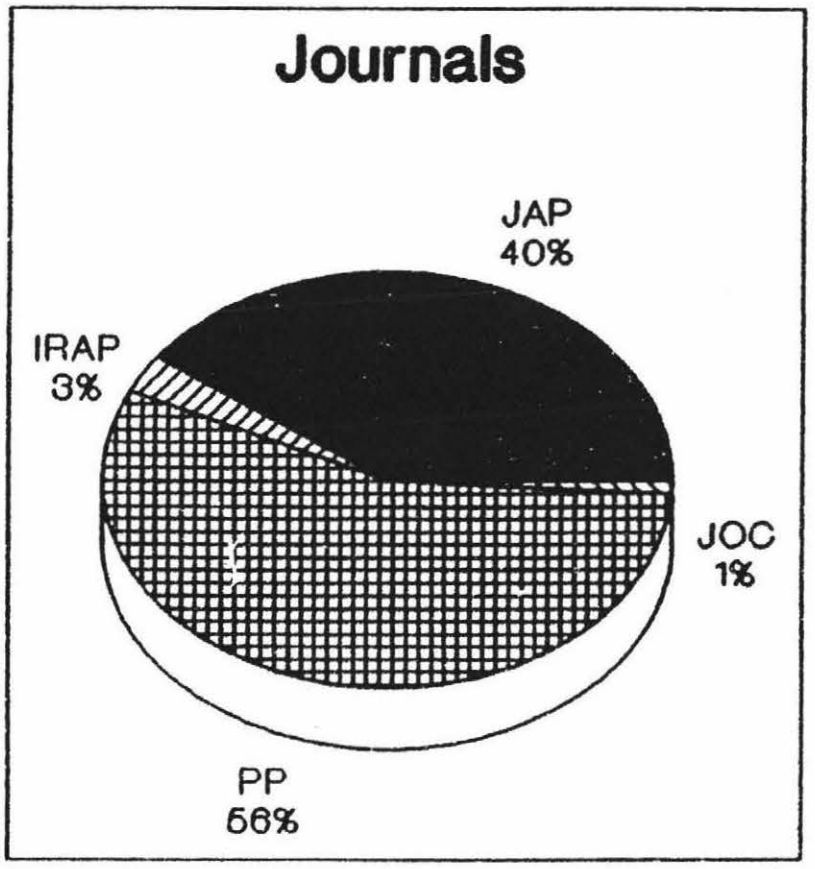
In order to compare the magnitude of the validity coefficient differences given the different calculation methods, different studies, criteria and test types, a chi-square test of statistical significance was used. Overall correlations of the Schmitt et al. (1984) study, the Ghiselli (1973) study, and the present study were tested for significant validity coefficient differences by converting the correlations using a Fishers Z transformation. In all cases the null hypothesis was that there were no significant differences between the correlations.

CHAPTER THREE

RESULTS

3.01 Journal Analysis

In all, 73 relevant articles were found, 41 from *Personnel Psychology*, 29 from the *Journal of Applied Psychology*, 1 from the *Journal of Occupational Psychology* and 2 from the *International Review of Applied Psychology* *. The proportions of articles collected from each of the journals used are shown in figure two.



JOC = Journal of Occupational Psychology, JAP = Journal of Applied Psychology
IRAP = International Review of Applied Psychology, PP = Personnel Psychology

Figure 2. Proportions of journal articles used in the meta-analysis.

---

* now known as *Applied Psychology. An International Review.*

A summary of relevant information pertaining to the characteristics of the studies used is shown in table three.

Table 3

Characteristics of the studies used in the meta-analysis.

			Test Type			
	NCP*	***Interest***	***Proj***	***Voc***	***Managerial***	***Misc***
Number of Independent Correlations	70	9	11	22	4	20
Average Sample Size	264	86	95	507	100	925
Range of Reported Sample Sizes	22-3964	33-267	34-193	11-8414	86-113	50-8497
Total Sample Size	18488	773	1047	11153	398	14506

* Non-projective Character and Personality Tests

Table three indicates that of the 73 articles collected, 24 studies were from the Schmitt et al. (1984) meta-analysis, 44 were from the Guion and Gottier (1965) review and 5 were from articles published between 1982 and October 1990. From the 73 articles, 246 independent coefficients were reported which were reduced to 136 summary coefficients by averaging coefficients from the same studies with similar characteristics.

The test-type with by far the greatest number of reported coefficients was the Non-projective character and personality tests. Managerial tests had the least number of reported coefficients and the smallest total sample size.

3.02 Analysis of Sample-types

In the sample-type analysis, the mean correlation was progressively reported from the sample weighting stage to the final stage where it was both sample weighted and corrected for attenuation. The variance before and after correction for unreliability and sampling error was also reported as well as a chi-square value, its level of significance, and a 90% confidence interval. In table four, data is presented summarizing the validity of studies over the various sample-types.

Table 4

Validity coefficients as a function of sample-type.

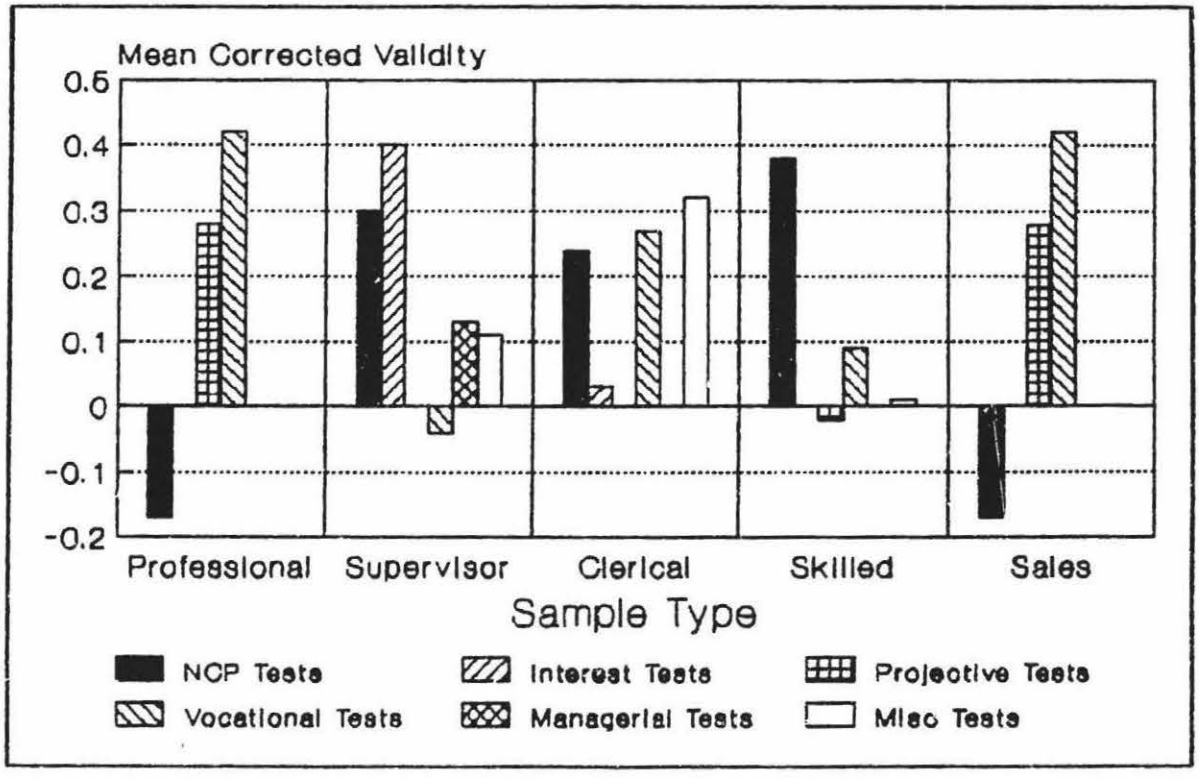
	Sample Type					
	Prof	***Super***	***Cler***	***Unskil***	***Skill***	***Sales***
Sample Size	21584	5152	612	1012	1486	1043
Mean r	.17	.13	.18	.11	.18	.10
Mean Sample Weighted r	.14	.14	.18	-.01	.15	.01
Sampling Error Var	.00273	.00600	.01555	.01301	.00776	.01163
Mean r Corrected for Attenuation	.17	.18	.23	-.01	.18	.01
Corrected Variance	.00974	.05483	.00342	.15429	.03330	.11526
% Var Due to Sampling Error	28%	11%	100%	8%	23%	10%
Number of Coeffs	61	32	10	13	12	12
Chi Square Value	212.45 *	213.33 *	11.51 NS	106.87 *	50.51 *	84.27 *
90% CI	.28-.35	-.28-.63	.12-.34	-.78-.76	-.18-.53	-.66-.67

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Closer examination of table four shows that the mean fully corrected r's appeared to be substantially higher than the mean sample weighted r's in the Professional, Supervisor, Clerical and Skilled labour categories and that of the six sample types used in the meta-analysis, Clerical workers had the highest validity followed by Supervisors and Skilled Labourers. Against the trend, Unskilled labourers showed a low negative correlation which was accompanied by a high corrected variance. Sales personnel followed along the same lines with very low validity and high corrected variance.

Total sample sizes indicate that there were notably more subjects used for studies using Professional workers as samples than for any of the other categories. Although Clerical workers had the highest mean validity, they also had the smallest total sample size. In relation to this small sample size, chi-square tests of homogeneity were significant for all but the Clerical sample and the 90% confidence interval was much wider for the Clerical sample than for any of the other samples. Therefore the implication that the Clerical sample had the highest mean validity should be treated very cautiously as a chance variation cannot be ruled out. Relating to the issue of chi-square tests of homogeneity, Hunter and Hunter (1990) stated that there are dangers in the use of statistical tests of homogeneity in the context of meta-analysis and it is usually undesirable to rely on their implications.

Following the general sample type analysis, all sample types were subdivided into test types and separate validities were obtained from them to show how each test type performed in relation to sample type. Summary results of this analysis are shown in figure three.



* Test-types with less than two reported correlation coefficients were not included.

Figure 3. Validity of tests within each sample-type.

The bar chart shows that none of the predictors in any of the samples achieved mean correlations greater than 0.5. No figures were available for the Unskilled sample because there was only one reported coefficient for each test type. However, the results did show that for Professional workers, Vocational tests were the best predictors of job success. For Supervisory staff, Interest tests were the most valid. Miscellaneous or composite tests had the highest validity in the Clerical staff category followed closely by Non-projective Character and Personality tests (NCP), and Vocational tests. NCP tests had clearly the highest validity in the Skilled labour samples and for Sales personnel, Vocational tests were the best predictors. Full details of each analysis can be seen in Appendix C.

3.03 Criterion-Type Analysis

Criterion-type data was analysed in the same way as the Sample type data. The results of this analysis can be seen in table five. Of the eight types of criteria used, personality tests correlated most highly with "Other" criteria such as creativity ratings, job tenure and employee suggestions. Of the specified criteria, personality tests correlated relatively highly with Performance ratings which was one of the most common criterion-types. Lower correlations were found with Achievement/Grades and Status Change/Promotion and the Achievement correlation was found to be non significant in the chi-square test.

A strong negative correlation was found in the Production criteria and a very low correlation was found in the Performance rating criteria which had the largest sample size. Wages and Composite criteria also had low correlations and were non significant. All of the three criteria that yielded non-significant results had sample sizes of less than 500 and were based on fewer than 10 studies. Low variance and a high degree of sampling error was found in the Achievement and Composite criteria.

Table 5

Validity of different criterion-types.

				CRI	*TYPE*			
	Perf	***Turn***	***Ach***	***Prod***	***Stat***	***Wage***	***Oth***	***Comp***
Spl Size	26542	735	429	506	20911	314	565	363
Mean r	.15	.18	.21	-.09	.19	.08	.23	.08
Mean Sample Weighted r	.22	09	.15	-.17	.15	.09	.26	.10
Sampling Error Var	.003	.011	.011	.006	.001	.009	.017	.022
Mean r Corrected for Atten	.013	.11	.19	-.22	.19	.11	.32	.12
Corrected Variance	.268	.199	.013	.143	.006	.124	.169	.012
% of Var Due To Sampling Error	20%	6%	84%	4%	17%	7%	10%	100%
Number of Coeffs	79	8	5	3	19	3	11	8
Chi Square Value	374.1 *	97.4 *	8.69 NS	49.2 *	101.3 *	0.69 NS	84.3 *	11.6 NS
90% CI	.02-.51	-.76 -.98	-.03 -.41	-.96 -.51	.03-.34	-	-.49 1.1	-.09 .33

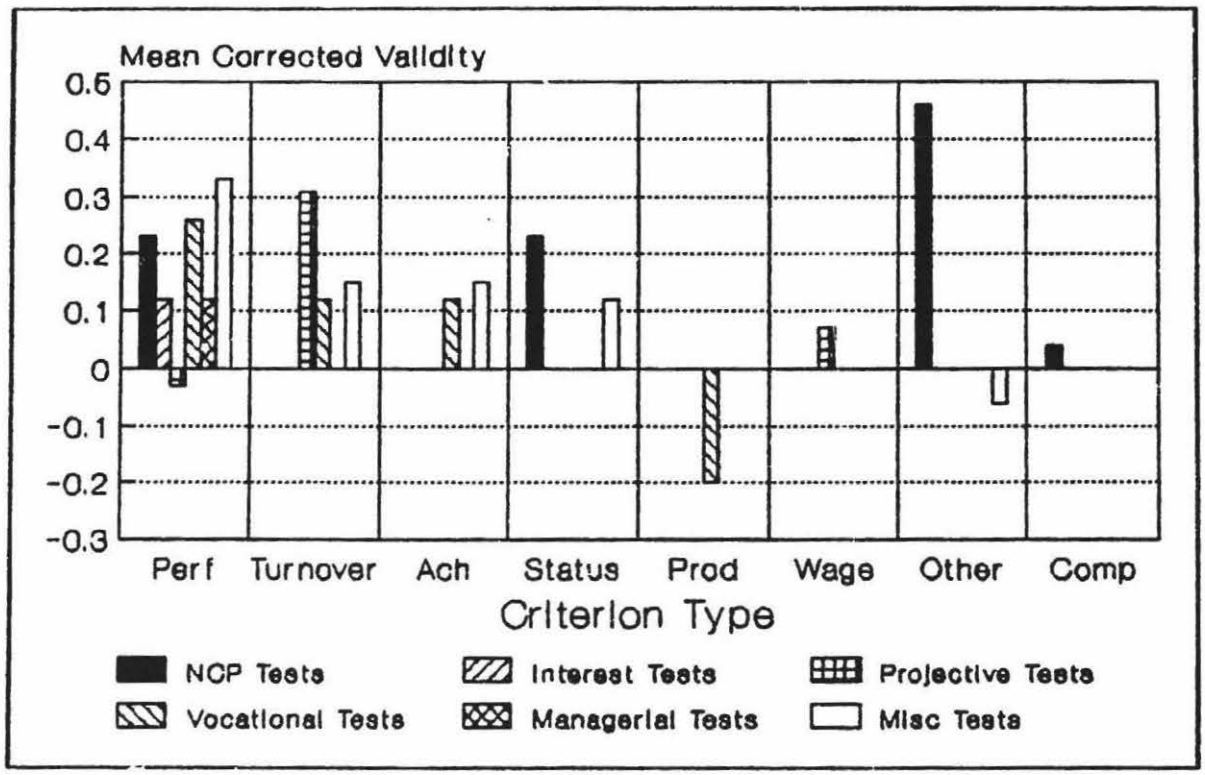
* = $p<.005$ ** = $p<.05$ NS = Non Significant

Subgroups of test-types were formed and analysed in the same way as the sample-type to find out which test-type was the best predictor within each criterion-type. The summary results of this analysis shows a patchy use of criteria in personality test validation studies (see figure 4). Once again as in the sample-type sub-analysis, none of the correlations exceeded 0.5.

Performance ratings was the only criterion to have studies involving all six test-types and in that group as well as the Achievement group, the Miscellaneous tests appeared to have the highest correlation. Non-projective character and personality tests had the highest correlations in the Status change, Other and Composite criteria classes. Projective tests faired best for the studies using Turnover and Wages as criteria. A strong negative correlation with Vocational tests was found in the Production category, however there were only three coefficients reported so one or two strongly negative correlations is enough to yield a negative overall result. Full details of each analysis can be seen in Appendix D.

It should be noted that sample sizes for the subgroup analyses were considerably smaller than for the overall group analyses so a large proportion of the results were not significant in the chi-square tests. This however, does not mean that the results can be overlooked. Instead they can be taken in a descriptive context, adding depth to the overall results, hence the use of bar charts to facilitate this process.

Sample sizes will increase as the entire data set becomes involved in the more global comparative analyses. The first of these is an analysis which was carried out to determine which test-type returned the highest overall correlation. All studies were divided into six test-type categories and were analysed separately.



* Test-types with less than two reported correlation coefficients were not included.

Figure 4. Validity of each test-type within each criterion-type.

3.04 Test-type Analysis

The results shown in table six indicate that out of all six test-types, Vocational tests had the highest overall correlation followed very closely by Miscellaneous tests which consisted of custom-made composite tests. NCP tests were also among the top three tests. The bottom three consisted of Projective, Managerial and Interest tests in descending order and all had considerably lower correlations than the others. The top three had a range of .21 - .25 whereas the bottom three ranged from .12 - .15.

All correlations were highly significant except for the Interest tests and Managerial tests categories which had relatively small sample sizes and very few reported coefficients. Sample sizes were the largest for Non-projective character and Personality tests, Miscellaneous and Vocational tests.

Table 6

Validity of different test-types.

	Test Type					
	NCP	**Int**	**Proj**	**Voc**	**Man**	**Misc**
Sample Size	18488	773	1047	11153	398	18506
Mean Sample Weighted r	.17	.11	.12	.19	.11	.19
Sampling Error Var	.0036	.0152	.0103	.0018	.0099	.0010
Mean r Corrected for Attenuation	.21	.12	.15	.25	.13	.24
Corrected Variance	.0218	-.0008	.0434	.0229	.0093	.0179
% Var Due to Sampling Error	16%	100%	24%	8%	100%	6%
No. of coeffs	70	9	11	22	4	20
Chi Square Value	358.76 *	8.77 NS	39.17 *	190.18 *	6.86 NS	250.3 *
90% CI	-.08 -.50	-	-.26 -.56	-.05- .54	-.06 - .32	-.03- .49

* = $p<.005$ ** = $p<.05$ NS = Non Significant

3.05 Re-Analysis of the Schmitt et al. data

One of the important issues surrounding the present study is the saliency of the technique used by Schmitt et al. (1984) where coefficients were corrected only for sampling error but not attenuation. The Schmitt et al. (1984) data was re-analysed in the present study and corrections for unreliability were made. A comparison of the Schmitt et al. (1984) results and the present study results is shown in table seven.

Results of the comparison show that the overall validity coefficient of .149 in the Schmitt et al. (1984) study increased to .181 when corrected for attenuation. Correspondingly, the variance and confidence intervals have widened in the corrected study.

Table 7

Comparison of Schmitt et al. (1984) data and method with present study method.

Comparison	***SNGK Study***	***Present Study***
Total Sample Size	23430	23430
No. of Coeffs	62	62
Final Coefficient	.149	.181
Variance	.00782	.01135
90% CI	-.0247 - .3219	-.0279 - .3893

The differences between the two mean correlations were tested using Fisher's Z transformation and were found to be statistically significant at both the 0.05 and the 0.01 levels (Z = 3.56) using a two-tailed test.

3.06 Overall Results

The overall analysis uses data from 1953 through to 1990 covering some 37 years of research. In table eight the overall validity coefficient is presented in two forms. The first, (A), is where corrections were made only for criterion unreliability. The second, (B), is where corrections for both criterion and predictor unreliability were made. The conclusions of the present study will be based on the results in column A, which were corrected for criterion unreliability but not predictor unreliabilty as recommended by Hunter and Hunter (1990). Given this, the overall results in table eight show that personality tests had a mean validity of .22 and the variance attributed to sampling error was found to be 12 percent. This result was highly significant at the .005 level. When corrections were made for both criterion and predictor unreliability, the validity rose slightly to .25 and the variance due to sampling error dropped to 10 percent (column B).

The overall results were calculated on the whole data set, most of which was used by Guion and Gottier (1965), Ghiselli (1973) and Schmitt et al. (1984). Some studies had very poor experimental design either due to the tests used or the sample population characteristics.

Table 8

Final validity coefficients for all personality studies from 1953 to 1990.

	A	*B*
Sample Size	50365	50365
Mean Sample Weighted r	.18	.18
Sampling Error Var	.00253	.00253
Mean r Corrected for Attenuation	.22	.25
Corrected Variance	.02062	.02550
% Var Due to Sampling Error	12%	10%
No. of coeffs	136	136
Chi Square Value	869.95 *	869.95 *
90% CI	-.06 - .50	-.06 - .56

* = $p<.005$ ** = $p<.05$ NS = Non Significant

3.07 Results of the INOUT Analysis

To show what the effects of more stringent data selection techniques may have, data was coded as to whether or not it should be included or whether it was undecided as described earlier. Table nine shows the resulting correlations using the three conditions. The first analysis used data that was deemed acceptable. The second analysis used data that was thought to be unacceptable and the third analysis combined data from both the unacceptable and undecided categories.

Table 9

Validity coefficients as a function of INOUT criteria.

	IN	***OUT***	***IN + UNDEC***
Sample Size	37205	12859	37506
Mean Sample Weighted r	.19	.15	.19
Sampiing Error Var	.00242	.00269	.00248
Mean r Corrected for Attenuation	.24	.18	.24
Corrected Variance	.02017	.02004	.02016
% Var Due to Sampling Error	12%	13%	12%
No. of coeffs	97	36	100
Chi Square Value	630.95 *	216.58 *	635.10 *
90% CI	-.04-.52	-.10-.45	-.04-.52

* = $p<.005$ ** = $p<.05$ NS = Non Significant

The results in table nine show that by excluding studies which were deemed unacceptable, the resulting coefficient did not differ much from the overall validity. In all, 36 coefficients were deemed unacceptable and when combined, had a mean correlation of .18. When the UNDECIDED group were added to the IN group, there was no noticeable difference in magnitude from the IN group result.

3.08 Results Summary

In summary, the results of the study showed that the sample-types that correlated most highly with measures of job success were the Supervisory and Skilled workers. The most successful criterion-type were the "Other" criterion measures followed by Achievement and Status change.

The Schmitt et al. (1984) data was corrected for criterion unreliability and the resulting validity coefficient was found to be significantly higher than the figure quoted by Schmitt et al. (1984) of .15.

The overall mean validity of personality tests was found to be .22 and it was found that Vocational tests with a mean validity of .25 were the best predictors of the six classes of personality measures.

CHAPTER FOUR

DISCUSSION

4.01 Journal Analysis

The analysis of journal proportions used in the present study showed that of the seven official journal publications looked at, 40% of the articles used in the present study were from the Journal of Applied Psychology and 56% were from Personnel Psychology. This result goes some way to justifying the Schmitt et al. (1984) use of only two main journals in their data collection procedure. These two main journals contained 96% of all validity data available in the area of personality tests for personnel selection. McDaniel et al. (1986) claimed that Schmitt et al. only used studies that were published in the fields two most prestigious journals. This does not appear to be a serious problem given the findings of the present study. However, this does not justify the use of only two journals by Schmitt et al. (1984). The dangers of doing so are still present and Schmitt et al. (1984) were fortunate that there were so few published studies outside the two main journals.

McDaniel et al. (1986) criticised the use of only published studies in the Schmitt et al. meta-analysis. Their claims are related to the "file drawer" problem advocated by Rosenthal (1979), whereby mainly significant studies are published while non-significant studies remain in the file drawers. As can be expected, a solution to this problem cannot be easily found, however, even if unpublished studies may be difficult to obtain, the Schmitt et al. (1984) study could have moved one step closer by covering all the journals instead of just two.

The present study has moved a step closer by using data references obtained through the DIALOG on-line computer search and by conducting a thorough manual search through seven journals relating to general and industrial psychology.

4.02 Sample-Type Analysis

The results of the sample-type analysis showed that for all types of personality tests, the highest significant validity coefficients were found in the Supervisory and Skilled worker groups. Although Clerical workers did in fact have the highest validity (.23), the result was non-significant.

These findings are comparable with the Schmitt et al. (1984) meta-analysis where it was found that Managerial (Supervisory) and Clerical groups had the highest validity coefficients across all sample-types. In direct comparison however, the magnitude of the coefficients in the present study were considerably lower (.18) compared to the Schmitt et al. (1984) study (.34-.39). The specific types of personality tests that were the best predictors of Supervisory and Skilled workers were Interest and Non-projective Character and Personality tests respectively. Along the same lines, the Ghiselli (1973) meta-analysis, which were simply mean unweighted correlations, also found Supervisors to have the highest validity (.34) followed by Sales personnel (.32).

4.03 Criterion-Type Analysis

The criterion categories used were the same as those used in the Schmitt et al. study. Among the eight criterion-types, the "Other" criteria achieved the highest significant validity (.32). This result highlights the benefits of proper criterion development and validation. Criteria that were used in the Other category were specifically designed for special tasks such as Navy underwater diving and architectural creativity.

Performance ratings as criteria were found to have the next highest validity (.27), which is in accord with the results of the re-analysis of the Schmitt et al. (1984) data performed by Hunter and Hirsh (1987). The present study found Composite and Vocational tests to be the best test-types using performance ratings as criteria. The Schmitt et al. (1984) study found that across all predictors, Wages and Status change were the most valid (.38 & .36 respectively).

The present study found Achievement and Status change to have much lower validities for those two criteria (.19 each). Hunter and Hirsh (1987) claimed an even lower validity of .13 using the Schmitt et al. (1984) data.

4.04 Test-Type Validity

The results of the test-type analysis revealed that Vocational tests, had the highest validity (.25) followed closely by Miscellaneous and NCP tests. The validity of all six tests in general, were low and ranged from .12 to .25. This reflects the general opinion that personality tests are not valid occupational predictors.

It is interesting to note that NCP tests include some of the most widely used pencil and paper psychological tests such as the MMPI, CPI and 16 PF. That category of tests had only the third highest validity (.21) but is the most researched and utilised test category. Advocates of these tests carry on using them, not because of their psychometric and predictive capabilities, but because of their past popularity and incorrect beliefs that they are useful.

Perhaps what should be focussed on during test development and before each application, are the relationships between the actual content of the test and what is being assessed or, in other words, content validity. For example the results seem to indicate that the more specific the test is to the job, the better the validity.

Content specificity could also be a critical factor as Vocational tests tend to be more job specific than Projective and Interest tests. Even more so are the Miscellaneous tests which are custom made for the task or job being assessed. Both of these test-types yielded the highest correlations and were of the job-specific nature, apart from the Managerial tests which, due to too small a sample size, could not be compared.

Vocational, Miscellaneous and NCP tests were also by far the most popular in terms of subject numbers and reported coefficients. The only other result that was not overly contaminated by sampling error was that of the Projective tests which as expected, returned a very low validity coefficient of .15.

The results of all the separate analyses on the coding groups support the first of the three major assumptions of the present study; that results and conclusions can vary greatly according to the decision rules used to categorise the data.

This implies that one overall validity coefficient can be misleading and damaging to a predictor which can perform better in some situations than in others. Breaking down personality testing into specific sample types, criterion-types and test- types has made it possible for more accurate summary statements to be made.

Personality tests should not be used on particular sample-types or criterion-types which appear to have very low or even negative correlations. The test-type category can be broken down even more, into specific tests so as to be able to identify tests which may be very good predictors against tests which may be very poor predictors. Combined analysis means tests which are poor predictors can overshadow tests which may be very good predictors.

4.05 Re-calculation of Schmitt et al. (1984) Meta-analysis

The results of the re-calculation of the Schmitt et al. (1984) data showed that if corrections were made for unreliability, there would be a significant increase in the final validity coefficient. This supports the second of the three main assumptions of the present study: that results and conclusions can vary greatly by the method used to calculate mean validity coefficients.

This result highlights the importance of correcting for both sampling error and unreliability in any form of meta-analysis. Any lesser correction will mean a misleading underestimate of the true validity. To add to this, the fully corrected result of the present study is still considered to be a slight underestimate of the true validity, because conservative estimates of reliability were used for studies which did not report criterion reliability coefficients (see Hunter & Hirsh, 1987).

4.06 Outcome of the Inout Analysis

One of the main issues surrounding the problem of the data selection is that of inclusion or exclusion of data that originated from poor experimental design. The results of the INOUT analysis showed that in this case, data that should have been excluded did not have a great affect on the final validity coefficient when it actually was excluded. However, the issue of careful data selection still stands because over 30% of the coefficients used in the present study were seen to be sub-standard for inclusion into a meta-analysis. The only reason why they were included was because other researchers such as Schmitt et al. (1984) had included that data in their studies.

The results of the INOUT analysis do not appear to support the last of the three assumptions of the present study; that results and conclusions can vary greatly according to criteria used to select study data.

4.07 Overall Results and Conclusions

From the overall results, the present study claims a figure of .22 for the validity of Personality tests using a meta-analytic procedure which corrected for both sampling error and criterion unreliability. This figure is claimed to be the most recent, accurate and global generalization of personality test validity yet, because it uses data covering 38 years of personality test validation research.

The final result is in accordance with the Hunter and Hirsh (1987) re-analysis of part of the Schmitt et al. (1984) data, in which the mean fully corrected validity coefficient was found to be .24. It also lends support to the Ghiselli (1973) conclusions that personality tests have a validity of .22. The implications of the present study are that the Schmitt et al. (1984) meta-analytic procedure was incomplete and their results were misleading and not a true indication of the validity of personality tests.

Despite claims of significant differences from the Schmitt et al. (1984) study, the overall conclusions about the validity of personality tests remains the same. From the Guion and Gottier (1965) review to the present study 25 years later, personality tests used as selection tools are poor predictors of job success and if used, should be treated with caution.

4.08 Editorial Policies and the Quality of Data Reporting

After having read through numerous journal articles, it was apparent that many studies had concentrated largely on their introductions but poorly described their methodology. Without an adequate description of how and what the authors used to obtain their findings, it makes it very difficult to decipher the statistics reported in results sections. It would appear that the reader was supposed to assume the results were unquestionable and go on to the discussion and read about what great findings the author has discovered.

When collecting data for a meta-analysis, data must be of a certain standard to allow it to be included into the analysis. This standard can only be determined by finding out exactly how the authors coded their data, what statistics they used and they used them. Often was the case when there would be holes in correlation matrices because the authors decided to mysteriously leave out data or decided to combine groups of data without any explanation.

This inadequacy of data reporting may in some cases, be out of the author's control. The culprit in some cases may be the editorial policies of the journals themselves. The *Journal of Applied Psychology* an example whereby stringent editorial policies limit the length of each article to approximately four published pages. In normal manuscript this is the equivalent of 16 double-spaced pages. Surely authors cannot adequately describe 70-100 pages of research in such a condensed form. The current editor of the Journal of Applied Psychology is, coincidentally, Neal Schmitt. He emphasised the importance of theoretical accuracy in accepting an article for publication. However, questions have been raised about the editorial policy of the *Journal of Applied Psychology* and previous editors have expressed the importance of technical adequacy (Schmitt, 1989).

Regarding the importance of technical versus theoretical adequacy, it can be said that even though both qualities are of great importance, "seeing is believing", and a high degree of technical accuracy should enable any article of research to be replicated. This is by far the best method of scientific confirmation.

4.09 Issues Surrounding Use of Significance Tests

The use of chi-square significance tests in the present study have been used for coefficients corrected for sampling error but not unreliability. Hunter and Schmidt (1990), recommended the use of chi-square tests for fully corrected correlations because they do not suffer from the Type I errors caused by variance due to artifacts that have not been corrected for. Even with fully corrected correlations, the use of significance testing has been cautioned by Hunter and Schmidt (1990). They state that it is undesirable to rely on tests of high statistical power such as the chi-square test, because they allow only for sampling error but not any other artifactual sources of variance. Some of these other sources of variance include transcriptional, computational and the reliability of measurements. These other effects create variance beyond sampling error and can falsely cause a chi-square test to be significant.

Hedges and Olkin (1985), caution that researchers should not become caught up in significance tests and fall into the usual trap of relying on the fact that if it is significant it must be important and if it isn't then forget about it. Instead, they suggest that researchers should evaluate the actual size of the variance.

4.10 Future Directions

The results of the present study have supported the Hunter and Hirsh (1987) claims and disputed the Schmitt et al. (1984) claims. The argument will not end here nor will it end in the near future. The area of meta-analysis is now in the performance stage where the basic engine being the concept of global validity generalization is already developed but needs to be refined. It is this refining process that disputes have and are continuing to arise. The statistical procedures involved are becoming more refined and complex. With the aid of powerful microcomputers, statistical complexity can be overcome and only a fraction of a second taken to analyse data. The main advantages of computerization are data storage, manipulation, and the ability to perform infinite analyses on any subset of the main data set in the space of a few minutes.

Once a particular method is universally agreed upon, a commercial statistical package can be designed to be user friendly enough for non-experts to simply enter and process data in any topic area and maintain a dynamic picture of the field. The key to the utility of meta-analytic procedures is the fact that more researchers must report validity coefficients in their studies.

REFERENCES

Bernardin, H.J. & Bownas, D.A. (1985). *Personality Assessment in Organizations (eds).* New York: Praeger.

Binning, J. F. & Barret. (1989). Validity of Personnel Decisions. *Journal of Applied Psychology 74*, 478-494.

Burke, M. J. (1984). Validity Generalization: A Review and Critique of the Correlation Model. *Personnel Psychology, 37*, 93-116.

Conrad, E. & Maul, T. (1981). *Introduction to Experimental Psychology.* New York: Wiley.

Cook, D. T. & Leviton, L. C., (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality, 48*, 449-472.

Fisher, C. D., & Gitelson, R. (1983). A meta-analysis of the correlates of role conflict and ambiguity. *Journal of Applied Psychology, 68*, 320-333.

Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Guion, R. M. & Gottier, R. F. (1965). Validity of Personality Measures in Personnel Selection. *Personnel Psychology, 13*, 135-164.

Hedges, L.V. & Olkin, I. (1985). *Statistical Methods For Meta-Analysis.* Orlando, FL: Academic Press.

Howell, W. C. & Dipboye, R. L. (1982). *Essentials of industrial and Organizational Psychology.* Illinois: Dorsey.

Hunter, J.E. & Hirsh, H.R.H. (1987). Applications of Meta-Analysis. *International Review of Industrial & Organizational Psychology.* New York: Wiley.

Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* London: Sage.

Hunter, J.E. & Schmidt, F.L. & Gregg. (1982). *Meta-analysis: Cumulating Research Findings Across studies.* California: Sage.

Kemery, E. R. et al. (1989). Meta-Analysis & Moderator Variables: A Cautionary Note. *Journal of Applied Psychology, 74,* 168-170.

Kleinbaum, D.G. & Kupper, L.L. (1978). *Applied Regression Analysis and Other Multivariate Methods.* Massachusetts: Duxbury.

McCormick, E. J. & Tiffin, J. (1974). *Industrial Psychology (6th ed).* New Jersey: Prentice-Hall.

McDaniel, M. A. Hirsh, H. R. Schmidt, F. L. Raju, N. S. & Hunter, J. E. (1986). Interpreting the results of Meta-Analytic research: A comment on Schmitt, Gooding, Noe and Kirsch (1984). *Personnel Psychology, 39,* 141-148.

Mitchell, J.V. (1985). *The Ninth Mental Measurements Yearbook.* Nebraska: Buros Institute of Mental Measurements.

Monahan, C. J. & Muchinsky, P. M. (1983). Three decades of personnel selection: A state-of-the-art analysis and evaluation. *Journal of Occupational Psychology, 56,* 215-225.

Osburn, Callender, Greaner & Ashworth. (1983). Statistical Power of the Situational Specificity Hypothesis in Validity Generalization Studies; a cautionary note. *Journal of Applied Psychology, 68,* 115-122.

Robertson, I. T. & Downs, S. (1989). Work Sample Tests of Trainability: A Meta Analysis. *Journal of Applied Psychology, 74*, 402-410.

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research.* California: Sage.

Rosenthal, R. (1979). The "File Drawer Problem" and Tolerance for Null Results. *Psychological Bulletin, 86*, 638-641.

Schmitt, N. (1989). Editorial. *Journal of Applied Psychology, 74*, 843-845.

Schmitt, N. Gooding, R. Z. Noe, R. A. & Kirsch, M. (1984). Meta-analyses of Validity Studies Published Between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Schmitt, N. & Robertson, I. (1990). Personnel Selection. *Annual Review of Psychology, 41*, 289-319.

Spector, P. E. & Levine, E. L. (1987). Meta-Analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72*, 3-9.

Wanous, J. P. et al. (1989). The Role of Judgement Calls in Meta-Analysis. *Journal of Applied Psychology, 74*, 259-264.

Zagar, R. Arbit, J. Falconer, J. & Friedland, J. (1983). Vocational Interests and Personality. *Journal of Occupational Psychology, 56*, 203-214.

APPENDIX A

Sources of Validity Data for the Meta-Analysis

Bass, B.M, Karstendiek, Barbara, McCullough, G., & Pruitt, R.C. (1957). Validity Information Exchange, No. 10-25. *Personnel Psychology, 10*, 343-344.

Biersner, R. J. & LaRocco, J. M. (1983). Personality Characteristics of US Navy divers. *Journal of Occupational Psychology, 56*, 329-334.

Borman, W. C., Rosse, R. L. & Abrahams, N. M. (1980). An empirical construct validity approach to studying predictor-job performance links. *Journal of Applied Psychology, 65*, 662-671.

Brayfield, A.H. & Marsh, Mary M. (1957). "Aptitudes, Interest, and Personality Characterstics of Farmers." *Journal of Applied Psychology, 41*, 98-103.

Bromer, J.A., Johnson, J.M., & Severansy, P. (1962). Validity Information Exchange, No. 15-02. *Personnel Psychology, 15*, 107-109.

Bruce, M.M. (1954). Validity Information Exchange, No. 7-079. *Personnel Psychology, 7*, 425-426.

Carron, T. (1969). Validity tests for chemical plant personnel. *Personnel Psychology, 27*, 307-322.

Campbell, J.T., Otis, J.L., Liske, R.E. & Prien, E.P. (1962). "Assessments of Higher Level Personnel": II. Validity of the Over-all Assessment Process." *Personnel Psychology, 15*, 63-74.

Campbell, J.T, Prien, E.P., & Brailey, L.G. (1960). "Predicting Performance Evaluations." *Personnel Psychology, 13*, 435-440.

Carter, G.C. (1952). "Measurement of Supervisory Behaviour." *Journal of Applied Psychology, 36*, 393-395.

Comrey, A.L. & High, W.S. (1955). "Validity of Some Ability and Interest Scores." *Journal of Applied Psychology, 39*, 247-248.

Cummin, P. C. TAT Correlates of executive performance. (1967). *Journal of Applied Psychology, 51*, 78-81.

Day, D. V. & Silverman, S. B. Personality and job performance: Evidence of incremental validity. (1989). *Personnel Psychology, 42*, 25-35.

Dubois, P.H. & Watson, R.I. (1954). Validity Information Exchange, No.8-16. *Personnel Psychology, 7*, 414-416.

Dugan, R.D. (1961). Validity Information Exchange, No. 14-01, *Personnel Psychology, 14*, 121-125.

Dunnette, M. D. & Aylward, M. S. (1956). Validity Information Exchange, No.9-21. *Personnel Psychology, 4*, 245-247.

Dunnette, M. D. & Kirchner, W. K. (1960). "Psychological test differences between industrial salesmen and retail salesmen." *Journal of Applied Psychology, 44*, 121-125.

Fitzpatrick, E.D. & McCarty, J.J. (1955). Validity Information Exchange, No. 8-35. *Personnel Psychology, 8*, 501-504.

Fletcher, C. (1987). Candidate personality as an influence on selection interview assessments. *Applied Psychology: An International Review, 36*, 157-162.

Ghiselli, E. E. Prediction and success of stockbrokers. (1969). *Personnel Psychology, 22*, 125-130.

Gluskinos, U. Brennan, T. F. Selection and evaluation procedure for operating room personnel. (1971). *Journal of Applied Psychology, 55*, 165-169.

Goodstein, L. D. & Schrader, W. J. (1963). "An Empirically-Devised Managerial Key for the California Psychological Inventory." *Journal of Applied Psychology, 47*, 42-45.

Graham, W. K. & Calendo, J. J. (1969). Personality correlates of supervisory ratings. *Personnel Psychology, 22*, 483-487.

Grant, D.L. (1954a). Validity Information Exchange, No.7-085. *Personnel Psychology, 7*, 557-558.

Grant, D.L. (1954b). Validity Information Exchange, No.7-086. *Personnel Psychology, 7*, 559-560.

Grant, D. L., Katkovsky, W. & Bray, D. W. (1967). Contributions of projective techniques to assessment of management potential. *Journal of Applied Psychology, 51*, 226-232.

Guilford, Joan S. (1952). "Temperament traits of Executives and Supervisors Measured by the Guilford Personality Inventories." *Journal of Applied Psychology, 36*, 228-233.

Gunderson, E. K. E., Rahe, R. H. & Arthur, R. J. (1972). Prediction of performance in stressful underwater demolition training. *Journal of Applied Psychology, 56*, 430-432.

Hall, W. B. & Mackinnon, D. W. (1969). Personality inventory correlates of creativity among architects. *Journal of Applied Psychology, 53*, 322-326.

Hilton, A.C., Bolin, S.F., Parker, J.W., Jr., Taylor, E.K., & Walker, W.B. (1955). "The Validity of Personnel Assessments by Professional Psychologists." *Journal of Applied Psychology, 39*, 287-293.

Hoiberg, A. & Pugh, W. (1978). Predicting navy effectiveness: Expectations, motivation, personality, aptitude, and background variables. *Personnel Psychology, 31*, 841-852.

Hughes, J.L. & Dodd, W.E. (1961). "Validity vs Stereotype: Predicting Sales Performance by Ipsative Scoring of a Personality Test." *Personnel Psychology, 14*, 343-355.

James, P. J., Campbell, I. M. & Lovegrove, S. A. (1984). Personality differentiation in a Police-Selection interview. *Journal of Applied Psychology, 69*, 129-134.

Johnson, J. A. & Hogan, R. (1981). Vocational interests, personality and effective police performance. *Personnel Psychology, 34*, 49-53.

Kaufman, H. G. (1972). Relations of ability and interests to currency of professional knowledge among engineers. *Journal of Applied Psychology, 56*, 495-499.

Keinan, G., Friedland, N., Vitzhaky, J. & Monan, A. (1981). Biographical, physiological, and personality variables as predictors of performance under sickness-inducing motion. *Journal of Applied Psychology, 66*, 233-241.

Laurent, H. (1970). Cross cultural cross validation of empirically validated tests. *Journal of Applied Psychology, 54*, 417-423.

McCarty, J.J. (1957). Validity Information Exchange, No. 10-15. *Personnel Psychology, 10*, 204-205.

McClelland, J. N. & Rhodes, F. (1969). Prediction of job success for hospital aides and orderlies from MMPI scores and personal history data. *Journal of Applied Psychology, 53*, 49-54.

McDermid, C. D. (1965). Some correlates of creativity in engineering personnel. *Journal of Applied Psychology, 49*, 14-19.

McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, M.A. & Ashworth, S. (1990). Project A Validity Results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 347-353.

MacKinney, A. C. & Wollins, L. (1960). Validity Information Exchange, No. 13-01. *Personnel Psychology, 8*, 443-447.

McNamara, W.J. & Hughes, J.L. (1961). "A Review of the Research on the Selection of Computer Programmers." *Personnel Psychology, 14*, 39-51.

Maher, H. (1963). Validity Information Exchange, No.16-01. *Personnel Psychology, 16*, 71-73.

Maher, H. (1963). Validity Information Exchange, No.16-02. *Personnel Psychology, 16*, 74-77.

Minor, J.B. (1960). "The Kuder Preference Record in Managerial Appraisal." *Personnel Psychology, 13*, 187-196.

Munson, J. M. & Posner, B. Z. (1980). Concurrent validation of two value inventories in predicting job classification and success for organizational personnel. *Journal of Applied Psychology, 65*, 536-542.

Nash, A. N. (1966). Development of an SVIB key for selecting managers. *Journal of Applied Psychology, 50*, 250-254.

Parry, M. E. (1961). Ability of psychologists to estimate validity of personnel tests. *Personnel Psychology, 21*, 139-147.

Perrine, M.W. (1955). "The Selection of Drafting Trainees." *Journal of Applied Psychology, 39*, 57-61.

Tenopyr, M. L. (1969). The comparative validity of selected leadership scales relative to success in production management. *Personnel Psychology, 22*, 77-85.

Tiggermann, M. & Winefield, A. H. (1989). Predictors of employment, unemployment and further study among school-leavers. *Journal of Applied Psychology, 62*, 213-221.

Toole, D. L., Gavin, J. F., Murdy, L. B. & Sells, S. B. (1972). The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology, 25*, 661-672.

Tziner, A. & Dolan, S. (1982). Validity of an assessment center for identifying female officers in the military. *Journal of Applied Psychology, 67*, 728-736.

Van Leeuwen, E. (1956). Validity Information Exchange, No. 9-36. *Personnel Psychology, 4*, 381-382.

Vincent, N.L. & Dugan, R.D. (1962). Validity information Exchange, No. 15-03, *Personnel Psychology, 15*, 223-225.

Wagner, E.E. (1960). "Predicting Success for Young Executives from Objective Test Scores and Personal Data." *Personnel Psychology, 13*, 181-186.

Watson, J. & Williams, J. (1977). Relationships between managerial values and managerial success of black and white managers. *Journal of Applied Psychology, 62*, 203-207.

Westberg, W.C., Fitzpatrick, E.D., & McCarty, J.J. (1954). Validity Information Exchange. No. 7-073. *Personnel Psychology, 7*, 411-412.

Wollowick, H. B. & McNamara, W. J. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology, 53*, 348-352.

APPENDIX B

CODING KEY

1 2 3 4 5 6 7 8 9

STUDYNO YEAR DESIGN SPLTYPE CRITYPE RELTYPEC INOUT TSTYPE XVAL

EG. 001 75 2321120

DESIGN 1=CONCURRENT
2=PREDICTIVE
3=PREDICTIVE WITH SELECTION

SPLTYPE 1=PROFESSIONAL
2=SUPERVISOR
3=CLERICAL
4=SKILLED LABOUR
5=UNSKILLED LABOUR
6=SALES

RELTYPE 1=INTER-RATER
2=INTERNAL CONSISTIENCY
3=TEST-RETEST
4=PARALLEL FORMS
5=SPLIT HALF

CRITYPE 1=PERFORMANCE RATINGS
2=TURNOVER
3=ACHIEVEMENT/GRADE
4=PRODUCTION
5=STATUS CHANGE OR PROMOTION
6=WAGES
7=WORK SAMPLE
8=OTHER
9=COMPOSITE

XVAL 0=NO
1=YES

INOUT 0=IN
1=OUT
2=UNDECIDED

TSTYPE 0=NONPROJECTIVE CHARACTER AND PERSONALITY
1=INTERESTS
2=PROJECTIVE
3=GENERAL VOCATIONAL
4=LEADERSHIP AND MANAGERIAL BEHAVIOUR
5=MISCELLANEOUS

APPENDIX C

Sample-type sub-analysis

Results for the breakdown of sample-types into test-types from figure three.

Table 10

Validity of each test-type for Professional workers.

Professional			***Test***	***Type***		
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	2706	67	1261	-	398	691
Mean r	.17	.31	-.01	-	.10	.05
Mean Sample Weighted r	.23	.31	-.03	-	.11	.08
Variance	.01668	.00902	.05361	-	.01678	.02832
Mean r Corrected for Attenuation	.30	.40	-.04	-	.13	.11
Corrected Variance	.01896	-.0270	.08273	-	.00931	.03522
Chi Square	50.57 *	.74 NS	67.72 *	-	6.86	19.83
90% CI	.033-.57	-	-.60-.52	-	-.06-.32	-.26-.47

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 11

Validity of each test-type for Sales personnel.

Sales			***Test Type***			
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	676	-	173	194	-	-
Mean r	-.04	-	.38	.16	-	-
Mean Sample Weighted r	-.14	-	.33	.22	-	-
Variance	.05611	-	.00987	.02575	-	-
Mean r Corrected for Attenuation	-.17	-	.42	.28	-	-
Corrected Variance	.07671	-	-.0042	.02714	-	-
Chi Square	39.35 *	-	2.13 NS	5.49 NS	-	-
90% CI	-.72-.37	-	-	-.044-.60	-	-

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 12

Validity of each test-type for Skilled workers

Skill	***Test Type***					
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	664	-	-	-	-	709
Mean r	.35	-	-	-	-	.01
Mean Sample Weighted r	.33	-	-	-	-	.01
Variance	.01643	-	-	-	-	.00010
Mean r Corrected for Attenuation	.38	-	-	-	-	.01
Corrected Variance	.01584	-	-	-	-	-.0071
Chi Square	13.70 **	-	-	-	-	.08 NS
90% CI	.18-.58	-	-	-	-	-

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 13

Validity of each test-type for Supervisory workers

Supervisor	***Test Type***					
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	2706	67	1261	-	398	691
Mean r	.18	.31	-.01	-	.10	.08
Mean Sample Weighted r	.23	.31	-.03	-	.11	.08
Variance	.01668	.00902	.05361	-	.01678	.02832
Mean r Corrected for Attenuation	.30	.40	-.04	-	.13	.11
Corrected Variance	.01896	-.0271	.08273	-	.00931	.03522
Chi Square	50.57 *	.74 NS	67.72 *	-	6.86 NS	19.83 *
90% CI	.033-.57	-	-.60-.52	-	-.06-.32	-.26-.47

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 14

Validity of each test-type for Clerical workers.

Clerical	***Test Type***					
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	299	-	79	-	-	141
Mean r	.22	-	.16	-	-	.26
Mean Sample Weighted r	.19	-	.21	-	-	.25
Variance	.01884	-	.02645	-	-	.00147
Mean r Corrected for Attenuation	.24	-	.27	-	-	.32
Corrected Variance	.01034	-	.00451	-	-	-.0186
Chi Square	6.04 *	-	2.29 NS	-	-	0.24 NS
90% CI	.04-.44	-	.14-.40	-	-	-

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Appendix D

Criterion-type sub-analysis

Results for the breakdown of criterion-types into test-types from figure four.

Table 15

Validity of each test-type for Performance criteria.

Performance	*Test Type*					
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	4231	773	10654	375	285	10224
Mean r	.17	.10	.19	-.02	.08	.12
Mean Sample Weighted r	.19	.11	.20	-.02	.10	.27
Variance	.02302	.01108	.00767	.01172	.02305	.00752
Mean r Corrected for Attenuation	.23	.12	.26	-.03	.12	.33
Corrected Variance	.02282	-.0001	.00993	.00611	.01622	.00915
Chi Square	104.54 *	8.76 NS	89.12 *	4.40 NS	6.71 NS	89.02 *
90% CI	-.07-.52	-	.07-.46	-.18-.13	-.13-.37	.14-.51

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 16

Validity of each test-type for Turnover criteria.

Turnover	*Test Type*					
	NCP	*INT*	*VOC*	*PROJ*	*MAN*	*MISC*
Sample Size	-	-	308	258	-	292
Mean r	-	-	.24	.24	-	.17
Mean Sample Weighted r	-	-	.01	.24	-	.12
Variance	-	-	.02285	.00000	-	.02164
Mean r Corrected for Attenuation	-	-	.12	.31	-	.15
Corrected Variance	-	-	.01657	-.01156	-	.01353
Chi Square	-	-	7.17 NS	0.00 NS	-	6.49 NS
90% CI	-	-	-.13-.38	-	-	-.08-.38

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 17

Validity of each test-type for Achievement criteria.

Achievement	*Test Type*					
	NCP	*INT*	*VOC*	*PROJ*	*MAN*	*MISC*
Sample Size	-	-	154	-	-	146
Mean r	-	-	.24	-	-	.17
Mean Sample Weighted r	-	-	.10	-	-	.12
Variance	-	-	.02285	-	-	.02164
Mean r Corrected for Attenuation	-	-	.12	-	-	.15
Corrected Variance	-	-	.01657	-	-	.01353
Chi Square	-	-	3.59 NS	-	-	3.25 NS
90% CI	-	-	-.13-.38	-	-	-.08-.38

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 18

Validity of each test-type for Production criteria

Production	***Test Type***					
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	-	-	345	-	-	-
Mean r	-	-	-.03	-	-	-
Mean Sample Weighted r	-	-	-.16	-	-	-
Variance	-	-	.13342	-	-	-
Mean r Corrected for Attenuation	-	-	-.20	-	-	-
Corrected Variance	-	-	.21313	-	-	-
Chi Square	-	-	48.37 *	-	-	-
90% CI	-	-	-1.1-.70	-	-	-

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 19

Validity of each test-type for Status change criteria.

Status	***Test Type***					
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	12988	-	-	-	-	7923
Mean r	.20	-	-	-	-	.10
Mean Sample Weighted r	.18	-	-	-	-	.01
Variance	.00493	-	-	-	-	.00002
Mean r Corrected for Attenuation	.23	-	-	-	-	.12
Corrected Variance	.00617	-	-	-	-	-.0004
Chi Square	68.30 *	-	-	-	-	20.16 *
90% CI	.01-.38	-	-	-	-	-

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 20

Validity of each test-type for Wages criteria.

Wages	***Test Type***					
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	-	-	-	201	-	-
Mean r	-	-	-	.05	-	-
Mean Sample Weighted r	-	-	-	.06	-	-
Variance	-	-	-	.00087	-	-
Mean r Corrected for Attenuation	-	-	-	.07	-	-
Corrected Variance	-	-	-	-.01520	-	-
Chi Square	-	-	-	0.17 NS	-	-
90% CI	-	-	-	-	-	-

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 21

Validity of each test-type for Other criteria.

Other	*Test Type*					
	NCP	*INT*	*VOC*	*PROJ*	*MAN*	*MISC*
Sample Size	399	-	-	-	-	132
Mean r	.33	-	-	-	-	-.05
Mean Sample Weighted r	.33	-	-	-	-	-.05
Variance	.09464	-	-	-	-	.10890
Mean r Corrected for Attenuation	.46	-	-	-	-	-.06
Corrected Variance	.11490	-	-	-	-	.15598
Chi Square	51.86 *	-	-	-	-	14.4 *
90% CI	-.20-1.12	-	-	-	-	-.83-.70

* = $p<.005$ ** = $p<.05$ NS = Non Significant

Table 22

Validity of each test-type for Composite criteria.

Composite	***Test Type***					
	NCP	***INT***	***VOC***	***PROJ***	***MAN***	***MISC***
Sample Size	234	-	-	-	-	-
Mean r	.04	-	-	-	-	-
Mean Sample Weighted r	.03	-	-	-	-	-
Variance	.03498	-	-	-	-	-
Mean r Corrected for Attenuation	.04	-	-	-	-	-
Corrected Variance	.01028	-	-	-	-	-
Chi Square	8.20 NS	-	-	-	-	-
90% CI	-.16-.23	-	-	-	-	-

* = $p<.005$ ** = $p<.05$ NS = Non Significant