DOCUMENT RESUME

ED 260 324 CG 018 417

AUTHOR Malouff, John M.; Schutte Nicola S.
TITLE A Review of Validation Research on Psychological Variables Used in Hiring Police Officers.
PUB DATE Apr 85
NOTE 73p.; Paper presented at the Annual Convention of the Rocky Mountain Psychological Association (Tucson, AZ, April 24-27, 1985).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Intelligence Tests; Interest Inventories; Interviews; Literature Reviews; Personality Measures; Personnel Data; Personnel Evaluation; *Personnel Selection; *Police; *Psychological Characteristics; *Research Methodology; *Research Problems; *Validity
IDENTIFIERS Minnesota Multiphasic Personality Inventory

ABSTRACT

This paper reviews the methods and findings of published research on the validity of police selection procedures. As a preface to the review, the typical police officer selection process is briefly described. Several common methodological deficiencies of the validation research are identified and discussed in detail: (1) use of past-selection research designs; (2) inappropriate comparison groups; (3) non-meaningful outcome variables; (4) alpha-inflated analysis; (5) over-emphasis of beta weights; and (6) the search for moderator variables. Validity evidence for several types of selection variables is discussed including biodata, measures of intellect, personality measures, interviews, interest inventories, and subjective background ratings. Of the 14 biodata categories researched only 5 were validated as predictors of poor police performance (prior involuntary termination, criminal and vehicle code convictions, having been married more than once, and short duration of prior jobs). Measures of intellect, subjective background ratings, and personality measures provided mixed evidence of validity. Some scales of the Minnesota Multiphasic Personality Inventory were found to have post-selection validity in more than one study. There was no meaningful evidence of validity for interest inventories or interviews as police selection procedures. (MCF)

A Review of Validation Research on Psychological Variables Used in Hiring Police Officers

John M. Malouff

University of Southern Colorado and
Medical University of South Carolina

and

Nicola S. Schutte

University of Southern Colorado

Requests for reprints should be sent to John Malouff, Psychology Department, University of Southern Colorado, Pueblo, CO 81001-4901

Abstract

The methods and findings of police-selection validation studies were reviewed. Several common methodological deficiencies were identified, involving (a) predictor data collected after the officers were hired, (b) inappropriate comparison groups, (c) contaminated outcome-variables, (e) non-meaningful outcome variables, (f) alpha-inflated analyses, (g) overemphasis of beta weights, and (i) the search for moderator variables. The validity evidence for several types of selection variables, including biodata, measures of intellect, measures of personality, interviews, interest inventories, and subjective background ratings was found to range from scant to promising.

In most jobs the costs of making a mistake in hiring involve low productivity, theft and early turnover. With police officers, however, the costs can reach monumental levels. For when an officer makes a mistake, someone may die, when an officer acts oppressively, trust in government may erode, and when an officer quits or is fired, $10,000 or more in training costs (Mills & Stratton, 1982) may be lost. Hence, the selection of police officers is or should be a central concern of all of the thousands of police agencies across the nation, including municipal police forces, sheriffs' offices, state patrols, military police and various federal agencies such as the Federal Bureau of Investigation.

Because of the importance of hiring good police officers, researchers have given considerable attention to the matter. The first reported study of the validity of a selection method was by Martin in 1923, and since then more than 40 reports of studies have been published.

Brief reviews of the earlier studies can be found in Cohen & Chaiken (1972), Ghiselli (1966, 1973), Kent & Eisenberg (1972), Lefkowitz (1977), Poland (1978), and Speilberger, Ward & Spalding (1979). Also, Henderson (1979) reviewed the validity of personality and aptitude scales, and Sparling (1975) reviewed the validity of education as a selection criterion.

However, many research design, statistical and report problems with the research were not discussed in any of the reviews (e.g., use of inappropriate comparison groups). Further, none of the reviews contained an in-depth examination of the different types of outcome and

predictor variables that have been used in the research. Finally, the reviews did not include any of over a dozen validation studies reported in the past few years. Sadly, many of the recent studies are replete with the same methodological errors made in prior research.

The purpose of this review is to provide a careful examination of the methods that have been used in research on police-selection and an up-to-date review of the results of research on all the types of psychological selection variables examined in the studies.

This review includes only published reports. Unpublished reports were excluded because they are difficult or impossible to obtain and because they have not been subjected to the scrutiny of the scientific community and so may contain methodological deficiencies that escape any single reviewer (see e.g., the scathing but different reviews of the Humm & Humm (1950) study by Blum (1964, pp. 106-107) and Ruch (1965)).

As a preface to the review, an outline of the typical police officer selection process is in order. Abramson (1974), in a survey of 266 college or university police departments, found that the vast majority required that applicants be at least 21 years old, be a high school graduate, have no criminal record, and pass a background investigation of their morals and personal character. Almost every department used an interview with detective or administrators as a screening device. A significant number of departments also used a test measure of intellect, a clinical interview by a psychologist/psychiatrist and/or a psychological test. Parish, Rios & Reilly (1979) surveyed 130 police departments and found that the

Minnesota Multiphasic Personality Inventory (MMPI) was the most used psychological test, being used by 43% of the departments. The next most popular tests were the California Personality Inventory, the Edwards Personal Preference Schedule and the 16 Personality Factor, in that order. None of these other tests were used by more than 16% of the departments.

Research Methods

Post-Selection Research

The most typical research design used in police-selection validation studies involves determining the correlation between selection variables and outcome variables. Three types of correlational research studies have been done. One type, which will be called pre-selection research, involves the collection of the selection data prior to the hiring of the subjects (e.g., the study of Marsh, 1962). A second type, which will be called post-selection predictive, uses data collected after the subjects were hired but before the collection of the outcome data (e.g., the study of Speilberger, Spalding, Jolley & Ward, 1979). The third type, which will be termed post-selection concurrent, uses selection data collected after the subjects were hired and at about the same time as the collection of the outcome data (e.g., the study Cascio, 1977). This third research design is the most common one in police-selection studies.

Both post-selection designs create major problems in interpreting results because of possible differences between actual applicants and

the officers who participate in the post-selection research. First, there is a likelihood of different levels of faking good. The motivation for presenting oneself favorably is far less in the case of individuals who have already been hired (Rothe, 1947). Several studies of various jobs found that persons in screening situations tend to produce substantially more socially desirable responses to psychological assessment (Elliot, 1976; Michalis & Eysenck, 1971; Kirchner, 1962; Bass, 1957; Heron, 1956; Herzberg, 1954; and Green, 1951), including biodata requests (Schrader & Osburn, 1977). Hence, variables that correlate with outcome measures in post-selection research may not do so with subjects who have more reason to fake good. This risk, of course, does not apply to selection variables that cannot be faked (e.g., measures of maximum performance; (Barrett, Phillips & Alexander, 1981), and it may apply only to a limited degree to variables that are rarely faked because of the chance of being caught (e.g., self-report of the highest educational degree earned.).

The second difference between pre-selection and post-selection studies involves the level of arousal. Officers already hired may be less aroused when completing a performance test such as an intelligence test. Differences in arousal can produce differences in performance, either increasing or decreasing it, depending on the individual and his optimum level of arousal (Hebb, 1949). Hence, performance tests that correlate with outcome variables in post-selection research may not do so when applied to applicants because with them arousal differences mask differences in ability -- which could be the crucial part of the

performance variable. On the other hand, performance tests found not to correlate with outcome variables in post-selection research might actually be useful in selection because they measure important differences in arousal control. They might only seem invalid because there is no arousal-control problem in post-selection subjects.

The final difference between pre-selection and post-selection studies involves differences in job experience. It is possible that personality types that are concurrently associated with good performance in veteran officers may be associated with subsequent poor performance in applicants. For example, officers who are good at their work may develop a personality trait of high self-appraisal. However, applicants who are high in self-appraisal may turn out to be officers who take needless risks.

Because of these difficulties in interpretation, post-selection research findings should be considered no more than suggestive with regard to the validity of their use in the selection of employees (Anastasi, 1976; Cascio, 1978; Dunnette, 1966; Guion, 1976; Siegal & Lane, 1974 and Zedneck & Blood, 1974). Pre-selection research is clearly preferable.

Inappropriate Comparison Groups

Several police-selection studies examined differences between officers who had resigned and officers who remained (Cross & Hammond, 1951; Finnigan, 1976; and Thweatt, 1972) in an attempt to establish what Anastasi (1976) called "validity through contrasted groups." Of these studies, three had serious confounds not mentioned in the research reports.

Finnigan (1976) attempted to determine whether a college education leads to better performance by police officers. The study compared a group of officers who had a college degree and a group of officers with only a high school diploma. However, in addition to differing as to education, the two groups differed in several other important respects. The college-educated officers were hired separately, were referred to in the police agency as "agents" and had to have a minimum of one year of field duty before qualifying to be hired. Further, although the report does not say so, it seems likely that the "agents" were paid more. Any of these other differences could account for differences between the two groups in outcome variables.

Thweatt (1972) compared a group of 50 experienced officers, sargeant and below, and a group of subjects who had resigned within 18 months of being hired. Apparently the groups were different regarding year when hired. This difference could account for differences between groups on selection variables.

Finally, Cross & Hammond (1951) compared current officers who had been employed "for at least one year" and officers who had resigned or been dismissed within the past three years. As in Thweatt's study, group differences in date when hired may have been a serious confound.

Outcome Variable Contamination

Outcome-variable contamination occurs when a predictor variable artifically influences an outcome measure. For example, one of the outcome variables in Finnigan's (1976) study was supervisor ratings. It seems likely that the knowledge by supervisors that one group was more

highly educated led by itself to some differences in ratings. This type of confound can occur in any pre-selection study that uses supervisor evaluations as outcome measures if the tested selection information is available to supervisors (McDonough & Monahan, 1975). The reports of most police-selection studies made no mention of this possible problem.

Unlawful and Unfair Selection Variables

Some police validity studies examined selection variables which could not lawfully or ethically be used. A prime example is race (e.g., Spencer & Nichols, 1971), which cannot lawfully be used in selection decisions by any police agency (42 U.S.C Sec 2000).

It may also be unlawful to use other biodata such as place of birth (e.g. Cohen & Chaiken, 1972) and length of residence (e.g. Levy, 1967) in the area of the police agency. Use of these biodata by police agencies would appear to violate the constitutional right to interstate travel See Shapiro v. Thompson (1969), which held that residency duration requirements for receiving free medical care violate this right

Another inappropriate variable is number of children, examined as a selection variable by Cohen & Chaiken, (1972). Use of this item by police agencies would appear to violate the constitutional rights of privacy and freedom of action with regard to procreation. See Griswold v. Connecticut (1965), in which the U.S. Supreme Court held that couples have a constitutional right of privacy which allows them alone to decide whether to use contraceptives.

Other questionable items are father's occupation, number of siblings, number of family members with a mental disorder (Cohen & Chaiken, 1972), birth order, number of parental residences (Leiren, 1973), (Levy, 1967; Speilberger, Spalding, Jolley, & Ward, 1979), and whether the applicant's father was home during the applicant's childhood (McDonough & Monahan, 1975). The use of some of these variables in selection would likely have the effect of screening out disproportionate numbers of minority applicants, who might tend to score low on these items. Moreover, the use of the items would be out of keeping with the American ideal of judging individuals by what they are or have done, rather than something that a parent did.

Meaningful and Non-Meaningful Outcome Variables

Conceptually inappropriate variables. Validation studies are necessarily no better than their outcome variables. It simply does not matter whether a selection variable correlates with a meaningless or completely ambiguous outcome variable.

Many different outcome variables have been used in police-selection research. Among those that are at least somewhat meaningful are the following: (a) involuntary termination (e.g. Levy, 1967); (b) citizen complaints (e.g., Cohen & Chaiken, 1972); (c) disciplinary accusations or actions (e.g., Cascio, 1977); (d) workers Compensation claims/times injured/days sick (e.g., McAlister, 1970; Snibbe, Azen, Montgomery & Marsh, 1973); (e) sick leave abuse (e.g., Spencer & Nichols, 1971); (f) preventable auto accidents (e.g., Fabricatore, Azen, Schoentigen & Snibbe, 1978); (g) resignation within two or three years (e.g., Azen,

Snibbe, Montgomery, Fabricatore, & Earle, 1974); (h) department awards/commendations (e.g., Baehr & Froemel, 1977); (i) letters of commendation from the public (e.g., Blum, 1964, study 4); (j) promotions (e.g., McDonough & Monaran, 1975); (k) supervisor ratings (e.g., Bartol, 1982); and (l) peer ratings (e.g., Henderson, 1979).

Several variables used for outcome measures are conceptually inappropriate. First, there is selection for hiring (e.g., Saxe & Reiser, 1976; Hogan & Kurtines, 1975; Spencer & Nichols, 1971). In some sense, selection for hiring can be viewed as a type of concurrent validity outcome measure. For instance, new depression scales are often evaluated against the MMPI depression scale. However, this type of validity check presupposes that the criterion measure has itself been shown to have predictive validity, that is, to be associated with performance measures. In the case of selection for hiring, there is no reason to believe that the selection is valid. Hence, selection for hiring is a meaningless outcome measure.

Second, there is training academy performance (e.g., Lester, 1979; Gordon & Kleinman, 1976; Hogan & Kurtines, 1975; Flynn & Peterson, 1972; Hogan, 1971; McAllister, 1970; Mormon, Hankey, Heywood & Liddle, 1966; Mormon, Hankey, Kennedy & Jones, 1966; Mills, McDevitt & Tomkin, 1966; Mullineaux, 1955; and Dubois & Watson, 1950). On the one hand, it is important that selected police candidates survive training in order to minimize training costs. On the other hand, training performance is not job performance, and to the extent that the two are unrelated, (a) it is unimportant whether a recruit finished first or last in his police

academy class, and (b) it is pointless to fail anyone, so prediction of termination during training is also unimportant. The question is, then, what is the relationship between training performance and job performance. Four relevant studies have been reported.

Cohen & Chaiken (1972) found that high academy grades were positively associated with many awards, promotion through civil service exam few complaints, few trials, few substantiated complaints against an officer, and few times sick. The study found no association with eight other outcome variables.

Gottlieb & Baker (1974) found an association between academy score and performance-based supervisor ratings.

Azen, Snibbe, Montgomery, Fabricatore & Earle (1974) found no significant association between academy variables of (a) academic performance, (b) physical performance, (c) marksmanship, and (d) instructor ratings and outcome variables of resignation within two years and supervisor ratings.

Leiren (1973) found no relationship between academy grades and later absenteeism, accidents, supervisor ratings, or commendations.

Hence, the evidence of an association between training performance and job performance is mixed. It may be that police departments vary widely in their training and evaluation of training performance. It therefore seems unwise to assume in any particular agency or study that police academy variables predict job performance. Thus training performance is generally not useful as a performance variable.

Third, there is the variable of no resignation ever (Levy, 1967 & 1973; Marsh, 1962; Snibbe et al. 1973; Blum, 1964, studies 3 & 4). It is reasonable for police agencies to want new police officers to stay with the department for at least a few years. The high cost of training makes rapid turnover through resignation a major problem. Therefore, an outcome measure of whether recruits stay at least two or three years is a meaningful one. However, it is doubtful whether longer tenure is meaningful. Burkhardt (1980) suggested that long tenure may be a negative outcome in that only the worst officers remain because (a) they cannot find better jobs, and (b) they fit into an existing police subculture that emphasizes alienation, cynicism, secrecy, isolation, and non-motivation. Although there is no direct empirical support for Burkhardt's first point, it has some intuitive appeal. Further, there is evidence that the subculture Burkhardt describes is a common one in police departments, and that it does tend to pressure recruits into becoming part of it or leaving the department (Van Maanen, 1975).

Even aside from Burkhardt's argument, it is unclear whether long tenure is a favorable outcome. Although training costs are saved and experience in the department may be helpful, some turnover may be good in that it helps to eliminate officers who are experiencing burn out and to add officers with new ideas and recent training.

The best view seems to be that turnover within 2 or 3 years is so costly that it is a valid outcome variable, but later turnover cannot be clearly interpreted as either good or bad. Hence, resignation after a few years is not a valuable outcome measure.

Fourth, there is the number of years as a police officer (Baehr & Froemel, 1977; Marsh, 1962). This variable has little to recommend it as an outcome measure. In addition to the problems mentioned above regarding resignation, it has virtually no meaning because it primarily measures the year in which officers were hired.

Fifth, there is the absolute number of positive or negative incidents or accomplishments with wide differences in opportunity (e.g., Spencer & Nichols, 1971). Outcome measures are meaningful only if opportunity is either equal for all subjects or is controlled for statistically. When the matter of opportunity is ignored and substantial differences exist, error variance occurs in the outcome variables, making it difficult to determine the true validity of selection variables.

In some cases the difference in opportunity can create a major confound. This may have occurred in the study of Baehr, Saunders, Froemel, & Furcon (1973), who found that black officers had twice as many arrests as white officers. One might speculate that blacks were better officers or that the black officers were assigned to minority neighborhoods with high crime rates. Assuming the latter, the researchers could have statistically controlled for the difference in opportunity by using as an outcome measure the number of arrests by an officer divided by the total number of arrests in the officer's precinct during the same time period.

Another likely victim of this type of error variance is any study using number of auto accidents as as outcome variable (e.g., Leiren,

1973). If all the subjects drove roughly an equal amount, there would be no problem. If, however, some drove much more than others, it makes little sense to use number of auto accidents as an outcome measure.

The question of opportunity is one that should be considered whenever an outcome variable is used that involves number of incidents or accomplishments. Researchers should not assume equality of opportunity even with regard to variables such as number of commendations. If as a practical matter commendations are given only to officers who go on patrol, there may be substantially less opportunity for officers who serve as administrators, guards or in other non-patrol positions.

In some cases it is very difficult or impossible to control for opportunity. For example, if the subjects frequently move from one precinct and job to another, determining opportunity can become a data nightmare. In these cases, researchers should at least point out the differences in opportunity so others can better interpret the findings.

Sixth, there is the number of accusations of wrongdoing by officers determined to be unfounded (Cascio & Real, 1977; Spencer & Nichols, 1971). This variable makes no sense as an outcome variable. Unfounded accusations are neither good nor bad indicators of performance.

Lack of coverage by outcome variables. In addition to the problem of inappropriate outcome variables, there is a problem of too few outcome measures relating to performance other than arrest-type actions. Job analyses done on the work of police officers have shown that they tend to spend about 85%-90% of their time doing work unrelated to crime,

including mediating family disputes and writing accident reports (Lefkowitz, 1977). None of the studies surveyed included any direct measure of these variables. Although supervisor ratings could be an indirect measure, it may be doubted whether supervisors either have informaton on the execution of these tasks or particularly care. Police supervisors may tend to pay more attention to crime-fighting aspects of performance (Burkhardt, 1980).

Some steps have been made in the direction of assessing performance unrelated to crime-fighting. Carr, Larson, Schnelle & Kirchner (1980) reported efforts to assess different types of police performance such as report writing and testifying. They used carefully constructed rating forms. As yet, there appears to be no reliability or validity data on approaches of this sort.

Low-Variance Problems With Outcome Measures. Rosen & Meehl (1955) noted the statistically based difficulty of predicting infrequent events. The same problem applies to outcome variables with low-variance.

Outcome variables with inadequate variance cannot always be identified in research reports because their variance foten is not reported. A likely candidate for inadequate variance in many studies is departmental ratings, because they tend to have such a leniency (ceiling) effect that there is little or no difference among subjects. Variables which surely had inadequate variance include whether an officer has been charged with a crime while employed (Cohen & Chaiken,

1972 -- 8% were charged) and being fired within a few years of being hired (McAllister, 1970 -- 5 were fired compared to 60 retained).

Researchers would be well advised to determine whether selection and outcome variables being considered for a study have sufficient variance to warrant the time and effort needed to collect and analyze the necessary data. Unfortunately, low-variance cannot always be predicted in advance. When it effectively scuttles an outcome variable, research reports should note this so that the selection variables tested will not be mistakenly written off by others.

Rating Problems. Ratings are the only type of outcome measure that have received significant attention with regard to their usefulness. Traditionally, the only type of rating used as an outcome measure was an absolute supervisor rating on a characteristic or job behavior. This is still a common outcome measure in police-selection studies. Several problems exist with these ratings. First, there are the leniency and low-variance problems mentioned above. Second, there is doubt about their meaningfulness. Third, ratings on different dimensions are very highly correlated, suggesting a hallo effect that makes use of the dimension-scores dubious.

Recent police studies by Cascio & Vanlezi (1978) and Baehr & Froemel (1977) found few or no associations between supervisor ratings and objective measures of performance, suggeting that the ratings may not be meaningful performance measures. Similar findings have been reported with non-police employes (e.g., Hausman & Strupp, 1955; and Seashore, Indik & Georgeopoulos, 1960).

Many attempts have been made either in police research or other selection research to overcome these problems. These attempts involve using different persons doing the ratings and different kinds of rating formats. One demonstrations study (Carlson & Sutton, 1979) explored the possibility of using citizen ratings of police officers, but this approach appears to be plagued with methodological problems (e.g., standardization of ratings) and cost problems.

Two police-selection studies (Bass et al. 1954; Henderson, 1979) used peer rankings as a alternative to, or in addition to, supervisor ratings. Two other studies (King, Hunter & Schmidt, 1980; Love, 1981) examined the validity of peer rankings or ratings in police-performance evaluations. King, et al. found that peer rankings were correlated with supervisor ratings and rankings. Love (1981) found that peer rankings and ratings were both correlated with supervisor rankings and ratings. However, another report of the study (Love, 1983) indicated that high peer rankings were also correlated with more on-the-job injuries. Love (1981) also found that the officers tended to dislike making the ratings.

Non-police studies on the validity of peer ratings as an outcome measure have produced mixed results (Gruenfield, 1981).

Two developments in rating formats, behaviorally anchored rating scales, (BARS) and paired comparison, have recently found their way into police-selection studies.

BARS are scales developed so that each as a series of ranked behavioral (performance) statements "anchoring" points along scale. The

process of creating these statements, as first described by Smith & Kendall (1963), is time-consuming and complicated. It was originally hoped that the scales would reduce hallo effects by making points more meaningful.

As yet, BARS have not clearly been shown in non-police studies to have any advantages that would justify the time and effort needed to create them (Gruenfield, 1981). The findings of two studies using police officers suggest that BARS do not improve outcome measurement. Cascio & Valenzi (1978) compared eight carefully created BARS with each other and 15 objective measures of police performance. The study found that the BARS had intercorrelations ranging from .84 to .91. This suggested that the ratings did not discriminate among different aspects of performance. The study also found that none of the BARS was significantly associated with the best linear composite of the objective measures, as determined through multiple regression. This suggests that the BARS lacked of validity.

Landy, Farr, Saal & Freytas (1976) collected BARS data in 58-police agencies. The overall results were that median intercorrelations among the BARS were about the same as interrater reliability. Meaning that scale ratings for an individual could be predicted nearly as well from a rating on another scale as from another rater's rating on the first scale. Hence, for an individual subject, any difference between scale scores was almost meaningless.

Rankings have been used instead of ratings in police-selection studies (e.g., Azen et al., 1974) in an effort to increase variance in

subject scores and, thereby, increase validity. A police study by Love (1981) provided some evidence that rankings are more valid.

Three type of rankings have been examined with police: paired comparisons, whole-group rankings and nominations. Obtaining paired-comparison ratings involves requiring a rater, usually a supervisor, to compare all possible pairs of subjects on the variable of interest. A total score is then derived for each subject, such as by adding the number of times he or she was named as superior (Fabricatore et al. 1978).

Whole group rankings have also been used (Mormon, Hankey, Liddle & Goldwhite, 1967) for this purpose, but not recently. This lack of use may be the result of concern about the appropriateness of using ordinal data in multiple regression analyses, which have become commonplace. Although some controversy remains, there appears to be growing acceptance of quantitative analyses of this sort with ranked data (Roscoe, 1975).

Nominations are similar to whole-group rankings except that each rater names only the few top performers in order of quality of performance. That saves raters from having to give low rankings to anyone. Total scores are based on all the ratings for each individual.

Gruenfield (1981) suggested that many of the problems associated with ratings occur because the raters lack (a) training in making ratings, (b) information on the subjects, and (c) incentive for making valid ratings. These problems generally have been ignored in police-selection research.

Alpha inflation

The most conspicous statistics problem in police-selection research involves ignored alpha inflation. Alpha inflation means that as either selection or outcome variables exceed one, the likelihood of any single correlation being significant at $p < .05$ or .01 because of chance increases. In other words, type 1 error becomes more likely. For example, Speilberger, Spaulding, Jolley & Ward (1979) apparently examined 147 scales of the Strong-Campbell Interest Inventory and found for males that four scales were associated with resignation without being eligible for rehiring. One would expect about that number of "significant" correlations by chance.

A related problem has occurred with the use of multiple regression. For example, McDonough & Monahan (1975) reported a multiple regression correlation of .50 between predictors and a performance rating, with 92 subjects. The report of the study expressed surprise that the correlation would be this high when the criterion rating had an interrater reliability of only .41. The ratio of 54 predictors to 92 subjects suggests that the multiple R was primarily the result of alpha inflation. Multiple regression significance tests control for alpha inflation, but none was reported in the study. A more subtle type of multiple-regression alpha inflation occurred in the study of Spencer & Nichols (1971) which examined dozens of selection variables. Only the top few were entered into a multiple regression analysis, thus allowing calculation of a misleading significance level (see Nunnally, 1978).

There are relatively simple solutions to these problems. First, statistically control for alpha inflation whenever multiple selection or outcome measures are used and report the overall significance levels. Second, if in an unusual case alpha inflation cannot be statistically controlled for, a statement of the problem would suffice. Third, whenever possible, cross validation or replication should be attempted to confirm significance.

Instability Of Beta Weights

In multiple regression police research, much is often made of a prediction equation found to predict outcomes relatively well. The exact equation may be used in other research (e.g., Mills & Bohannon, 1980) or conceivably in making hiring decisions. If the equation is cross-validated, this may be reasonable. Otherwise, beta weights should be interpreted and used with caution, as they can be quite unstable, especially if the predictors are highly intercorrelated (Kerlinger & Pedhauzer, 1973, p. 77).

Dawes (1979) convincingly summarized evidence that improper linear models (with randomly selected weights) succeeded as well in cross validation as did proper beta weights. The article recommended the simplest approach: ignore beta weights and assume each selection variable has the same weight. Cattin (1978) pointed out, however, that proper beta weights are superior when (a) the ratio of predictors/N is small, (b) there are substantial differences among the weights, (c) the R squared of the regression equation is relatively high, and (d) predictors are not highly intercorrelated.

Moderator Variables

There is controversy over whether moderators are useful in police-selection research. Moderators are variables which are not predictors but which increase the predictive ability of predictors. For example, Baehr et al. (1973) found that when data for black and white officers were analyzed separately the correlation between predictors and outcome variables was substantially higher. Speilberger, Spaulding, Jolley, & Ward (1979) found substantial differences in which selection variables were useful when males and females were separated.

Schmidt and Hunter (1978) argued, however, that variables such as race and sex have no intuitive appeal as moderators. On a rational basis, one might ask why would a selection variable (e.g., intelligence) predict performance only with whites or only with males. On a statistical basis, one must ask whether group differences in validity are meaningful. Schmidt, Berner & Hunter (1973) examined all single group police and non-police studies on the role of race in selector validity, and concluded that not one of the studies found a difference between groups large enough to be statistically significant. They noted that large differences usually are the result of chance variation especially likely to occur when one of the groups has a relatively low number of subjects. Hunter, Schmidt and Hunter (1979) later used a meta-analysis to examine dozens of race-moderator studies and found no overall significant difference in validity. Hence, there appears to be little or no statistically sound evidence to support a role for moderators in selection research with police officers or other workers (Dunnette & Borman, 1979; Ghiselli, 1966).

A final note on this subject has to do with the legal requirements for the use of moderators. Some of the impetus for the use of race as a moderator has come from misinterpretations of federal law. Baehr (1979) and Cascio (1977) stated incorrectly that regulations issued by the Equal Employment Opportunity Commission under Title VII of the Civil Rights Act require that selection measures be validated for both whites and minorities. In fact, the regulations require such validation onl if the measure is used so as to discriminate against minorities and only if technically feasible. One would think that it would be a rare case in which both qualifications are met. The use of formal and informal quotas prevents discrimination, and technical feasibility of dual validation rarely exists because even with qutoas few numbers of minority groups are hired.

Findings of Police Selection Studies

The review of findings will be presented according to type of selection variable. The types included are biodata, measures of intellect, measures of personality and psychopathology, interviews, interest inventories, and subjective background ratings.

Situational tests, although used in some police departments (Mills, McDevitt & Tonkin, 1966), are omitted because no research relevant to their validity has apparently been reported. Polygraphs likewise have not been empirically evaluated as employee selection devices. For different views of the ability of polygraphs to detect lies, see Sackett and Decker (1979), Lykken (1979), and Podlesney and Raskin (1977).

Some published studies are excluded from the review of findings because they had either inadequate research designs (Cross & Hammond, 1951; Finnigan, 1976; and Thweatt, 1972) or no meaningful performance outcome variable (Flynn & Peterson, 1972; Hogan, 1971; Milles et al. 1966; Mormon, Hankey, Heywood & Liddle, 1967; Mormon, Hankey, Kennedy & Jones, 1966; and and Mullineaux, 1955). Also exluded are published reports of studies that are so lacking in statistical or other important information as to be uninterpretable (Cascio & Real, 1979; Colarelli & Siegel, 1964; Furcon, Froemel & Baehr, 1973; Levy, 1973; Mills & Stratton, 1982; and Spencer & Nichols, 1971).

It is worth noting that many of the research reports examined for this review failed to provide even the most rudimentary listing of selection variables tested (e.g. Spielberger, Spaulding, Jobley, & Ward, 1979). Too common are the presentation of nonsignificant correlation coefficients without significance levels (e.g., Henderson, 1979) and the presentation of multiple regression statistics without 0-order correlation coefficients (e.g., Fabricatore, Azen, Shoentgen & Snibbe, 1978).

All reported associations were significant at $p < .05$ or better unless otherwise indicated.

Biographical Items (Biodata)

For the purposes of this review, biodata are items concerning an individual's life experiences. Although some researchers have referred to data such as written responses to questions regarding motivation as being biodata (e.g., Levy, 1973), these are more properly characterized as personality or psychological-state measures.

Because of the nature of biodata, most the studies used pre-selection designs. The studies that used biodata collected after the subjects were hired will be noted.

In considering the findings, one should keep in mind that many of the items may have been found to be uncorrelated with outcomes in studies which did not report nonsignificant biodata (e.g., Levy, 1967; and Speilberger, Spaulding, Jolley, & Ward, 1979). A number of specific items listed above under Unlawful Selection Variables will be ignored because their use in selection would or might be unlawful.

Age. Bartol (1982) used a group of officers divided into those rated by supervisors as average, below average and above average to assess the relationship between several predictors, including age as a continuous variable, and supervisor ratings. The study found that the below-average officers were older when hired than the other two groups. There was no significant difference between average and above average.

Five studies found no evidence that age at hiring was a valid predictor of police performance. Levy (1967) assessed age divided into ranges of under 24, 24-26, 27-29 and 30-above and found no association with involuntary termination. Azen et al. (1974) assessed age as a continuous variable and found no correlation with voluntary termination within two years. McDonough & Monahan (1975) found no association between age as a continuous variable and resignation, being fired, or being promoted within two years. Gottlieb & Baker (1974) found no association between age as a continuous variable and supervisor ratings. Marsh (1962) assessed the validity of age with unspecified age

categories and found no association with supervisor ratings, being discharged or auto accidents.

Snibbe et al. (1973) used 95 of Marsh's 550 subjects 10 years later in a follow-up study and found that younger applicants were more likely to be promoted beyond patrol officers. There was no association between age and supervisor ratings, auto accidents, workers' compensation claims or injuries.

One study found evidence that officers who were older when hired performed better. Cohen & Chaiken (1972) used 14 outcome measures to examine age in terms of three categories: 18-24, 25-29, and 30-above. The study found that being older was associated with few total complaints, few civil complaints and little absenteeism, but also with no career advancement.

Overall, the findings regarding age are wildly mixed. Older applicants may be better, worse or the same as other applicants. The only replicated finding is that police officers who were older when hired were less likely to be promoted, even when they otherwise performed as well as or better than the other officers. This merely raises doubts about the meaningfulness of promotion as an outcome variable.

Court Appearances. Cohen & Chaiken (1972) found an association between appearances in civil court prior to being hired and harassment. Considering that 14 outcome measures were used in the study, and alpha inflation was not controlled for, one should put little confidence in this finding. The study also found that receiving a summons to testify

in a civil proceeding before being hired was associated with later departmental awards. This predictor was also associated with being tried for misconduct and substantiated cases of misconduct. However, the relationship with these two outcomes was such that officers who had received one summons to testify had far fewer of the two negative outcomes than did officers with no summons or more than one summons. This sort of relationship is difficult to interpret as being other than a chance finding.

Criminal convictions and arrests. Levy (1967) found that vehicle code violations and more serious offenses prior to being hired were associated with involuntary termination. These findings were cross-validated in the study. Cohen & Chaiken (1972) found no association between petty offenses and any of 14 performance measures, but did find that arrests before becoming an officer were associated with fewer harassment charges while an office. Because this finding was in the unexpected direction and 14 outcome measures were used, one should be cautious in placing any confidence in it.

Debts. Cohen & Chaiken (1972) and Levy (1967) examined the validity of number of debts and both studies found no associations with performance.

Education. The most commonly studied biodatum in police-selection studies is amount of education. Because many police departments require at least a high school degree (Abramson, 1974), education usually has been studied by evaluating the association between performance and years of education past high school.

Two studies found evidence that more education is associated with better performance. Cascio (1977) studied the predictive ability of education with a group of deputies divided into 825 whites, 60 blacks, and 55 "Spanish-surnamed." Inexplicably, two orientals were treated as "Spanish-surnamed." Forty-four outcome measures were used, including BARS which assessed job knowledge, initiative, dependency, attitude, relations with others and communications. The other measures were objective measures, of which the report listed only the ones that correlated significantly with education.

For the white officers, 13 meaningful objective measures were associated with education. More education was associated with a lower incidence of 12 negative outcomes: injuries, injuries by assault and battery, disciplinary actions due to accidents, verbal discourtesy allegations, preventable accidents, use of force, internal reviews, legal investigations, "personnel" complaints, false-arrest allegations, physical force allegations, and times sick per year. However, more education was also associated with fewer commendations.

For the black officers, more education was associated with fewer injuries, fewer preventable accidents, fewer injuries, fewer physical force allegations and more commendations.

With the "Spanish-surnamed" subjects, there were two significant correlations, one favoring more education (association with fewer preventable accidents) and the other cutting against it (association with more false-arrest allegations).

Hence, there was considerable evidence with regard to white officers and black officers that more education led to better performance. This was not true with regard to "Spanish-surnamed" officers. As was discussed above under moderator variables, this is unlikely to be the result of actual differences in validity.

Cohen & Chaiken (1972) examined education with 14 outcome variables and found that two of the outcome measures were associated with more education: career advancement by civil service exam and lack of civil complaints against the officer.

Two studies found that more education was associated with poor performance. Levy (1967) using educational levels ranging from less than 11 years to more than 17, found that of the officers hired in the past 10 years, those who were retained had substantially less education than those who were fired. The study cross-validated the finding. Gottlieb & Baker (1974) found that college education was associated with low supervisor ratings.

Overall, the results are wildly mixed as to the validity of education as a predictor of police performance. This conclusion is in line with that of Caplan & Schmidt (1977), who reviewed the research on education as a predictor of performance in jobs generally and found little evidence of validity. The contradictory research findings also fit well with conflictng conceptual arguments (see Lefkowitz, 1977 & Sparling, 1975) that more education would be positively associated with performance (e.g., through added understanding of human behavior), negatively associated (college graduates may find routine work unchallenging) or unrelated.

Experience as a police officer. Leiren (1973) found an association between prior experience as a police officer and high supervisor ratings, but no relationship with absenteeism, commendations from the public, and auto accidents. Considering alpha inflation, one can put little confidence in the one "significant" finding.

Caplan & Schmidt (1977), in reviewing employee selection in general, found little evidence of validity for prior experience.

Involuntary terminations. Levy (1967) found that having ever been fired in a prior job was positively associated ($p < .01$) with involuntary termination as a police officer. The study cross-validated the finding.

Marital status. Levy (1967) found that officers who had been married at least twice were more likly to be fired. This finding was cross-validated in the study. Azen et al. (1973) found no effect for number of marriages with regard to resignation within two years, absenteeism, supervisor ratings in absolute and paired-comparison formats.

Four studies found that the variable of being married or unmarried at the time of hiring was unrelated to performance as a police officer (Cohen & Chaiken, 1974; Gottlieb & Baker, 1974, Levy, 1967; and Speilberger, Spaulding, Jolley & Ward, 1979).

Military disciplinary actions. Cohen & Chaiken (1972) found that military disciplinary actions were positively associated with complaints against an officer, departmental trials and substantiated complaints but not with 11 other performance variables.

Military Experience. Azen et al. (1974) found that prior military experience was associated with continuation with the department for at least two years. However, Cohen & Chaiken (1972) found no association with any of 14 performance variables, Levy (1967) found no association with involuntary termination, and Gottlieb & Baker (1974) found no association between prior military experience and supervisor ratings. McDonough & Monahan (1975) found no association between "type of military service" and resignation, being fired or being promoted within two years.

Participation in athletics. Speilberger, Spaulding, Jolley, & Ward (1979) found that for white males, lack of participation in athletics was associated with termination without being rehirable. For females, there was no association. Shealy (1976), however, found a positive association between prior participation in athletics and supervisor evaluations that an officer was corrupt. Hence, as with regard to education and age, the results have been inconsistent, making a conclusion on the value of the predictor impossible at this point.

Prior jobs. Levy (1967) found that long duration of prior jobs was associated with staying on the force as opposed to being fired. This finding was cross-validated in the study. Cohen & Chaiken (1972) found no association between long duration of prior jobs and any of 14 outcome measures.

Cohen & Chaiken (1972) and Levy (1967) also investigated the predictive validity of status of prior occupations and number of prior jobs. Neither study found any association with outcome measures.

Training in police science. Levy (1967) found no association between having taken police science courses and involuntary termination.

Psychological disorders. Cohen & Chaiken (1972) found no association between reports of prior psychological disorders and any of 14 outcome variables.

Summary of biodata. Overall, the studies provided encouraging empirical evidence that five biodata are associated with poor performance by police officers. The variable of prior involuntary terminations was cross-validated as a predictor of involuntary termination as a police officer (Levy, 1967), and no contrary findings have been reported. Criminal convictions and vehicle-code convictions, having been married more than once, and short duration of prior jobs were all cross-validated as predictors of involuntary termination (Levy, 1967), but subsequent studies of each failed to replicate the findings.

It seems likely that a combination of these five biodata would be more successful than any one item in predicting police-officers performance. This conclusion is consistent with the conclusion of reviews of non-police validity studies that groups of items can validly predict job performance (Dunnette & Borman, 1979; Owens, 1976; and Tenopyr & Oeltjen, 1982).

Measures of Intellect

Pre-selection validity. In one of the seven reported pre-selection studies of published intelligence tests, McDonough & Monahan (1975) assessed the validity of the Otis Intelligence Test and an unpublished

civil service exam by comparing groups of fired, promoted and non-promoted officers. There were two significant group differences. Both tests discriminated between fired and promoted and between non-promoted and promoted. However, supervisors were aware of the scores, so outcome-variable contamination may have contributed to the results. Also, the report did not say whether promotion was based on the results of promotion tests, on which more intelligent officers would likely do better.

Blum (1964, study 3) assessed the pre-selection validity of several measures of intellect, including the Otis, Army Alpha form SB, Army Alpha (Intelligence Test), Terman-Merrill Intelligence Test, Moss Social Intelligence Test, Moss Mental Alertness Test and O'Rourke's Policeman Aptitude Test. The study found no association between any of the tests and being fired or promoted.

Marsh (1962) assessed the pre-selection validity of a civil service exam and its parts: word memory, sentence completion, number series completion, arithmetic reasoning, cube and block counting, practical judgment and memory. The study used as outcome measures departmental rating and number of auto accidents. The results were that high scores on the overall test on sentence completion and on number series completion were positively associated with high departmental ratings, and that no correlations were found between test scores and auto accidents.

Snibbe et al. (1973), using 95 of Marsh's subjects 10 years later, assessed the pre-selection validity of the total test score. The study

found a positive association between high test scores and high rank but no association with whether still on patrol, supervisor ratings, auto accidents, workers' compensation claims or injuries.

Dubois & Watson (1950) assessed the pre-selection validity of the Army General Classification Test (First Civilian Ed.) and found no correlation with departmental rating. Gottlieb & Baker (1974) investigated the pre-selection validity of the Schrammel General Ability Test and found no association with supervisor ratings.

Post-selection predictive validity. Cohen and Chaiken (1972) studied the post-selection predictive validity of the Otis Intelligence Test and found that high scores predicted promotion by written test but were unassociated with 13 other performance variables.

Speilberger et al. (1979) assessed the post-selection predictive validity of the Nelson-Denny Reading Test with groups of females and white males. Although the test is not a measure of intellect, it is a test of maximum performance like intelligence tests, and the study, therefore, will be discussed here. Termination during the probationary period without being eligible for rehiring was the sole criterion. For white males, there was an association between termination and low scores on the total test, on comprehension, on vocabulary and on reading rate. For females, there was an association between termination and low comprehension.

Blum (1964, study 4) assessed the post-selection predictive validity of the Otis with several outcome measures: days lost due to illness, periods in which sick leave was taken, minor disciplinary

charges, injuries, accidents, serious charges, commendations, and citizens' commendation. The study found that high test scores were associated with more departmental commendations, but also with more auto accidents and injuries.

Blum (1964, study 2) investigated the post-selection predictive validity of the AGCT and the McCardell Test of Practical Judgment and found no association between scores on either test and departmental ratings on character and performance.

Concurrent validity. Henderson (1979) investigated the concurrent validity of three measures of intellect in a two-part study. In one part, 151 officers took the Culture Fair Test and the numerical and verbal subtests of the Differential Aptitude Test as volunteers. In the other part of the study, 234 officers took the Culture Fair and the numerical and verbal subtests of the SRA as part of seeking promotions. For both parts of the study the outcome variables were ratings by superiors and by peers. The results were reported without significance levels. We have determined the significance of the correlation coefficients through a table provided by Snedecor & Cochran (1967) and found that in each study high scores on the numerical test were associated with high ratings by superiors and peers, but that scores on the Culture Fair verbal tests were not associated with superior and peer ratings.

Bass, Karstendiek, McCollough & Pruitt (1954) assessed the concurrent validity of two Moss tests (Social Intelligence and Memory), the Wonderlik Personnel Test, and two perceptual flexibility tests by

Thurstone (Gestalt Completion and Hidden Figures). The data for two groups of deputies from different agencies were analyzed separately. The outcome measure was a rating by the chief for one department and peer ratings in the other department. The report provided zero-order correlations without significance levels. Conversion of these into significance levels with a conversion table (Snedecor & Cochran, 1967) shows that none of the test scores were associated with performance rating.

Leiren (1973) evaluated the concurrent validity of a civil service exam and found no association with letters of commendation, number of auto accidents, supervisor ratings or absenteeism. Cascio (1977) investigated the concurrent validity of the California Mental Capacity Questionnaire, using 15 objective performance measures. The multiple correlation was nonsignificant.

Experimental tests. Leiren (1973) investigated the concurrent validity of seven experimental measures of intellectual functioning with four outcome measures. The findings were that high verbal-reasoning scores were associated with commendations, few accidents, and high supervisor ratings, that high numerical-reasoning scores were associated with supervisor ratings, and that high vocabulary scores were associated with few accidents. The other 23 correlations were nonsignificant, including all involving the following tests: spacial visualization, "best trend name," and "letter triangle." Absenteeism was not associated with scores on any test.

Crosby, Rosenfield & Thornton (1979) developed a test intended to assess several aspects of intelligence and tested the concurrent validity of the test in four police departments. In each department, the test was assessed against supervisor ratings on 15 dimensions. The only results reported were in terms of number of significant correlations between high test scores and high ratings, listed according to department: (a) all 15 correlations significant; (b) 10 significant; (c) none signifitant; and (d) 3 significant in the unexpected direction.

Using 30 officers, Martin (1923) assessed the concurrent validity of eight experimental measures of intellect, a civil service exam and two parts of the Army Alpha Intelligence test (opposites and common sense). The only finding reported is that a multiple regression equation based on seven of the tests produced a R-squared coefficient of .55. No significance level was reported. Using a formula for testing R-squared (Kerlinger & Pedhazur, 1973, formula 3.12), one finds that the R-squared was not significant.

Summary of measures of intellect. It might well be thought that the key studies were the four that used a pre-selection design. Only McDonough and Monahan (1975) found that high test scores predicted a good outcome and that was promotion. One must suspect high test scores led to promotion in the study of McDonough and Monahan because (a) promotion was based on the scores, or (b) promotion was based on later tests, and officers who did well on the pre-employment tests also did well on the promotion test. Hence, the ability of an intelligence test to predict promotion in the study should not be interpreted as validating the test.

The post-selection studies of published tests also provided little evidence of validity for measures of intellect. Only eight of 52 correlations calculated for various outcome measure showed a positive, significant association between high test scores and good performance. Further, there were two correlations showed an association between high test scores and poor performance. Thus, one is drawn to the tentative conclusion that measures of intellect are not valid predictors of police performance. This conlusion stands in contrast to that of Ghiselli (1966, 1973), who reviewed a large number of studies done on the validity of measures of intellect as predictors of performance in a wide array of jobs. The reviews concluded that there is empirical support for the general validity of the measures. However, it is possible that the job of a police officer is different from other jobs in ways that make high intelligence test scores unimportant, or both good and bad.

In the future, researchers could produce optimally useful findings by evaluating only tests that have been published, with demonstrated reliability and validity as a measure of intellect or a related construct.

MMPI. The MMPI is the only measure of personality and psychopathology that has been evaluated in more than one pre-selection study, so the MMPI will be reviewed before similar measures.

The most recently reported MMPI study was done by Bartol (1982), who assessed pre-selection validity with ratings by the chief of police as to whether officers were average, above-average or below-average. Of the 13 MMPI scales, only one discriminated among groups. Subjects with

high Mf scores were more often rated as below average than average or above average.

The study also compared the groups without the usual K-corrections on Hs, Pd, Pt, Sc and Ma. Without the corrections all five of the scales discriminated between average and below and between above and below average. None discriminated between average and above average.

Bernstein, Schoenfield & Costello (1982) evaluated the pre-selection validity of the 13 validity and clinical MMPI scales and also the ? scale. The report provided only statistics showing the multiple regression correlation between all the MMPI scales and each outcome measure. Therew were significant multiple correlations for the 14 scales in predicting four performance variables: disciplinary days, citizen complaints, sick days and injuries. There was no significant multiple correlation with disciplinary actions, grounded citizen complaints, chargeable auto accidents, or avoidable auto accidents.

Saxe & Reiser (1976) assessed the pre-selection validity of the MMPI and found that high scores on Hy and Pt were associated with voluntary termination within three years. However, low scores on L, K and Pa were also associated with termination.

McDonough & Monahan (1975) assessed the pre-selection validity of the MMPI scales and found no difference between groups of fired, non-promoted, resigned and promoted officers.

Azen et al. (1974) assessed the predictive post-selection validity of the MMPI with three outcome measures: resignation during training or within the first two years thereafter, an absolute rating by supervisors

and a paired-comparison rating by supervisors. Only high Mf scores were associated with resignation. The officers were divided into two groups for determining the correlaton between scale scores and ratings. In one group, high Pa and Ma were associated with low absolute ratings by supervisors but not with low paired-comparison ratings. There were no significant correlations for th other group of subjects.

Blum (1964, study 4) assessed the predictive post-selection validity of the MMPI with 10 outcome measures: departmental commendations, public commendations, charges of misconduct, substantiated charges of serious misconduct, charges of minor misconduct, auto accidents, injuries, number of periods when sick leave taken, total days lost due to illness and ratings of assignment progression. Several associations were found; all were between high scale scores and negative outcomes. No significance levels were provided in the report, but these can be determined through a table for translating correlation coefficients into significance levels (Snedecor & Cochran, 1967). The significant associations were F with charges of minor misconduct and substantiated charges of serious misconduct; Mf with periods in which sick leave was taken; Pa with substantiated charges; Pt with charges of minor misconduct and substantiated charges; Sc with charges of minor misconduct, serious charges and substantiated charges; and Ma with charges of minor misconduct and substantiated charges of serious misconduct.

Schoenfield, Kobos & Phinney (1980) assessed the predictive post-selection validity of the MMPI by comparing scale scores of 23

officers whom supervisors would not rehire and 46 officers who were rated acceptable, had a commendation and had no disciplinary actions against them. There were no significant differences between groups on any scale. The study also assessed the ability of two psychologists to distinguish between the two groups by clinical interpretations of MMPI scores. Neither psychologist identified members of the two groups with accuracy greater than chance.

Costello, Schoenfield & Kobos (1982) used the same two groups and a randomly selected intermediate group of 92 officers to assess the post-selection predictive validity of the Goldberg MMPI Index (L + Pa + Sc + Hy + Pt). The results were that the index discriminated between the two extreme groups but not between either of those and the intermediate group.

Merian, Stefan, Schoenfield & Kobos (1980) used the data for the same 92 officers to assess the predictive post-selection validity of the 566 MMPI items. The results were that five items were cross-validated at $p < .10$ as discriminating between the low-rated and exemplary officers. On the basis of chance, one might expect about 50 items to be significant in one comparison and then 5 of those to be cross-validated. Hence, the finding suggested that MMPI items were not valid predictors of performance.

Speilberger, Spaulding, Jolley & Ward (1979) assessed the predictive post-selection validity of the MMPI L scale and an experimental sociopathy scale made up of MMPI items. The study found no association with termination without being eligible for rehiring.

Blum (1964, study 2) studied the predictive post-selection validity of clinical judgment of MMPI profiles. The outcome measure was derived by dividing subjects into poor, medium and good performance groups based on departmental ratings, commendations and disciplinary actions. The results were the opposite of those expected. Predictions of failure were associated with better performance.

Overall, not a single scale was found to be valid in more than one pre-selection study. Mf was found to be a valid selection variable in three studies, two of which were post-selection studies. High scores of Mf were associated with low supervisor ratings (Bartol, 1982), resignation (Azen et al., 1974) and number of periods in which sick leave was taken (Blum, 1964, Study 4). One can only speculate whether high Mf officers (artistic, creative, effeminate) performed poorly or were just out of place in a tough-guy police subculture. At any rate, the scale deserves further validation research. Scales Pd, Pt and Sc, are also deserving of further evaluation, as they showed validity in more than one study.

It is unclear what to make of the finding of Bartol (1982) that Hs, Pd, Pt, Sc and Ma were valid predictors only if not K-corrected. It could be that K-corrections tend to mask valid predictor information in the scale scores. Researchers might profitably explore this avenue.

Four scales were shown repeatedly to lack usefulness: L, K-corrected Hs, D and So. These scales would seem undeserving of further validation attempts.

Personality and Psychopathology Measures Other Than the MMPI

California Personality Inventory (CPI). McDonough & Monahan (1975) assessed the pre-selection validity of the 18 CPI scales by comparing groups of officers who, two years after being hired, had been fired, had resigned, had been promoted, or had not been promoted. A total of 80 group comparisons were made. Of these, only two were significant at $p < .05$. High Responsibility and Socialization scores were associated with being fired as opposed to being still employed. These findings were in the unexpected direction, but in view of the number of correlations done, one would expect two to be significant at $p < .05$ by chance.

Speilberger, Spaulding, Jolley, and Ward (1979) assessed the predictive post-selection validity of the CPI. The study found that low scores on the following scales correlated with termination without being eligible for rehiring for white males: Dominance, Capacity for Status, Sociability, Achievement through Conformance and Intellectual Efficiency; for females: Capacity for Status, Well-Being, Responsibility, Self-Control, Tolerance and Good Impression.

Mills and Bohannon (1980) used supervisor ratings of leadership and suitability for police work to assess the concurrent validity of the CPI. The study found that Tolerance, Intellectual Efficiency and Achievement through Independence were positively associated with both outcome measures. Socialization, Communality and Flexibility were associated with suitability only. The other scales and Gough's (1969) leadership index (.372 Dominance + .696 Social Presence + .345 Well-Being + .274 Achievement via Independence - .133 Good Impression) were found not to be associated with either outcome variable.

Hogan (1971) assessed the concurrent validity of the 18 CPI scales and found that high scores on six scales were associated with good supervisor ratings: Well-being, Responsibility, Self-Control, Good Impression, Achievement via Conformance, Achievement via Independence, and Psychological Mindedness.

Hogan & Kurtines (1975) assessed the concurrent validity of the CPI with a performance variable of number of disciplinary actions. The report provided no 0-order correlations regarding individual scales but did state findings regarding a social maturity index (Gough, 1966; .148 Dominance + .334 Responsibility + .512 Socialization + .227 Flexibility - .317 Good Impression - .274 Communality) and a leadership index (Gough, 1969). The study found no association between either scale and number of disciplinary actions.

Although several studies of the CPI have been reported, only one used a pre-selection design, and it produced inconsistent findings. Several scales, including Achievement via Conformance, Self-Control, Tolerance and Intellectual Efficiency, were found to have concurrent validity in two studies, making them deserving of evaluation by pre-selection studies.

Edwards Personal Preference Schedule (EPPS). Azen et al. (1973) investigated the concurrent validity of the EPPS scales and found that low scores on one of the 15 scales, Introception, were significantly associated with termination during training or within two years thereafter.

Leiren (1973) also assessed the concurrent validity of the EPPS scales. Four outcome variables were used: Supervisor rating, absenteeism, letters of commendation (source unspecified) and number of auto accidents. Only one of the 60 correlations was significant. High scores on scale F I (aggressivley self-assured) were associated with high supervisor ratings.

Henderson (1979) assessed the concurrent validity of the EPPS scales with groups of 75 white officers and 40 black officers. The report provided correlation coefficients without significance levels. By converting these into significance levels through a table (Snedecor & Cochran, 1967), one discovers that none of the scales were associated with either supervisor or peer ratings for whites, but for blacks high Order and Autonomy scores and low Heterosexuality scores were associated with high scores on both ratings. High endurance scores were also associated with high supervisor ratings of blacks.

In summary, three post-selection studies produced no replicated evidence of validity for any EPPS scale. Therefore, the value of the EPPS in police-selection must be doubted. However, only pre-selection studies of the EPPS could provide the data needed for a firm conclusion on the usefulness of the scale in selecting police officers.

16 Personality Factor Questionnaire (16PF). Fabricatore et al. (1978) evaluated the concurrent validity of the 16 PF by doing a canonical correlation analysis with all 16 scales and four performance measures: supervisor's paired-comparison rating, supervisor's absolute rating, number of official reprimands and number of preventable

accidents. The resulting correlation, R = .27, was significant. No 0-order correlations were provided.

Henderson (1979) assessed the concurrent validity of the 16PF with groups of 151 police-officer volunteers and 234 officers who completed the scale as part of screening for promotion. In the volunteer group, high Anxiety scores were associated with high peer ratings; for the volunteers, high Brightness scores were associated with high supervisor ratings. Note that the first of the findings is counterintuitive. Given that 32 correlations were calculated, one should put little faith in the validity of the anxiety and brightness findings.

The results of the two post-selection 168F studies were mixed, making pre-selection studies of the scale appropriate.

F Scale of Authoritarianism. Blum (1964, study 4) assessed the predictive post-selection validity of the F scale and found a significant relationship between high F Scale scores and minor disciplinary charges, days lost due to illness and low number of citizen commendations. There was no association with number of auto accidents, serious disciplinary charges, cases with substantial evidence of misconduct, departmental commendations, periods in which ill or number of injuries.

Bass et al. (1954) found no association between F Scale scores and supervisor ratings in two groups of officers. McDonough & Monahan (1975) found no relationship with resignation, termination or promotion.

The evidence of concurrent validity for the F scale is mixed, making pre-selection research appropriate.

Rorschach and Draw-A-Person. McDonough & Monahan (1975) assessed the pre-selection validity of a psychologist's ratings of responses to a group-administered Rorschach and a Draw-A-Person. The study compared groups of fired, resigned, promoted and non-promoted officers and found that the Rorschach ratings discriminated between fired and promoted officers only. Draw-A-Person rating failed to discriminate between any two groups.

Blum (1964, study 4) assessed the predictive post-selection validity of the same tests using his interpretations of responses. The interpretations were not associated with any of 10 outcome variables.

The two studies of projective tests produced overall mixed evidence of validity for clinical interpretations of the Rorschach, suggesting that further evaluation of it may be warranted. Further evaluation of the Draw-A-Person technique seems unwarranted, however.

Rotter Internal-External (IE) Scale. Speilberger et al. (1979) also investigated the post-selection predictive validity of the IE Scale and found no relationship with termination of males or females without being eligible for rehiring. Leiren (1973) found a concurrent relationship between high externalization on the Rotter IE Scale and more auto accidents, but no correlation with absenteeism, letters of recommendations or supervisor's ratings. Together, the studies provided only meager evidence that the IE Scale is valid as a predictor of police performance. However, only pre-selection studies would allow a firm conclusion as to the validity of the scale.

Cornel Word Form 2 & Rosensweig Picture Frustration Study. Dubois & Watson (1950) assessed the pre-selection validity of the Cornel Word Form 2 (neurosis) and the Rosensweig Picture Frustration Study (direction of aggression and type of frustration reaction) and found no relationship with departmental ratings.

Gordon Person Profile. Bass et al. (1954) investigated the concurrent validity of the Hypersensitivity, Ascendency, Sociability and Responsibility scales of the Gordon Person with Profile. The report provided correlation coefficients without significance levels, but one can determine with a table provided by Snedecor & Cochran (1967) that the correlations were all nonsignificant. No association was found with peer or supervisor ratings in a group of 22 city officers or in a group of 37 deputy sheriffs.

Guilford-Martin Temperament Inventory. Marsh (1962) assessed the concurrent validity of the Guilford-Martin Temperament Inventory and found that high General Activity scores were associated with high supervisor ratings but not with involuntary termination, or rate of auto accidents. None of the other four scales were associated with any outcome variable. Snibbe et al. (1973), using 95 of Marsh's subjects 10 years later, assessed the validity of the Guilford-Marting General Activity Scale and found that scores on it were not associated with any of five outcome variables.

Humm-Wadsworth Temperament Scale. Humm & Humm (1950) assessed the post-selection predictive validity of ratings made by "trained" psychometrists on the basis of a number of selection criteria including

the Humm-Wadsworth Temperament Scale. Low ratings were associated with involuntary termination. For lengthy criticisms of the study and its findings, see Blum (1984, pp. 106-107) and Ruch (1965).

Rotters' Incomplete Sentences Blank. Bass et al. (1954) investigated the concurrent relationship between scores on Rotter's Incomplete Sentences Blank and peer ratings and found no association.

State-Trait Anxiety Scale. Speilberger, Spaulding, Jolley, and Ward (1979) assessed the predictive post-selection validity of the State-Trait Anxiety Scale and found no association with resignation without being eligible for rehiring.

Experimental Scales. Shealy (1979) developed a social judgment scale and assessed its concurrent validity with an unspecified number of officers. The scale consists of 15 items intended to produce scores for moral knowledge, socialization, empathy, conscience (internal v. external rules) and resistance to peer pressure for immorality. The study found that low conscience scores were associated with supervisor ratings of corruption. Total scores and scores on the other variables were unrelated to the ratings.

Hogan (1971) developed an empathy scale and found no concurrent association with supervisor ratings. Bass, Karstendiek, McCollough, and Pruitt (1954) assessed the concurrent validity of an empathy scale with 22 deputies and found no association with supervisor ratings.

Baehr & Frommel (1971) assessed the concurrent validity of an experimental psychoanalytic instrument called the Arrow-Dot Test, which produces scores for id, ego and superego. The study used a

cross-validation design with nine outcome variables ranging from supervisor paired-comparison ratings to sustained complaints. The results were reported in terms of multiple regressions done with all three test scores. Although four outcome variables were found to be associated with the scores in one group, cross-validation was nonsignificant.

There have been four police-officer selection studies of the TAV, (Toward-Away-Versus) which contains parts involving preferences, proverbs and sayings, judgment; an adjective check list and "personal data." Each part has approximately 300 items and produces scores on three dimensions related to Karen Horney's work: Toward people (cooperation), away (withdrawal, creative) and versus (competition, aggression) (Mormon, Hankey, Heywood & Liddle, 1966), for a total of 15 part/dimension scores.

All four studies examined concurrent validity. The first (Hankey, Mormon, Kennedy & Heywood, 1965) involved four parts of the TAV, all except "personal data." The outcome measures were supervisor ratings on three dimensions, so in all, 36 correlations were determined. Of these, six were significant.

The second study (Mormon, Hankey, Kennedy & Heywood, 1965) used the first four parts of the TAV and found that four of the 12 scores produced were associated with supervisor ratings of performance.

The third study (Mormon, Hankey, Heywood & Kennedy, 1965) involved the first four parts of the TAV and two outcome measures: supervisor's rankings of performance and police hours/hazardous arrest. Only two of

the 24 correlations were associated with rankings, and they were apparently significant in the unexpected direction.

The fourth study (Mormon, Hankey, Heywood, Liddle & Goldwhite, 1967) used all five parts of the TAV and two outcome measures: supervisor rankings and ratings. Only the three personal data scores were associated with ratings.

Overall, 13 of 102 correlations were significant in the expected direction, but only one score was found to be significant in two studies, Adjective Check List -- Versus. However, the three personal data scores were found to have concurrent validity in the only study in which they were tested. Hence, the Adjective Check List-Versus and the three personal data scores may merit pre-selection research.

Summary of personality and psychopathology measures other than the MMPI. Few of the studies of personality and psychopathology measures other than the MMPI used pre-selection designs, and the pre-selection studies did not provide any replicated evidence of validity. Several CPI scales and one TAV scale were found to have validity in more than one post-selection study, but the findings should be considered suggestive only in view of the limitations of post-selection research.

More pre-selection studies need to be done before one can safely conclude that any one personality/psychopathology measure is or is not a valid predictor of police performance. Further research on personality/psychopathology measures seems warranted in view of the widespread use of the measures in actual police selection (Abramson, 1974; Parisher, et al. 1979) and the conclusion of Ghiselli (1966; 1973) that personality

tests have been shown to be valid selection variables in a wide array of jobs.

Interest Inventories

Strong-Campbell. Speilberger, Jolley, Spaulding, Jolley, and Ward (1979) examined the predictive post-selection validity of the Strong-Campbell Interest Inventory (SCII), which has 23 interest scales and 124 occupation scales (Anastasi, 1976, p. 531). Using resignation within two years without being eligible for rehiring as the outcome variable, the study found that for male officers, high scores on four scales were associated with remaining on the force: business management, office practices, female army officer and male army officer. For females, there were no associations. For neither sex were scores on police-officer or highway-patrol scales associated with remaining on the force. Because 294 correlations were apparently calculated, one would expect more than four to appear significant merely by chance. Hence, the study suggests that the scale would not be useful as a selection variable.

Blum (1964, study 4) assessed the predictive post-selection validity of the Strong-Vocational Interest Blank (SVIB), the predecessor of the SCII. It had 54 occupational scales and four non-occupational scales (Anastasi, 1968). The study, using 10 outcome measures, found that five job interests were associated with one or more outcome measures: psychologist interests with auto accidents; physicist interests with auto accidents; journalist interests with charges of serious misconduct and injuries; and policeman interests with days lost

due to illness. High physicians interests and carpenter interests were each associated with bad performance on at least one outcome variable and good performance on another. Thus, nine of apparently 580 correlations were significant, less than the number one would expect by chance. Also noteworthy is that none of the "significant" scales matched a "significant" one in the study of Speilberger et al. (1979). Dubois & Watson (1950) also assessed the concurrent validity of the SVIB and found no association with service rating.

Kudor Preference Record. Sterne (1960) carried out a concurrent validity study of the five scores provided by the Kudor Preference Record and found no association with supervisor ratings.

Marsh (1962) found no concurrent association between the five scores and involuntary termination, supervisor ratings or auto accidents. Snibbe, Azen, Montgomery, Marsh (1973), using 95 of Marsh's subjects 10 years later, assessed the validity of the Mechanical scale and the Social Services scale of the Kudor and found high Mechanical scores were associated with two measures of promotion and with high supervisor ratings, but not with workers' compensation claims or injuries. Social Service scale scores were not associated with any of the five performance variables.

Summary of Interest Inventories. Overall, the studies suggest that interest inventories are not useful in selecting police officers. This should come as no surprise, since the inventories were designed not for screening but for vocational counseling. Further, studies have shown that responses to the items can easily be faked (Kirchner, 1962; Bridgeman & Hollenbeck, 1961).

Interviews

Landy (1976) assessed the pre-selection validity of interviews done by a three-person panel. During a 45-minute interview, each applicant was rated on nine dimensions and also given a recommendation as to hiring. The nine interview ratings were reduced to three factors through factor analysis: manifest motivation, communication and personal stability. Supervisor ratings on nine dimensions were factor analyzed and then reduced to four outcome factors: professional maturity, technical competence, demeanor and communication. The study found that the hiring recommendation of the interviewers was unrelated to any of the outcome factors. The study also found, however, that the motivation selection factor was associated with the competency and demeanor peformance factors and that the personal stability selection factor was associated with the communication and competency performance factors.

McDonough & Monahan (1975) assessed the pre-selection validity of one-hour "psychiatric interviews." The study found that the interview ratings failed to discriminate among groups of fired, promoted and non-promoted officers. It is noteworthy that the interviewer ratings had an interrater reliability of only .41, making the variable of little value.

Dubois & Watson (1950) assessed the pre-selection validity of ratings based on an interview done with five police-officer interviewers. Each applicant was rated on five dimensions: appearance, manner, speech, adaptability and general impression. These ratings were

then summed for an interview score, which was found to be unrelated to service ratings.

Marsh (1962) assessed the pre-selection validity of "civil service" interview scores and found no association with involuntary termination, supervisor ratings or rate of auto accidents.

Overall, the four pre-selection interview studies provided virtually no support for the validity of interviews in the selection of police officers. This conclusion is similar to that of Tenopyr & Oeltjen (1982) and Guion (1976), who found little or no research that supported the validity of interviews in screening for any job. It is unclear what to make of the validity of Landy's interview factors. Perhaps studies should be done of such factors, rather than of global ratings. Ordinary Interview ratings, however, appear to be undeserving of further research.

Subjective Background Ratings

Cohen & Chaiken (1972) examined the pre-selection validity of background ratings made by an officer who reviewed the records of applicants and interviewed the applicants and sometimes their friends, neighbors, and employers. The study used 14 outcome measures and found that high ratings were positively associated with civil service promotion, few total complaints, few trials, few substantiated complaints, and few times sick. There was no association with promotion by other than civil service, awards, criminal complaints, days sick, injury claim disapprovals or arrests.

McAlister (1970) also assessed the pre-selection validity of background ratings. The study found no evidence of validity with regard to awards, supervisor rating or five measures relating to sick leave and injuries.

The two studies thus produced mixed results that suggest that further exploration of this variable would be appropriate.

Discussion

The review of police selection validation studies leads to the following recommendations regarding research methods and reports: (a) Use only pre-selection designs; (b) when comparing different performance groups to test the validity of a variable, make sure that the groups do not differ with regard to a confounding variable; (c) eliminate if possible, outcome variable contamination, and, if that is not feasible, at least assess the extent of the problem and mention it with the findings; (d) use only meaningful outcome measures, which do not include selection for hiring, training performance, or long tenure; (e) control for different levels of opportunity with outcome measures like number of arrests, and if this is not possible, assess the extent of the opportunity variance and report it with the findings; (f) use outcome variables that cover as many as possible of the important functions of police officers, including non-arrest functions such as mediating family disputes and writing reports; (g) use only outcome measures that have enough variance to allow a possibility of a significant relationship with the selection variable being tested; (h) when using ratings as

outcome variables, consider using alternatives to absolute supervisor ratings, including peer ratings and various rating formats that may lend themselves to more precise ratings, such as paired-comparisons and rankings; (i) when using ratings, make sure that the raters have the training, information and incentive needed to make valid ratings; (j) in reporting results, follow the Guidelines for Reporting Criterion-Related and Content Validity of the Office of Federal Contract Compliance (Anatasi, 1976) and report 0-order correlations for all combinations of selection and outcome variables; (k) recognize and report problems of alpha inflation; (l) place little faith in beta weights unless the rules set out by Cattin (1978) are met; and (m) do not put great effort into a hunt for racial or gender moderators.

In reaching a conclusion as to the validity of the selection variables examined, one must establish a standard for what qualifies as validation. A reasonable standard is that for a selection variable to be considered valid, it must be cross-validated in one study or validated in two or more studies.

Only five biodata items have met this standard. Cross-validated predictors of poor-police-officer performance are (a) prior involuntary termination, (b) having been married more than once, (c) vehicle-code violations, (d) more serious criminal offenses, and (e) short duration of prior jobs.

Several pre-selection and post-selection studies of measures of intellect provided overall only slight evidence of validity. Two pre-selection studies of subjective background ratings likewise produced only mixed evidence of validity.

Five pre-selection studies of the MMPI produced no replicated findings of validity, but some MMPI scales were found to have post-selection validity in more than one study: Mf, Pd, Pt and SC. Because of inherent deficiencies in post-selection research, the findings of post-selection validity should be considered suggestive only. Few pre-selection studies of other personality and psychopathology measures have been reported and these provided, at best, mixed evidence of validity. However, several scales were found to have post-selection validity in more than one study: The Achievement via Conformance, Self-Control, Tolerance and Intellectual Efficiency scales of the CPI and the Adjective Check List-Versus scale of the experimental TAV.

Studies of interest inventories and interviews consistently found no meaningful evidence of validity.

It would seem that future attempts to find pre-selection predictors of police-officer performance would best be directed at biodata, the MMPI, to other measures of personality, measures of intellect, and subjective background ratings. Interviews and interest inventories appear, at this time, not to be promising variables for further validation research regarding police officers.

References

Abramson, S.A. (1974 July). A survey of campus police departments. The Police Chief, 54-57.

Anatasi, A. (1968). Psychological testing (3rd ed.). London: Macmillan.

Anatasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

Azen, S.P., Snibbe, H.M., Montgomery, H.R., Fabricatore, J.M., & Earle, H.H. (1974). Predictors of resignation and performance of law officers. American Journal of Community Psychology, 2, 79-86.

Baehr, M. (1979). Impact of civil rights legislation and court actions on personnel procedures and practices. In C.D. Spielberger (Ed.) Police selection evaluation. Washington: Hemisphere.

Baehr, M., Froemel, E.C. (1977). The Arrow-Dot Test as a predictor of police officer performance. Perceptual & Motor Skills, 45, 683-693.

Baehr, M., Saunders, D.R., Froemel, E.C., & Furcon, J.E. (1973). The prediction of performance for black and white patrolmen. In J.R. Snibbe & H.M. Snibbe (Eds.). The urban policeman in transition. Springfield, Ill.: Charles C. Thomas.

Barrett, G.V., Phillips, J.S., & Alexander, R.A. (1981). Concurrent and predictive validity designs: a critical reanalysis. Journal of Applied Psychology, 66, 1-6.

Bartol, C.R. (1982). Psychology and American Law. Belmont, CA: Wadsworth.

Bass, B.M. (1957). Faking by sales applicants of a forced-choice personality inventory, Journal of Applied Psychology, 41, 403-404.

Bass, B.M., Karstendiek, B., McCollough, G., & Pruitt, R.C. (1954). Validity information exchange no. 7-024. Personnel Psychology, 7, 159-160.

Bem, D.J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. Psychological Review, 81, 506-520.

Bernstein, I.H., Schoenfield, L.S., & Costello, R.M. (1982). Truncated Component regression, multicollinearity and the MMPI's use in a police officer-selection setting. Multivariate Behavioral Research, 17, 99-116.

Blum, R.H. Police selection. (1964). Springfield, Ill.: Charles C. Thomas,

Bridgeman, C.A., & Hollenbeck, G.P. (1961). Effect of simulated applicant status on Kudor Form D occupational interest scores. Journal of Applied Psychology, 45, 237-239.

Burkhart, B.R. (1980, February). Conceptual issues in the development of police-selection procedures, Professional Psychology, 121-129.

Caplan, J.R., & Schmidt, F.L. (1977). The validity of education and experience ratings. Paper presented at the Annual Meeting of the International Personnel Management Association Council, Kansas City.

Carlson, H.M., & Sutton, M.S. (1979). Some factors in community evaluation of police street performance. American Journal of Community Psychology, 7, 583-591.

Carr, A.F., Larson, L.D., Schnelle, J.F., & Kirchner, R.E. (1980). Outcome measure of police performance: Some steps toward accountability, Journal of Community Psychology, 8, 165-171.

Cascio, W.F. (1977). Formal education and police officer performance. Journal of Police Science and Administration, 5(1), 89-96.

Cascio, W.F. (1978). Applied psychology in personnel management. Reston, VA: Reston.

Cascio, W.F., & Real, L.J. (1979). The civil service exam has been passed: Now what? In C.D. Spielberger (Ed.) Police selection and evaluation. Washington: Hemisphere.

Cascio, W.F., & Valenzi, E.R. (1978). Relations among criteria of police performance. Journal of Applied Psychology, 63(1), 22-28.

Cattin, P. (1978). A predictive-validity-based procedure of choosing between regression and equal weights. Organizational Behavior and Human Performance, 22 93-102.

Cohen, R., & Chaiken, J. (1972). Police background characteristics and performance: Summary report (No. R-999-DDJ). New York: Rand Institute.

Colarelli, N.J., & Siegel, M.A. (1964). A method of police personnel selection. Journal of Criminal Law, Criminal Justice and Police Science, 55, 287-289.

Costello, R.M., Schoenfield, L.S., & Kobos, J. (1982). Police applicant screening: An analogue study. Journal of Clinical Psychology, 38, 216-221.

Crosby, A., Rosenfield, M., & Thornton, R.F. (1979). The development of a written test for police applicant selection. In C.D. Speilberger (Ed.), Police selection evaluation. Washington: Hemisphere.

Cross, C.A., & Hammond, K.K. (1951). Social differences between "successful" and "unsuccessful" state highway patrolmen. Public Personnel Review, 12, 159-161.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34, 571-582.

Dubois, P.H., & Watson, R.I. (1950). The selection of patrolmen. Journal of Applied Psychology, 34, 90-95.

Due, F.D. Screening police applicants. The Police Chief, 1975 (Feb), 73-74.

Dunnette, M.D. (1966). Personnel selection and placement. Belmont, CA: Brooks/Cole.

Dunnette, M.D., & Borman, W.C. (1979). Personnel selection and classification systems. Annual Review of Psychology, 30, 477-525.

Elliott, A.G.P. (1976). Fakers: A study of managers' responses on a personality test. Personnel Review, 5, 33-37.

Elliot, A.G. (1981). Some implications of lie scale scores in real-life selection. Journal of Occupational Psychology, 54, 9-16.

Fabricatore, J., Azen, S., Shoentgen, S., & Snibbe, H. (1978). Predicting performance of police officers using the Sixteen Personality Factor Questionnaire. American Journal of Community Psychology, 6(1), 63-70.

Finnigan, J.C. (1976, August). A study of relationships between college education and police performance in Baltimore, Maryland. The Police Chief, 60-62.

Flynn, J.T., & Peterson, M. (1972). The use of regression analysis in police patrolmen selection. Journal of Criminal law, Criminology and Police Science, 63, 564-569.

Furcon, J. E., Froemel, E.C., & Baehr, M.E. (1973). Psychological predictors and patterns of patrolman field performance. In J.R. Snibbe & M.M. Snibbe (Eds.), The urban policeman in transition. Springfield, Ill.: Charles C. Thomas.

Ghiselli, E.E. (1966). The validity of occupational aptitude tests. New York: Wiley.

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

Gordon, M.E., & Kleiman, L.S. (1976). The prediction of trainability using a work sample test and an aptitude test: A direct comparison. Personnel Psychology, 29, 243-253.

Gottlieb, M.C., & Baker, C.F. (1974). Predicting police officer effectiveness. Journal of Forensic Psychology, 6, 35-46.

Gough, H.G. (1966). Appraisal of social maturity by means of the CPI. Journal of Abnormal Psychology, 71, 189-195.

Gough, H.G. (1969). A leadership index on the CPI. Journal of Counseling Psychology, 16, 283-289.

Green, R.F. (1951). Does a selection situation induce testeess to bias their answers on interest and temperament scales? Educational and Psychological Measurement, 11, 503-515.

Griswold, V. (1965). Connecticut, 381 U.S., 479.

Gruenicld, E.F: (1981). Performance appraisal: Promise and peril. Ithaca, NY: Cornel University.

Guion, R.M. (1976). Recruiting, selection, and job replacement. In M.D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand-McNally.

Hankey, R.J., Mormon, R.R., Kennedy, P., Heywood, H.C. (1965, March, April). TAV selection system and state traffic officer job performance. Police, 10-13.

Hausman, H.J., Strupp, H.H. (1955). Non-technical factors in supervisors' ratings of job performance. Personnel Psychology, 8, 201-217.

Hebb, D.O. (1949). The organization of behavior. New York: Wiley.

Henderson, N.D. (1979). Criterion-related validity of personality and aptitude scales: A comparison of results under volunteer and actual test conditions. In C.D. Speilberger (Ed.), Police selection evaluation. Washington: Hemisphere.

Heron, A. (1956). The effects of real-life motivation on questionnaire response. Journal of Applied Psychology, 40, 65-68.

Herzberg, F. (1954). Temperament measures in industrial selection. Journal of Applied Psychology, 38, 81-84.

Hogan, R. (1971). Personality characteristics of highly rated policeman. Personnel Psychology, 24, 679-686.

Hogan, R. & Kurtines, W. (1975). Personalogical correlates of police effectiveness. Journal of Psychology, 91, 289-295.

Humm, D.G., & Humm, K.A. (1950. Humm-Wadsworth Temperament Scale appraisals compared with criteria of job success in the Los Angeles police department. Journal of Psychology, 30, 63-75.

Hunter, J.E., Schmidt, F.L. & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 21-35.

Kent, D.A., & Eisenberg, T. (1972, February). The selection and promotion of police officers: A selected review of recent literature. The Police Chief, 20-29.

Kerlinger, F.N., & Pedhauzer, E.J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart & Winston.

Kim, J., & Kohopt, F.J. (1975). Multiple regression analysis: Subprogram regression. In H. Nie, C. Hull, J. Jenkins, K. Steinorenner & D. Bent (Eds.), Statistical package for the social sciences. New York: McGraw-Hill.

King, L.M., Hunter, J.E., & Schmidt, F.L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. Journal of Applied Psychology, 65, 507-516.

Kirchner, W.K. (1962). "Real-life" faking on the Strong Vocational Interest Blank by sales applicants. Journal of Applied Psychology, 46, 128-130.

Landy, F.J (1976). The validity of the interview in police officer selection. Journal of Applied Psychology, 61, 193-198.

Landy, F.J., Farr, J.L., Saal, R.E., & Freytas, W.R. (1976). Behaviorally anchored scales for rating the performance of police officers. Journal of Applied Psychology, 61, 730-738.

Lefkowitz, J. (1977). Industrial-organizational psychology and the police. American Psychologist, 32, 346-364.

Leiren, B.D. (1973). Validating the selection of deputy marshals. In J.R. Snibbe & H.M. Snibbe (Eds.), The urban policeman in transition. Springfield, Ill.: Charles C. Thomas.

Lester, D. (1979). Predictors of graduation from a police training academy. Psychological Reports, 44, 362.

Levy, R.J. (1967). Predicting police failures. Journal of Criminal Law, Criminology and Police Science, 58, 265-276.

Levy, R.J. (1973). A method for identification on the high-risk police applicant. In J.R. Snibbe & H.M. Snibbe (Eds.), The urban policeman in transition. Springfield, Ill.: Charles C. Thomas.

Love, K.G. (1981). Comparison of peer assessment methods: reliability, validity, friendship bias and user reaction. Journal of Applied Psychology, 66, 451-457.

Love, K.G. (1983). Empirical recommendations for the use of peer rankings in the evaluation of police officer performance. Public Personnel Management Journal, 25-32.

Lykken, D.T. (1979). The detection of deception. Psychological Bulletin, 86, 47-53.

Marsh, S.H. (1962). Validating the selection of deputy marshals. Public Personnel Review, 23, 41-44.

Martin, E.M. (1923). An experiment in new methods of selecting policemen. Journal of Criminal Law, 14, 671-681.

McAllister, J.A. (1970, March-April). A study of the prediction and measurement of police performance. Police, 58-64.

McDonough, L B , & Monahan, J. (1975) The quality of community caretakers. A study of mental health screening in a sheriff's department Community Mental Health Journal, 11, 33-43

Meehl, P.E. (1971). Law and the fireside inductions: Some reflections of a clinical psychologist Journal of Social Issues, 27, 65-100.

Meehl, P.E., Rosen A. (1955) Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. Psychological Bulletin, 52, 194-216.

Merian, E.M , Stefan, D., Schoenfield, L.S , & Kobos, J.C. (1980). Screening of police applicants: A five item MMPI research index. Psychological Reports, 47, 155-158

Michaelis, W., & Eysenck, H.J. (1971). The determination of personality inventory factor patterns and intercorrelations by changes in reall life motivation. Journal of Genetic Psychology, 118, 223-234.

Mills, C.J , & Bohannon, W.E. (1980). Personality characteristics of effective state police officers. Journal of Applied Personality, 65, 680-684.

Mills, R B., McDevitt, R.J., & Tonkin, S. (1966). Situational tests in metropolitan police recruit selection. Journal of Criminal Law, Criminology and Police Science, 57, 99-104.

Mills, M.C., & Straton, J.C. (1982, February). The MMPI and the prediction of police job performance. FBI Law Enforcement Bulletin, 10-15.

Mormon, B.R., Hankey, R.D., Heywood, H L., & Kennedy, P.K. (1965, July/August). Multiple relations of TAP selection system predictors to state traffic officer performance. Police, 41-44.

Mormon, R.R., Hankey, R.D., Heywood, H.L., & Liddle, L.R. (1966, July/August). Predicting state traffic officer cadet academic performance from theoretical TAV selection system scores. Police, 54-58.

Mormon, R.R., Hankey, R.D., Kennedy, P., Heywood, H.L. (1965, May/June). Predicting state traffic officer performance with TAV selection system theoretical scoring keys. Police, 70-73.

Mormon, R.R., Hankey, R.D., Kennedy, P., & Jones, E.M. (1966, July/August). Academic Achievement of state traffic officer cadets related to TAV selection system plus other variables. Police, 30-34.

Mormon, R.R., Hankey, R.D., Heywood, H.L., Liddle, L.R., & Goldwhite, M. (1967, January/February). Multiple prediction of police officers' ratings and rankings using theoretical TAV system and certain non-test data. Police, 19-22.

Mullineaux, J.E. (1955). An evaluation of the predictors used to select patrolmen. Public Personnel Review, 16, 84-86.

Nunnally, J.C. (1978). Psychometric theory. New York: McGraw-Hill

Owens, W.A. (1976). Background data. In M.D. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally.

Parisher, D., Rioa, B., & Reilly, R.R. (1979). Psychologists and psychological service in urban police departments: A national survey. Professional Psychology, 10, 6-7.

Podlesny, J.A., & Raskin, D.C. (1977). Physiological measures and the detection of deception. Psychological Bulletin, 84, 782-799.

Poland, J.M. (1978). Police selection methods and the prediction of police performance. Journal of Police Science and Administration, 6(4), 374-392.

Roscoe, J.T. (1975). Fundamental research statistics. New York: Holt, Rinehart & Winston.

Rosenthal, R., & Rosnow, R.L. (1969). The volunteer subject. In R. Rosenthal & R.L. Rosnow (Eds.), Artifact in behavioral research. New York: Academic Press.

Rothe, F.H. (1947). Distributions of test scores of industrial employees and applicants. Journal of Applied Psychology, 31, 480-483.

Ruch, F.L. (1965). The Humm-Wadsworth Temperament Scale. In O.K. Buros (6th ed.), Mental Measurements Yearbook, 252-254.

Sackett, P.R., & Decker, P.J. (1979). Detection of deception in the employment context: A review and critical analysis. Personnel Psychology, 32, 487-506.

Saxe, S.J., & Reiser, M. (1976). A comparison of three police applicant groups using the MMPI. Journal of Police Science and Administration, 4, 419-424.

Schmidt, F.L., & Hunter, J.E. (1978). Moderator research and the law of small numbers. Personnel Psychology, 31, 215-231.

Schmidt, F.L., Berner, J.G., & Hunter, J.E. (1973). Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 62, 529-540.

Schoenfield, L.S., Kobos, V.C., & Phinnery, I.R. (1980). Screening police applicants: A study of reliability with the MMPI. Psychological Reports, 47, 419-425.

Schrader, A.D., & Osburn H.G. (1977). Biodata faking: Effects of induced subtley and position specificity. Personnel Psychology, 30, 395-404.

Seashore, S.E., Indik, B.P., & Georgopoulos, B.S. (1960). Relationships among criteria of job performance. Journal of Applied Psychology, 44, 195-201.

Shapiro, V., Thompson, (1969). 394 U.S. 618.

Shealy, A.E. (1979). Police corruption: Screening out high-risk applicants. In C.D. Speilberger (Ed.), Police selection evaluation, Washington: Hemisphere.

Siegel, L., & Lane, I.M. (1974). Psychology in industrial organizations. Homewood, Ill.: Irwin.

Smith, P.C., & Kendall, L.M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Smith, D.H., & Stotland, E. (1973). A new look as police officer selection. In J.R. Snibbe & H.M. Snibbe (Eds.), Urban policeman in transition. Springfield, Ill.: Charles C. Thomas.

Snibbe, H.M., Azen, S.P., Montgomery, H.R., & Marsh, S.H. (1973). Predicting job performance of law enforcement officers: A ten and twenty-year study. In J.R. Snibbe & H.M. Snibbe (Eds.), Urban policeman in transition. Springfield, Ill.: Charles C. Thomas.

Snedecor, G.W., & Cochran, W.G. (1967). Statistical methods (6th ed.), Ames: Iowa State University Press.

Sparling, C.L. (1975). The use of educational standards as selection criteria in police agencies: A review, Journal of Police Science and Administration, 3, 332-334.

Spencer, G , & Nichols, R: (1971, June). A study of Chicago police recruits: Validation of selection procedures. The Police Chief, 50-55.

Spielberger, C.D. Spaulding, H.C., Jolley, M.T., & Ward, J.C. (1979). Selection of effective law enforcement officers: The Florida police standards research project. In C.D. Spielberger (Ed.) Police selection and evaluation. Washington: Hemisphere.

Spielberger, C.D., Ward, J.C., & Spaulding, H.C. (1979). A model for the selection of law enforcement officers. In C.D. Speilberger (Ed.), Police selection and evaluation. Washington: Hemisphere.

Sterne, D.M. (1960). Use of the Kudor Preference Record, Personal, with police officers, Journal of Applied Psychology, 44, 323-324.

Tenopyr, M.L., & Oeltjen, P.D. (1982). Personnel selection, Annual Review of Psychology, 33, 581-618.

Thweatt, W.H. (1972, November). Improving police selection on a shoestring budget. The Police Chief, 60-63.

Van Maanen, J. (1975). Police Socialization: A longitudinal examination of job attitudes in an urban police department. Administration Science Quarterly, 20, 207-228.

Zednick, S., & Blood, M.R. (1974). Foundations of behavioral science research in organizations. Monterey: Brooks/Cole.