Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

TRAINABILITY ASSESSMENTS AND WORK SAMPLES:

A FIELD STUDY AND A META-ANALYSIS.

A thesis presented in partial fulfillment of

the requirements for the degree of

Master of Arts in Psychology

at Massey University

Karl PAJO

1987

This Thesis is Dedicated

to Faye and Karl

ABSTRACT

This study compared a work sample test with a trainability test for the prediction of typing students grades. A meta-analysis of the work sample literature was also carried out. Participants in the work sample trainability test comparison were 89 female first year Polytechnic typing students. Students were randomly assigned to either the work sample group or the trainability test group. Tutors then administered the relevant predictor and data was collected. Scores on the predictors were later correlated with the students grade in their second terms test. All the obtained correlations were found to be highly significant although the results unexpectedly revealed that the error score on the work sample was the best predictor overall. It was suggested that the tutors inexperience in administering trainability tests, their greater familarity with work samples and certain deficiencies in the criterion may have contributed to the unexpected trend in the data. Meta-analysis was used to cumulate and average results from many different studies which examined work samples. Studies which utilised training criteria were analysed seperately from those which employed job proficiency criteria. Results from the analysis showed substantial remaining variance following correction for statistical artifacts. The studies were then grouped according to Robertson and Kandola's (1982) classification of work samples in order to identify potential moderator effects. Meta-analysis of subgroups revealed that for all categories, with the exception of group discussion/decision making, considerable variance still remained following correction for statistical artifacts. It is suggested in the

discussion that further research on work samples is required, particularly the development of a classificatory system which can accurately and reliably distinguish between types of work samples. Possibilities for future research on trainability tests are also explored.

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr Mike Smith for his encouragement, advice and patience throughout this study. I would also like to express my gratitude to Dave George who was always willing to provide assistance. My thanks also go to Dr John Spicer who helped resolve several knotty statistical problems.

I am grateful to the staff of the Manawatu Polytechnic who freely gave of their time and facilities and enabled this study to take place.

Finally, it is with deep appreciation that I acknowledge the support of my parents and all my friends. Larry and Sharon who often had to put up with me until the early hours of the morning. Mike, Julie, Steven (Thumis), Fiona, Steven (Sudagura), Chris, Mark, Wendy, Yvonne, Scott, Josie, Sharon, and Gerald with whom many good times were had. Also my ex-flatmates, Mark, Leeanne, Joanne and Jin who tolerated my many foibles.

TABLE OF CONTENTS

Page

LIST OF APPENDICES

Page

APPENDIX

LIST OF TABLES

Page

LIST OF FIGURES

Page

FIGURE

CHAPTER ONE.

INTRODUCTION

Personnel psychology, which constitutes the psychology of personnel decisions (Landy, 1985), has always been a traditional area of interest for many industrial and organisational psychologists. As early as 1917 tests were being developed to predict the success of employees on the job. Both the first and second world wars provided further impetus for the testing movement as a whole and allowed personnel psychologists the opportunity to develop and refine their skills (Grant, 1980). The use of tests and other selection procedures burgeoned and they became increasingly sophisticated.

Hakel (1986) points out in his review of personnel selection that there has been substantial progress in the last few decades. He notes that in the early 1960's personnel selection could best be considered "pragmatic, empirical, and atheoretical." Since then personnel research has diversified and other facets of selection have begun to receive some of the attention they merit. Tenopyr and Oeltjen (1982) note that there was a long overdue upsurge in research on job analysis. Hakel's (1986) review indicates that this research continues unabated. The measurement of performance, particularly the cognitive processes underlying performance assessments, has been a topic that has also generated a substantial body of literature. Utility analysis (e.g. Hunter and Hunter, 1984; Schmidt, Hunter, McKenzie, & Muldrow, 1979)

has developed to the point where psychologists can safely describe the monetary savings to be gained from the implementation of efficacious selection procedures. The requirement in many countries that selection should not adversely affect minority groups has also prompted a significant amount of work on the part of numerous psychologists. Meta-analytic procedures have now been developed and offer a comprehensive means for assessing cumulative results. These developments have been paralleled by the continuing refinement of existing predictors and the development of alternative selection strategies.

Today the personnel practitioner has a considerable array of selection tools available from which to make a choice. While the techniques available have multiplied, the fundamental goal of personnel selection has remained unchanged. The primary objective is still the identification of those applicants who best suit the organisation's requirements. This usually involves the prediction of an applicant's likelihood of success on the job or during training. Unfortunately, despite continuing research and development, the reliability and validity of many predictors can hardly be considered reassuring.

Work samples however, offer a promising alternative selection strategy which has been receiving increasing attention from psychologists and others interested in matching people to jobs. Historically, research on work samples has tended to be rather sparse and fragmentary. Downs (Note 1) has attributed this to several factors. The zeitgeist of the time was one that favoured the development of predictors that conformed

with traditional psychometric test properties. There was an implicit belief amongst psychologists that the predictor should be different from the criterion (Wernimont & Campbell, 1968). Psychological tests were to be used as indicators of predispositions to behave in certain ways rather than being regarded as examples of the typical behaviour of individuals (Robertson & Kandola, 1982). This notion was contested when Wernimont and Campbell (1968) argued that validity would be enhanced if predictors resembled more closely the criterion behaviour they were attempting to predict. They described this concept as behavioural consistency and it is what Asher and Sciarrino (1974) have labelled point-to-point correspondence between predictor and criterion space.

Other more practical difficulties also tended to preclude the widespread acceptance of work samples when used for the prediction of competency on the job. Work samples are disadvantaged in that they have to be individually designed and validated for each particular job. Furthermore, they are often costly to set up, particularly if complex machinery or simulations are required, and are also expensive in terms of manpower and materials used. In contrast paper-and-pencil tests can be used off the shelf, are generally easier to administer and cheaper. Given that such tests were considered capable of measuring those abilities important to performance on the job it is not surprising that work samples were regarded with suspicion (Downs, Note 1).

A further bar to the ready acceptance of work samples as predictors was their past history of use in organisations. Work samples had

traditionally been associated primarily with the verification of the acquisition of skills (Downs, Note 1). Essentially their role was to function as an achievement test assessing competence and providing a basis for certification. In some cases they were also used as a criterion to validate other selection methods.

Thus, the practical problems in developing and administering work samples, their past history of use predominantly as criteria to which organisations and individuals had become accustomed, and the adherence to traditional psychometric test concepts, all conspired to inhibit the propagation of work samples as a selection tool. However, proponents of work samples refused to be deterred and continued with a great deal of enthusiasm to investigate this potentially useful predictor. Their work generated a great deal of interest and a considerable body of literature (see reviews by Asher & Sciarrino, 1974; Howard, 1985; Robertson & Kandola, 1982; Downs, Note 1; Gill, 1979; Karren, 1980; Robertson & Downs, 1979; Hunter & Hunter, 1984).

There are many reasons for this upsurge in research on work samples, albeit that it has been somewhat sporadic at times. The gradual alteration of the belief that predictors and criterion should be different has been alluded to. Tests were used to sample behaviour based on the premise that the best indicator of future performance is past performance. Wernimont and Campbell (1968) and Asher and Sciarrino (1974) argued that prediction of future behaviour would be facilitated if tests more closely resembled the behaviour to be predicted (the criterion). Thus, the notion of behavioural consistency

or point-to-point correspondence was established. The development of work samples was also given a boost by the failure on the part of more established selection procedures to reach acceptable levels of validity despite in some cases, several decades of research and experimentation.

One prominent example of a commonly used selection device is the employment interview. There have been several reviews of its reliability and validity. One of the earliest was that carried out by Wagner (1949). He reported a median validity coefficient of .27 for the 22 studies reviewed. Reilly and Chao (1982) summarised the data from 12 studies and came up with a mean validity coefficient of only .19. Such disappointing results have since become typical of research in the field (e.g. Mayfield, 1964; Ulrich & Trumbo, 1965; Wright, 1969; Schmitt, 1976). Recent reviews (Arvey, 1979; Arvey & Campion, 1982; Reilly & Chao, 1982) have also pointed to the interviews susceptibility to bias and distortion and particularly the fact that it may act as a vehicle for discrimination against women and minority group members. All in all, the literature suggests that the interview may not be an efficacious method for selecting personnel.

Similar conclusions can be reached regarding the use of references. Reference reports are commonly requested by many organisations (Muchinsky, 1979). The veracity of reference reports however, is questionable and their widespread use difficult to justify. In a review of the available literature Muchinsky (1979) concludes that reported validity coefficients ranged from unacceptable to mediocre. Reilly and Chao (1982) report an average validity coefficient for the

studies they reviewed of .14. It appears that reference reports are unlikely to contribute appreciably to the validity of employee selection decisions.

Psychological tests are commonly used for personnel selection in occupational settings. Ghiselli (1973) reviewed the validity of aptitude tests during the period 1920 to 1971. Tests were classified into broad categories. These included;

a) tests of intellectual ability

b) tests of spatial and mechanical ability

c) tests of perceptual accuracy

d) tests of motor ability

e) tests evaluating personality and/or interests.

The average predictive validity for each category of test was then calculated for different occupational groupings. Average validity coefficients for both training and proficiency criteria rarely exceeded .30. However, as Ghiselli (1973) notes, various artifacts such as restriction of range in predictor and criterion scores, errors in sampling, and unreliability of predictor and criterion measures would mean that such estimates are likely to be conservative hence underrating the true predictive power of the tests.

A more recent review incorporating meta-analytic procedures capable of correcting for such artifacts was conducted by Hunter and Hunter (1984). Using formulas developed by Hunter, Schmidt, and Jackson (1982) they were able to correct the variance across different studies

for sampling error and wherever possible also corrected for the effects of error of measurement and range restriction. Ghiselli's (1973) review of ability tests was reanalysed using these more sophisticated procedures. The conclusions reached by the authors was that most of the variance in results across studies was due to sampling error. Furthermore, the validity figures they computed were markedly higher than those obtained by Ghiselli (1973). The average validity of cognitive ability tests for different job families ranged from .27 to .61. The average validity for tests of psychomotor ability ranged from .17 to .44. Multiple correlations computed using combined cognitive and psychomotor ability scores tended to be uniformly high across all the job families. Excluding the job of sales clerk, the validity for the combined tests of ability ranged from .43 to .62. Hunter and Hunter (1984) determined that the average validity of cognitive and psychomotor ability tests combined was .53.

In addition to Ghiselli's (1973) study Hunter and Hunter (1984) reanalysed the data from several other reviews. Relevant figures are presented in table one.

TABLE 1 META-ANALYSIS DERIVED FROM (A) DUNNETTE (1972) (B) REILLY AND CHAO (1982).

PREDICTORS	No OF CORRELATIONS	AVERAGE VALIDITY
A Cognitive Ability	215	.45
Perceptual Ability	97	.34
Psychomotor Ability	95	.35
Biographical Inventories	115	.34
Interviews	30	.16
Education	15	.00
Job Knowledge	296	.51
Job Tryout	20	.44
B Biographical Inventory	44	.38
Interview	11	.23
Expert Recommendation	16	.21
Reference Check	7	.17
Academic Achievement	10	.17
Self Assessment	7	some
Projective Tests	5	little
Handwriting Analysis	3	none

from Hunter and Hunter (1984)

The table shows that in general, most selection instruments are poor predictors. The major exceptions appear to be tests of ability, biographical inventories, and work samples.

The use of biographical data appears to be a promising approach to selection. However, there are shortcomings associated with its use. Empirically keyed biodata scores are prone to attenuation of validity over time (e.g. Wernimont, 1962; Hunter & Hunter, 1984). Furthermore, unless a cross validated research design is used, the process of deriving a biographical inventory is one that is prone to massive capitalization on chance (Hunter & Hunter, 1984). It is also possible that applicants may intentionally falsify their responses (e.g. Goldstein, 1971).

A straightforward appraisal of the value of work samples has been impeded by the failure on the part of researchers to come to a clear agreement about what actually constitutes a work sample. Depending on the text consulted, one's impression of a work sample may differ substantially (e.g. Cronbach, 1966; Guion, 1965). This confusion in the literature has persisted (e.g. Howard, 1983; Landy, 1985) although attempts at a rapprochement have been made (Downs, Note 1; McCormick & Tiffin, 1976; Thornton & Byham, 1982). The basis for most disagreement has centreed on how broad or narrow the definition of a work sample should be. Downs (Note 1) abstracted what she considered were the key features of work samples commonly agreed upon by most writers in the field. Using those features she derived a definition of a work sample test as

> "a performance test based on work or job related elements, the design of which allows for measurement or objective assessment of the skills involved in all, or crucial aspects of the job. This measurement may be used to measure past learning or predict potential to learn in the future." (page 2)

Such a definition is quite broad in scope and would include a variety of tasks or tests that vary along a dimension of "fidelity" or relatedness to actual work performance. Examples of work samples could thus range from business games, in-basket tests, leaderless group discussions, through to trainability tests, job simulations and measurement of performance at the job station. Acceptance of such a definition would go a long way towards clearing up many of the misunderstandings currently rife in the literature and would set the field on a theoretically sounder basis.

Useful distinctions within the domain of work samples can still be made. For example, Asher and Sciarrino (1974) classify work samples as either motor or verbal. A motor work sample is a task involving the manipulation of things (e.g. performance on an aircraft simulator, piecing together an electronic circuit board). A verbal work sample is a task containing problems which are primarily people or language oriented (e.g. an in-basket test, leaderless group discussion). Their review demonstrated that motor work samples were superior in predictive power to all other predictors except for biographical data. The verbal work sample tended to be consistently less efficient in its ability to forecast job proficiency than the motor work sample but was still superior to most of the other predictors. When the relevant criterion was changed to "success in training" then the verbal work sample was

clearly superior to the motor work sample.

Robertson and Kandola (1982) differentiate between four categories of work sample;

1) Psychomotor. Tasks involving the physical manipulation of objects.

2) Individual, situational decision making. Tasks in which the applicant is required to make decisions similar to those made in the job being tested for. This category can vary along a dimension of realism with close approximations being in-basket tests while more abstract cases could involve the presentation of hypothetical situations and asking the applicant how he/she would respond.

3) Job-related information. Typically paper-and-pencil tests, their purpose is to evaluate applicant knowledge in areas considered to be directly relevant to work performance.

4) Group discussions/decision making. A group of individuals are required to discuss a particular topic and their performance during the discussion is assessed.

They found that psychomotor work samples and job-related information tests had the highest median validity coefficients (.39 and .40 respectively) and the greatest proportion of coefficients above .40. Situational decision making was the poorest of the four categories with the lowest median validity coefficient (.28), the greatest proportion of coefficients below .30 and the smallest proportion above .40. Comparison to other psychological tests showed that psychomotor work samples were superior to all other types except for biographical data.

Group discussion measures also produced quite high validity coefficients in comparison with other tests. An interesting feature of Robertson and Kandola's (1982) analysis is that the high validities obtained by job-related information tests seemed to be mainly confined to situations where training criteria were used. When one considers only the criteria of job performance then the median validity coefficients for psychomotor, group discussion and situational decision making work sample tests outstripped those of job-related information tests.

Hunter and Hunter (1984) compared a number of alternative predictors of job performance using meta-analytic procedures. Abstracting data from many studies, including other meta-analyses, and using the criterion of job performance, as measured by supervisor ratings, they compared predictors used for entry level jobs where training followed hiring and predictors used for decisions regarding promotion or certification. Work samples were second only to an ability composite in predictive power for entry level jobs (mean validity of .44). They were the most efficient predictor used for promotion or certification decisions (mean validity of .54).

Gordon and Kleiman (1976) have conducted one of the few studies which has directly compared a work sample with a standardised test. They used recruits from three separate classes that attended a police training academy. The training program was a 20 week course during which recruits were instructed in the fundamentals of police work. Recruits were administered a work sample after approximately two weeks

on the course which covered areas such as introduction to law enforcement, the relationship of the police department to other civic agencies, department rules and regulations and organisation of the department. The recruits were also administered a standardised intelligence test. Correlations with the trainability criterion (sum of the grades achieved during the training course) revealed that in all cases the work sample achieved significant validity coefficients (ranged from .52 to .72) whereas only one of the validity coefficients for the intelligence test was significant (range from .15 to .56).

Mount, Muchinsky, and Hanser (1977) compared the predictive and concurrent validity of a work sample with two traditional paper-and-pencil tests under controlled laboratory conditions. The work sample consisted of following a diagram and constructing a model from mechanical parts. The criterion was the assembling of a more complex model. Using the number of parts correctly assembled as the dependent measure the authors found that in all cases the validity of the work sample was higher than that of the paper-and-pencil tests. Furthermore, even when all three predictors were combined using multiple regression there was only a slight improvement in the validity coefficient obtained over and above that of the work sample alone.

Similar results have been obtained by Sylvia Downs in her work on trainability assessments. An early study (Downs, 1970) involved the development of a trainability test for sewing machinists in a children's clothing factory. The company's existing selection procedures (a form board and a pin board) were compared with the

trainability test. The trainability test was highly predictive of success at the end of training while the other selection procedures failed to achieve any significant predictive validity. The results were so convincing that the company immediately terminated its old selection methods and embraced the new test whole-heartedly.

Smith (1977) compared university entrance examination marks, the mechanical aptitude test and the space relations test from the Differential Aptitude Test Battery and a specially designed trainability test for the prediction of the practical performance of dental students. Using students from three separate academic years he found that the trainability assessment was highly correlated with performance on a combined criterion of conservation test marks and final conservation exam marks. In fact, the trainability test surpassed all other predictors except for the DAT mechanical reasoning test.

Siegal and Bergman (1975) constructed trainability tests (what they called miniaturised job training and evaluation) and compared them with standard US Navy paper-and-pencil tests for the prediction of performance by low aptitude naval recruits. Scores on six trainability assessments and three navy tests were correlated with judges' ratings of recruits' performance on several job related tasks after nine months' fleet experience and after 18 months' fleet experience. For the first follow-up, five of the six job performance criteria were predicted better by the trainability tests than by the navy selection tests. For the second follow-up, some attenuation of predictive power

for the trainability tests was apparent and the navy predictors were superior for five of the six criterion tasks. Siegal and Bergman (1975) note that such attenuation is not unusual and may simply reflect that the trainability tests are more appropriate for predicting success on initial job entry rather than subsequent improvement. Cohen and Penner (1976) have expressed some reservations regarding the methodology of Siegel and Bergman's (1975) study, particularly the failure to cross-validate the predictors and the large number of drop-outs in the sample used.

Other authors have used trainability tests and although they were not compared with alternative predictors, high validity coefficients have been reported for such diverse jobs as carpentry (Robertson & Mindel, 1980), welding (Downs, 1968; cited in Robertson & Downs, 1979; Robertson & Mindel, 1980), fork truck operating (Downs, 1972), electronic assembling (Smith, 1972), industrial sewing (Downs, 1972), metal use and fitting (Smith & Downs, 1975), brick laying, capstan operating and centre lathe turning (Robertson & Mindel, 1980), catering and forestry work (figures reported in Downs, Note 1), and naval recruits training to be firemen, seamen and airmen (Siegal, 1983).

Changes in beliefs about the functions work samples can fulfil and the clear demonstration in many studies of their superiority over other predictors coupled with the high validity coefficients attained have served to popularise work samples as a viable selection procedure. Concomitant with this was a rise in interest about fairer selection spurred on in many cases by legal changes and social pressures

requiring that tests should not exhibit adverse impact. Researchers interested in work samples were able to capitalize on this since in many cases traditional selection procedures were proving to unfairly discriminate against women and/or ethnic minorities (see Arvey, 1979; Einhorn & Bass, 1971). It was argued that work samples would not be prone to such effects since nothing could be fairer than selecting an applicant based upon his or her performance on a sample of the work he or she would actually be required to do. Several studies have subsequently confirmed that work samples do appear to be a fair method of selection.

Schmidt, Greenthal, Hunter, Berner, and Seaton (1977) compared a work sample for metal trades skills with a well constructed content-valid, written achievement test for the same technical area. The written achievement test and each of its component subtests showed large and significant minority-majority differences. The work sample showed a considerably smaller difference between minority-majority workers. Schmidt et al (1977) explain that the small gap exhibited was primarily due to differences in work speed and suggest that this minimal amount of adverse impact could be reduced by decreasing the weighting of the work speed sub-score in the work sample. The authors also point out that both minority and majority examinees saw the job sample tests as significantly fairer, clearer, and more appropriate in difficulty level.

Hamner, Kim, Baird, and Bigoness (1974) conducted a laboratory study in which they examined the way the sex and race of the rater and the sex

and race of the ratee influence assessments of ratee performance on a simulated work sampling task. Their results suggested that sex-race stereotypes do influence assessments of behaviour on a work sampling task although unexpectedly the ratings of women's performance were inflated rather than deflated. Brugnoli, Campion, and Basen (1979) criticised the research of Hamner et al (1974) on the grounds that the work sample selected failed to represent important performance factors and hence may have encouraged raters to rely on stereotypes when evaluating applicants. They also argue that Hamner et al (1974) should have used an evaluation device more specific to the behaviours being observed rather than a global rating scale and non-behavioural anchors. They then designed an experiment to examine the role of evaluation specificity and task relevance in explaining racial bias in the use of work samples. They found that bias was not evident when subjects used behavioural recording forms or when evaluations were based on observations of relevant job behaviour. They conclude that if work samples are carefully developed and raters focus on and record relevant behaviour then the potential for bias in the use of work samples appears small.

Bray and Howard (1983; cited in Howard, 1983) report large racial differences when the paper-and-pencil School and College Ability Test was used. Use of an in-basket exercise showed considerably less adverse impact, and performance in group discussions showed almost none. Cascio and Phillips (1979) constructed motor and verbal work sample tests for use by a US city government. No significant difference in selection rates for minority versus majority workers was

reported. Downs (1970) found no significant difference in the trainability assessment ratings and criterion ratings of United Kingdom applicants and overseas applicants for the job of sewing machinist. She also notes that there was a high degree of agreement between ratings of overseas applicants on the trainability test and their criterion performance ratings . However, separate validity data for the two groups was not presented and, as Robertson and Kandola (1982) point out, studies that fail to report validity data are only of limited use. Some exceptions are studies by Grant and Bray (1970), Field, Bayley, and Bayley (1977), and Kesselman and Lopez (1979) who all report improved validity accompanying the use of appropriate work sample tests and reduced adverse impact for minority groups (see Robertson and Kandola, 1982 for a review).

Favourable applicant reaction and other ancillary functions of work samples have also contributed to their increased usage. Downs (1970) reports that both instructors and applicants preferred a trainability test over existing selection procedures. Instructors liked the test because they felt more involved in the selection procedure. Applicants liked the test because they felt it was fair and enabled them to demonstrate their capabilities. Schmidt et al (1977) reported that both minority and majority subjects in their study considered the job sample test as significantly fairer, clearer, and more appropriate in difficulty level than a written test covering the same content area.

There is also some evidence that work samples could function as realistic job previews. Wanous (1977) in a review of realistic job

previews has concluded that they allow applicants to make more informed choices hence diminishing subsequent dissatisfaction and increasing the probability that applicants will remain on the job. While not all the literature is consistent with such a view (e.g. see Reilly, Brown, Blood & Malatesta, 1981), studies using work samples do seem to offer some support. Downs, Farr & Colbeck (1978) examined the data from sewing machinist trainability tests administered throughout the United Kingdom during the period 1973-1975. All applicants who sat the test were invited to start work regardless of the grade received. The authors found that the individual's trainability assessment grade (ranging from A-highest to E-lowest) influenced the decision about whether or not to start work. Fully 90.8% of those graded A accepted the companies offer while 81.1% of those graded B, 75.6% of those graded C, 54.6% of those graded D and only 23.1% of those graded E accepted offers of employment. The evidence suggests that the trainability tests allowed applicants to accurately gauge their own performance and encouraged self-selection based on those judgments.

Farr, O'Leary & Bartlett (1973) found that for white subjects the administration of a pre-employment work sample resulted in more accurate expectancies about task requirements and a commensurately lower voluntary turn-over rate. The failure to find similar results for black subjects was explained in terms of the differential importance of factors in the work situation. It was argued that black applicants may have paid more attention to such facets of the environment as pay and interpersonal relations whereas whites may have focussed exclusively on the task related factors portrayed in the work

sample.

Additional support for the notion that work samples may encourage the self assessment of ability comes from a study by Downs (note 2). She administered a trainability assessment to Royal Navy helicopter pilots. Results of the assessment showed that pilots' ratings of their own abilities were clearly affected by the trainability test. She concluded that the test helped applicants to judge whether or not they would like the job and enabled those who did well to assess themselves more realistically.

Campion (1972) asserts that work samples may have additional advantages of reducing the possibility of response sets and faking and being less prone to charges of invasion of privacy. While such a claim seems inherently plausible the paucity of relevant data in the literature means that such statements remain to be substantiated.

The use of work samples is not completely without drawbacks. Several authors (e.g. Downs, Note 1; Howard, 1983; Robertson & Downs, 1979; Smith & Downs, 1975) have enumerated their disadvantages. These include the fact that work samples (particularly psychomotor work samples, job-related information tests and trainability tests) tend to be job specific. This means that they have to be individually designed and validated for different jobs. Furthermore, they require continual monitoring in order to ensure that their reliability and validity is not affected by changes in job content over time. In cases where machinery is required it can be costly to set up or to construct

appropriate simulations. Many work samples can only be administered individually or in small groups and require skilled assessors to evaluate performance. They usually take longer to administer and also use more materials than equivalent paper-and-pencil selection tests. Howard (1983) also notes that work samples may not be particularly useful for assessing a candidates' range of knowledge. Finally, some studies (Downs, 1977; Siegal & Bergman, 1975; Smith & Downs, 1975) suggest that the predictive validity of work samples may be prone to attenuation over time (perhaps due to changes in job content as noted above). However, as Hunter and Hunter (1984) point out, there exist very few predictors which do not become less efficient with the passage of time.

The present study elaborates on previous research examining the value of work samples. More specifically it is composed of two parts. The first involves a direct comparison of the predictive validity of a work sample and a trainability test. Robertson and Downs (1979) distinguish between standard work samples and trainability tests. A trainability test is a specialised type of work sample designed to evaluate an applicant's potential to **learn** a task or to succeed in **training**. Such tests typically include standardised instructions and a period of demonstration during which the instructor teaches the applicant the task. While the applicant is being instructed in the task he or she is permitted to ask questions and to practice. The applicant is then tested on the material he or she has been taught by being asked to perform the task unaided. The applicant's performance on the task is assessed by the instructor who uses a standardised error checklist and

rating scale. Thus the trainability test differs from the normal work sample in several important ways.

1) it incorporates a structured learning period during which the applicant is encouraged to ask questions and practice the task.

2) the assessor uses an error checklist rather than simply evaluating the product of performance on the work sample.

3) the applicant is only tested on what he or she has been taught during the learning period, hence, it does not assume any prior experience.

Research on trainability tests has shown that they are very good predictors of success in training and often subsequent performance on the job (see Robertson & Downs, 1979 for a review). The question remains as to whether or not trainability tests tap important performance dimensions that work samples do not. Trainability tests tend to be more complex and time consuming than equivalent work samples. Employers may be reluctant to accept trainability tests on face value unless it can be clearly demonstrated that they are superior to work samples for predicting training outcomes. A study by Gordon and Kleiman (1976) found that performance on a work sample administered to police recruits was significantly related to grades achieved at the end of training. In other words, Gordon and Kleiman (1976) were able to predict trainability using a work sample. While such evidence is suggestive it is by no means conclusive. There has been no study to date that has specifically compared a work sample with a trainability test.

Statement of Hypotheses.

For the first part of the present study it is hypothesised that a trainability test designed to predict training success for typing students will prove to be superior to a work sample administered for the same purpose. Such a hypothesis is based upon the fact that trainability tests are specifically designed to forecast training outcomes and their prior history of success in that endeavour.

The second part of the study is a partial replication and extension of work done by Robertson and Kandola (1982). It consists of an examination of the predictive validity of different types of work samples. Robertson and Kandolas' (1982) categorization of work samples will be used with the addition of a separate trainability test group. More sophisticated meta-analytic formulas will be used to analyse the data rather than simply calculating distributions of validity coefficients and median validity coefficients.