An Evaluation of Home Office Extended Interviews for Police Personnel

Robert Tucker Feltham

PhD

The University of Aston in Birmingham

Submitted August 1986

The University of Aston in Birmingham

"An Evaluation of Home Office Extended Interviews for Police Personnel"

Robert Tucker Feltham

Submitted for the degree of PhD, 1986

SUMMARY

In this thesis the validity of an Assessment Centre (called 'Extended Interview') operated on behalf of the British police is investigated. This Assessment Centre (AC) is used to select from amongst internal candidates (serving policemen and policewomen) and external candidates (graduates) for places on an accelerated promotion scheme. The literature is reviewed with respect to history, content, structure, reliability, validity, efficiency and usefulness of ACs, and to contextual issues surrounding AC use. The history of, background to and content of police Extended Interviews (EIs) is described, and research issues are identified. Internal validation involved regression of overall EI grades on measures from component tests, exercises, interviews and peer nominations. Four samples numbering 126, 73, 86 and 109 were used in this part of the research. External validation involved regression of three types of criteria - training grades, rank attained, and supervisory ratings - on all EI measures. Follow-up periods for job criteria ranged from 7 to 19 years. Three samples, numbering 223, 157 and 86, were used in this part of the research. In subsidiary investigations, supervisory ratings were factor analysed and criteria intercorrelated. For two of the samples involved in the external validition, clinical/judgemental prediction was compared with mechanical (unit-weighted composite) prediction. Main conclusions are that: (1) EI selection decisions were valid, but only for a job performance criterion; relatively low validity overall was interpreted principally in terms of the questionable job relatedness of the EI procedure; (2) EIs as a whole had more validity than was reflected in final EI decisions; (3) assessors' use of information was not optimum, tending to over-emphasize subjectively derived information particularly from interviews; and (4) mechanical prediction was superior to clinical/judgemental prediction for five major criteria.

KEY WORDS: Assessment Centres, United Kingdom, Police Personnel, Predictive Validity, Management.

ACKNOWLEDGEMENT

The research reported in this thesis would have been i
without the advice and considerable help with data c
provided by Chief Superintendent John Linnane of th
Extended Interview Office (Home Office).

LIST OF CONTENTS

LIST OF FIGURES

9 Pearson Correlation Coefficients - Interviews with EI Final Mark - Samples 4 - 7

10 Hierarchical Multiple Regression of EI Final Mark on al EI Component Measures Entered in Sequence of EI Procedure - Samples 4 - 7

11 Pearson Correlation Coefficients - Final Mark and a Unit-Weighted Composite with Five Criteria - Sample 2 (Special Course)

2 Pearson Correlation Coefficients - Final Mark and a Unit-Weighted Composite with Five Criteria - Sample 3 (Graduate Entry)

LIST OF TABLES

Chapter 1 - ASSESSMENT CENTRES AND THEIR HISTORY

The term 'Assessment Centre' (AC) describes a family of psychological assessment systems used for management selection, promotion, training and career development. As a working definition it can be said that an AC consists of the 'assessment of a group of individuals by a team of judges using a comprehensive and integrated series of techniques' (Fletcher, 1982; p.42). These techniques include psychometric tests, group exercises, written exercises, in-baskets, individual role-plays, interviews and peer assessments.

The AC was never really 'invented'. It is rather the cumulative result of the work of a number of personnel practitioners over time. This said, a recognizable historical starting point can be found in multiple assessment procedures developed for officer selection in pre-war Germany. This inspired the British War Office Selection Board (WOSB) which was in turn the forerunner of the Civil Service Selection Board and similar selection procedures in the armed services and in Commonwealth countries. The first application of ACs in the United States was also WOSB derived. This was for the selection of wartime secret agents by the Office of Strategic Studies (OSS). Later, in the '50s and '60's, AC programmes at American Telephone and Telegraph set the standard for widespread applications in both private and public sector organizations. This US type of AC has inspired a number

of applications in British industry in recent years.

In the following sections a closer look will be taken at each of these main historical developments.

German Origins

Limitations on the size of the German army imposed after World War I by the Versailles Treaty, combined with the desire of German militarists to build a 'nucleus army of leaders' from amongst an immense number of volunteers (Farago, 1972, p.45), provided the impetus for the development of a new technology of psychological assessment for use in officer selection. This development was helped by the fact that military psychology was expanded in Germany in the 1920s and 1930s, in contrast to English speaking countries where it lapsed completely (Vernon and Parry, 1949). That the technology was a major departure from established psychometrics-centred approaches may have been due in part to the influence of the Berlin based Gestalt school. While primarily a movement in experimental psychology it had a great effect on psychological thinking in general (Thomson, 1968). Its holistic approach to psychological phenomena is very much in line with the principles on which selection for the commissioned ranks in the German army came to be based.

Most of the developmental work was done in the 1920s. During the 10 years preceding World War II programmes underwent relatively little change (Fitts, 1946). In 1939 Simonheit, director of German army psychology, produced a statement of the principles said to underlie officer selection. These have been translated by Ansbacher (1941), and three are quoted here:

> 'The whole personality must be considered ... One must not be led too hastily by the first impression ... [or] compile a list of the proper attitudes for a soldier and expect someone to have them all ... The question is rather whether the candidate will be likely to live up to the best in his own personality.
>
> ... The examination must keep close to everyday life ... The method of intelligence tests has been abandoned; tasks of a serious character which are in rapport with daily life are given instead ...
>
> ... The candidate's conduct should be observed throughout the entire examination. The candidate's way of performing a task is considered more prognostic than his achievement' (p.379)

It is the concept of assessment as holistic, continuous, and focused on performance on a range of tasks chosen to represent salient aspects of the real world that distinguishes the German approach from what had gone before. However a concentration on the assessment of somewhat elusive qualities of leadership, such as 'will power, ... mental energy, sustaining power, and readiness to act to the limit of mental capacity' (Farago, 1972, p.51), led to extreme subjectivity. The style of the approach is indicated by Ansbacher (1941):

> 'In tests where scores are obtained, these are almost incidental. Judgment is subjective. From many samples of behaviour symptoms are observed; from a number of symptoms conclusions as to a personality trait are drawn; and judgment

of a trait is not made until it seems to fit with the
of the total personality' (p.380)

Descriptions of the procedures used are given by Ans
(1941), Fitts (1946), Vernon and Parry (1949) and Farago (
The general pattern was that candidates were assessed o
lays, made up of two days of assessment and a 'rest d
etween during which candidates continued to be closely obs
he board consisted of a colonel (who made the final decisi
edical officer and three psychologists/psychological exami
During the early years ... psychologists were sel
arefully and high professional standards were maintained
on, however, the demand for trained men to staff the expa
sting program exceeded the supply and standards were low
dividuals whose specialization had been in ... other fi
re given brief training courses and assigned as psycholog
aminers' (Fitts, 1946, p.152). Individual candidates
located to a psychological examiner who administered all
lividual tests, interviewed the candidate, and mad
commendation regarding selection and classification.

ts and exercises were gouped as follows:

ellectual and other abilities were assessed us
ective-type tests and open-ended tests such as writin
cription of a cinema film and answering practical problems
y form.

Character and temperament were assessed with projective and situational tests. Examples of the latter are: a lengthy choice reaction test designed to test such things as power of sustained attention and emotional control; 'command series' tasks which involved carrying out complex orders requiring agility, attention, powers of memory etc.; and 'leadership sample' tests which consisted of instructing a group of soldiers in some mechanical task, e.g. making a coat hanger out of a piece of wire.

Expression analysis was a term used to encompass measures of handwriting, literary style, style of speech, and facial expression in reaction to distractions or painful stimuli, e.g. electric shock, which were taken as personality indicators.

Life-history (recorded information plus interview) was studied for motivational, self-evaluative and attitudinal characteristics.

The assessment was rounded off with a group discussion of some topic by all the candidates designed to show up their competitiveness and other social reactions.

These officer selection programmes were discontinued around 1941/1942 due to a number of political and practical difficulties (Fitts, 1946). Their legacy is the originality of the approach rather than any proven merit. Fitts (1946) was unable to find any evidence of adequate follow up studies after extensive inquiry, and attributes the lack of interest in validation to the

:ance on non-quantitative methods and lack of familiarity wi
:istical method. However the use of interviews a
entional objective tests (albeit not objectively scored)
unction with group discussions, written problems of
tical nature, and naturalistic military tasks provided tl
iration for the first War Office Selection Boards in Britai

ffice Selection Boards

iptions and historical accounts of War Office Selectio
s (WOSBs) are provided by Garforth (1945), Harris (1949)
ernon and Parry (1949). The development of WOSBs appears tc
been motivated by widespread dissatisfaction with the systen
arter- to half- hour interviews used to select temporary
rs in the first two years of war. According to Vernon and
(1949):

> system worked fairly effectively so long as there was a
> e supply of good material, e.g. from the public schools.
> when this source began to dry up, the boards, being faced
> recruits whose social and educational backgrounds were
> ely unfamiliar, were unable to discriminate effectively.
> table candidates were often passed and sent to ...
> ning] where large proportions failed, with unfortunate
> ts on the morale of the remainder. Moreover, so many
> dates who might have succeeded were rejected by the
> s, often - according to their own accounts - on flimsy
> ds such as Grammar School education or socialist opinions,
> recruits lost confidence in the system, and there was a
> danger of insufficient officers being forthcoming. Again,
> was less opportunity than in 1914-18 for selection on the
> of performance in battle' (pp.52-53).

In early 1941 there were experiments in multiple assessment encouraged by an ex-military attache in Berlin who had observed some of the German selection techniques. On the basis of this experience the first WOSB had been set up by January 1942. Twelve boards using similar methods were in operation in Britain by October 1942, and later on boards were set up for the British army overseas. By the end of the war some 140,000 candidates had been assessed, of whom about 60,000 passed.

While there were many variations between boards, they seem mostly to have conformed to a general outline. Boards consisted of a President and Deputy President (Colonel/Lieutenant-Colonel), Military Testing Officers (MTOs; Major/Captain rank) and a Psychological Department consisting of a Psychiatrist (in charge), a Psychological Officer (not necessarily a qualified Psychologist), and Sergeant-testers. An MTO would take charge of a group of 7 to 12 candidates for three days, mess with them, observe their 'natural social behaviour', and put them through a variety of practical/situational tests. These tests consisted of such things as lecturettes, obstacle courses, 'command situations' in which the candidate was placed in charge of a group faced with some specified task, and leaderless group tests. Leaderless group tests consisted of unstructured group discussions and 'progressive group tasks' which involved 'management of men and materials ... [and] several sub-tasks or obstacles of progressive difficulty and increasing frustration' (Harris, 1949, p.28). During the three days candidates also

completed intelligence tests, projective tests, biographic
nedical questionnaires, and peer assessments. The
interviewed by the more senior army officers and sometimes
sychiatrist. At the end of the procedure there was
conference in which each candidate was discussed by
resident, MTO and Psychiatrist before the President gave a
ecision. For most of the war the Psychologist's role appe
ave been a rather peripheral research/advisory one w
nvolvement in day-to-day selection.

ie first WOSBs appear to have been somewhat hastily ass
th many variations between Boards. Vernon and Parry
ngle out the situational tests for particular criticism:

'None of the leaderless groups or other M.T.O. techniques
standardised tests; no measurements were taken, and only
limited extent were observations recorded under any sta
scheme' (pp.64-65)

ertheless there was a process of gr
inement/improvement/standardization of methods during
rse of the war. The leaderless group tests evolved in 194
team of psychiatrists and psychologists represented a g
ovement in situational testing. These tests also came
ide a common meeting place for board members, who found
rmation gleaned complementary to their interviews.
ings were symptomatic of a general trend toward grea
aboration of assessors. To begin with they studied
idates independently until final conference, resulting
ous disagreements at a time when it was too late to m

further investigation of doubtful candidates. The introduction of collaboration and mutual consultation at all stages was found to have considerable advantages (Vernon and Parry, 1949).

Morris (1949) and Vernon and Parry (1949) cite some evidence as to the validity/utility of war-time WOSBs. In 1942 some selection boards were using the new multiple assessment procedures while others continued to use the old interview methods. Candidates were then sent to training units where instructors were unaware of the selection method applied in individual cases. 37% of the 491 candidates selected by the old methods were marked 'below average' and 22% 'above average', compared with 25% and 35% respectively for the 721 selected using the new methods. Later in the war assessments of over 500 officers from 16 WOSBs were found to correlate 0.165 (0.35 after correction for range restriction) with assessments of field performance. After the war Reeve (1971) conducted some further validation studies for the years 1946-53. In a twelve month follow-up of 3965 officer cadets selected in 1947, WOSB and training grades were found to correlate 0.217 ($p< 0.001$). In two similar six month follow-ups of cadets selected in 1950, WOSB and training grades were found to correlate 0.280 ($N=649$, $p<0.001$) and 0.153 ($N=684$, $p<0.001$). While the validity coefficients in general are not large (uncorrected coefficients accounting for between 2% and 8% of criterion variance), they are at least positive, and the 1942 comparison provides a good demonstration of the practical benefits that even a selection method with low

.dity can bring.

ell as the validity evidence there is some evidence as t
reliability (Vernon and Parry, 1949). In an experiment i
early part of the war 116 candidates attended two boards
night apart. Board reliability, as represented by
elation of Final Grades, was 0.67. In a more refined but i
ways more limited experiment in 1945, two teams o
ienced assessors interviewed the same 125 candidates and a
bility coefficient of 0.8 was found for their final
ments. These experiments are discussed in more detail in
er four.

biliity and validity coefficients, however, do not fully
t the utility of WOSBs in the context of wartime manpower
ges. They won wide acceptance in the army and stimulated
tment of candidates. The 'combination of slightly more
selection procdedures with greater attractiveness to
ates resulted in the sending of two-and-a-half times as
bove-average cadets to ... [training units] as the old
would have done, within five months after the
shment of new boards' (Vernon and Parry, 1949, p.124).
was also a general boost to morale as, despite
logical shortcomings, the 'Army was led to believe that it
getting the best possible officers' (Vernon and Parry,
66).

If one compares WOSBs with their German predecessors, three main developments are apparent. Firstly there was a move away from the extreme subjectivity which characterized the German approach (though WOSB assessments remained subjective). Secondly the importance of validating the procedures was recognized. And thirdly there was a move towards assessors working as integrated teams as opposed to individually until the end; this set the pattern for later British ACs.

WOSBs continued in operation for applicants for national service and short service commissions until 1961. After that assessment of applicants was centralized at the Regular Commissions Board. WOSBs were the direct forefrunner of the British Civil Service Selection Board (CSSB), which will shortly be described, and ACs in Commonwealth countries which are WOSB and CSSB derived. In the next section, however, a look will be taken at the Office of Strategic Studies programme which developed WOSB methods in a US application.

The OSS Programme

The development of the Office of Strategic Services assessment programmes is described in a book by the OSS Assessment Staff (1948) and by Mackinnon (1977) who was Director of the original programme.

The OSS was a US government agency set up in 1942 to deal such things as secret intelligence, operations behind enemy nd black propaganda. There were various methods of recrui o OSS but, according to MacKinnon (1977), all were:

'without benefit of any professional or uniform scre process. Nobody knew who would make a good spy or an effe guerrilla fighter. Consequently, large numbers of misfits recruited from the very beginning, and this might continued had it not been for several disastrous operat such as one in Italy for which, on the assumption that it t dirty men to do dirty works, some OSS men had been recru directly from the ranks of Murder, Inc. and the Philadel Purple Gang. The need for professional assistance in selec was obvious' (pp.14-15)

the suggestion of an OSS official back from London in Oct 43, the agency gave consideration to the adoption of WOSB st sessment procedures. By the beginning of 1944 the fi gramme was in operation. While OSS programmes were hig ivative of WOSB to the extent that some situational tests w rowed complete, MacKinnon (1977) points to the influence . Murray in shaping their development. His experiments sonality assessment at the Harvard Psychological Clinic in 1)s (Taft, 1959; Thornton and Byham, 1982) had led him ir gnostic committee assessments of personality using su niques as interviews, questionnaires, objective a ective tests, group discussions and problem-solvi cises.

The first OSS programmme - Station S - lasted three and a half days and was used to select personnel for overseas assignments. Later, in Winter 1944, a one day programme - Station W - was added, broadening the scope of selection to include appointments to headquarters and rear bases overseas. During the period of their operation the two stations assessed 5,391 recruits. The Station S programme will be briefly described here.

Station S was surrounded in extreme secrecy. Candidates were driven to it in a completely closed van, and while there had to maintain a cover story concealing their true identity, from assessors as well as other candidates. The reason for these measures was concern about possible infiltration by foreign agents. However, the net effect was to turn the whole programme into one large scale simulation, given that many candidates would eventually be living abroad under cover (Mackinnon, 1977).

Specific traits/dimensions were rated on each candidate's final report, and these also provided a rationale for the assessment procedures used. The dimensions were: motivation, practical intelligence, emotional stability, social relations, leadership, physical ability, observation and reporting, propaganda skills, and maintaining cover. The core of the programme was a detailed life-history psychiatric interview which utilized information from various questionnaires and from a projective test. There were also pencil and paper ability and aptitude tests and a range of situational tests.

The situational tests are described in detail by Assessment Staff (1948) and Mackinnon (1977). Examples Map Memory Test where a candidate had to assume he had time to memorize a map at a secret rendezvous, and t answer multiple-choice questions on it; the Constructio which a candidate had to direct two stooge helpers (o and sluggish, the other agressive, critical an impractical suggestions) in an outdoor construction ta Stress Interview which was a simulated interrogation a candidate had supposedly been caught going through night in a government office where he had no right to be

The organization of OSS assessment is described in this from MacKinnon (1977):

'For each assessment class, usually consisting of 18 a the staff was divided into teams of two senior staff (professionals with PhD or MD degrees) and one junio member (enlisted men who had had some training in psyc Each team was assigned to a group of five to seven can The senior members conducted the life history intervi junior member administered special individual tes interpreted the projective test protocols. Otherwise, of both senior and junior members was the same - nam develop as a group as complete a conception as possible candidate in the subgroup assigned to them.

During the various situational tests, the behaviours participant were carefully noted by both senior an menbers of the responsible team, each staff member rati assessee on the variables relevant to the particular e Usually immediately following each situational test, the team met to discuss their impressions of the candidates ratings they had assigned to them. The purpose of the was to come to agreement upon the ratings to be assi each assessee on each of the rated variables.' (p.25)

The predictive validity of OSS programmes was assessed using four appraisal measures based to differing degrees on field performance data. Correlations with overall OSS ratings, after correction for restriction of range, varied from 0.08 (N=53) to 0.37 (N=88) for Station S and 0.15 (N=158) to 0.53 (N=83) for Station W (OSS Assessment Staff, 1948). The highest correlations for each station were obtained using the 'Overseas staff' appraisal measure; this was considered to be the most valid criterion as it was based on the ratings of several observers. Using the correlations with this measure, Wiggins (1973) reanalysed the data to assess the utility of the OSS programmes. He estimated that 63% of pass/fail decisions would have been 'correct' if Station S had used random selection, whereas in reality 77% were 'correct'. Corresponding figures for Station W were 66% and 84%. Since selection ratios were high (75%) and candidates were carefully screened prior to assessment (both factors likely to reduce utility), the figures appear to demonstrate that OSS programmes served a useful purpose.

In assessing the overall contribution of OSS assessment to the development of AC technology, the most significant innovation seems to have been the introduction of systematic rating of behaviour on specified dimensions. While both the German programmes and early WOSBs had involved the tacit assumption that certain types of characteristics were being assessed, this had not been incorporated into procedures in any structured way. From 1945 WOSBs also started to include ratings according to a

standard personality profile (Harris, 1949). Today dimensions are widely though not universally used However, as will be discussed in chapter 2, the ratio their use may be questioned.

The Civil Service Selection Board (CSSB).

According to the Davies Report (1969), when World War II an end:

> 'the Civil Service Commission was faced with the filling the vacancies which had accumulated since 193 Administrative Class and the Senior Branch of the Service. The traditional type of academic examina seemed inappropriate to candidates whose studies h disrupted by the war and it was therefore decided to principle the methods of the War Office Selection Bo number of differences, however, were immediately estab CSSB, unlike WOSB, was not made solely responsibl selection: it was preceded by a written qualifying exam and was succeeded by a Final Selection Board to which i recommendations. There was much more emphasis at intellectual capacity, and on paperwork to test it, and of physical aptitude were eliminated.' (para. 3.3)

he report goes on to describe how CSSB's design was 'base etailed analysis of the work of the Administrative C ssistant Secretary level carried out in 1945 by Dr. E. ith help from Sir Cyril Burt. Information was gathered ities of 505 Assistant Secretaries and their Departments ked to assess the relative importance for the work of e urteen different job headings' (para. 6.10). Accordi stey (1977):

'The new "extended interview" procedure, as it was called, was regarded as experimental during the "Reconstruction" period from 1945-48. At the end of this period the Home Civil Service decided to re-introduce the traditional full written examination plus Board interview, to be called Method I, in parallel with the new extended interview procedure, to be called Method II. The Foreign Office were sufficiently convinced of the merits of the new procedure for this to continue as the sole method of entry to senior posts in the Foreign Service' (pp.150-151).

Method I continued in operation for some time, but numbers of candidates opting to be assessed in this way progressively declined until it was dropped in 1969. CSSB, however, has continued up to the present day. Its original function, i.e. assessment of candidates for administrative and Foreign Office appointments, is still its principal function though it has widened its brief considerably to include, for example, assessment for the Tax Inspectorate and Hong Kong Civil Service.

The CSSB methods of the Reconstruction period are described by Vernon (1950) and Wilson (1948). Briefly, candidates were assessed in groups of seven at a residential centre over a period of 48 hours by two administrative civil servants (in the roles of 'Chairman' and 'Observer') and a psychologist. Assessment consisted of a battery of cognitive tests, questionnaires on interests and leisure pursuits, projective tests, peer nominations, interviews of each candidate by each assessor, and 'practical exercises' about which Vernon (1950) writes:

'All three members listened to the first session - a Group Discussion - and gave a preliminary grading on a five-point scale ... Subsequent exercises were designed to resemble the work of a higher Civil Servant. They included sitting on a Committee, writing an Appreciation of a dossier, and the

exposition and handling of a Problem in Commit candidate also gave a Short Talk on a subject of choice, and there was a Second Group Discussic sessions were attended by the Chairman and Observer, of them by the Psychologist, who gave new gradings a Although the staff attempted to assess each separately, their gradings were inevitably infl earlier observations; i.e. they tended to become sum the general suitability of the candidates in the li the available evidence' (p.76)

Finally, following consultation, each assessor awarded based on judgement of all the evidence. The overall was normally an average of these three grades (Vernor CSSB had discretion to fail the weaker candidates (Anst but for most the CSSB mark and a report went 'as a con recommendation to the Final Selection Board' (Wils pp.209-10).

The interviews and practical exercises which formed the l the early CSSB procedure remain substantially the sam giving CSSB a distinctive character. The description in 6 of the assessment procedures used by the Home Office, w CSSB derived and very close to current CSSB method demonstrate the high degree of continuity.

The first major study of CSSB validity was reported by (1950). He found that CSSB overall grade correlate (N=106) with a rating of potential after a two week t course. In later follow-ups using factor analysis general performance ratings as criteria, he found that CSS correlated 0.254 (0.509 after correction for restric range; N=147) with performance of administrators after one

at 0.215 (0.499 after correction; N=123) with performance of Foreign Office staff after one year, and at 0.164 (0.505 after correction; N=202) with performance of administrators after two years. There was considerable overlap in the samples from which these coefficients were derived, so the findings are not independent. Vernon also calculated validity coefficients for the Final Selection Board (FSB) grades (informed by CSSB's findings) which were in most cases higher than the corresponding CSSB coefficients. For example the correlation of FSB grade and performance of administrators after two years was 0.287 (0.563 after correction; N=202).

Anstey (1966,1977) continued Vernon's (1950) follow-up of those who had entered the Civil Service by way of CSSB during the Reconstruction period. There were important methodological differences however. Anstey used rank attained as the criterion whereas Vernon used a specially designed performance appraisal form. (Differences in rank would not have emerged by that time.) Also Anstey calculated the predictive validity of the FSB mark, not the CSSB mark, which makes it impossible to judge the extent of CSSB's contribution. Anstey (1966) found a correlation of FSB mark and rank attained of 0.305 (0.596 after correction for range restriction; N=350) for administrators with 16-18 years' service. Anstey (1977) continued this follow-up to 30 years' service, when the administrators 'were nearing the end of their careers; it could reasonably be assumed that their rank reached constituted a fair criterion of the progress made by the individual' (p. 152).

He found a correlation of 0.354 (0.660 after correction; N=301) between FSB mark and rank attained.

In addition to these studies Anstey gave evidence to the the Fulton inquiry into the Civil Service (Fulton, 1968; pp.106-153) concerning the utility of the CSSB procedures in comparison with the 'Method I' system of selection. He demonstrated that the performance and potential of administrators who had entered by CSSB between 1948 and 1963 was on average superior to that of those who had entered via Method I. However, as Anstey pointed out, CSSB competitions were held earlier in the year than Method I competitions and many failed CSSB candidates went on to sit Method I. Thus the findings might be explained in terms of differences in the quality of the candidate groups. Another Fulton study, by Dr. J.F. Pickering, which followed up failed CSSB and Method I candidates in terms of later salary and career success outside the Civil Service, adds weight to this interpretation in indicating that the CSSB candidates were the more successful group (Fulton, 1968; pp. 37-98).

Despite some shortcomings in research methodology, in making an overall assessment of CSSB's contribution to the historical development of ACs it would seem fair to say that it was one of the first ACs for which good evidence of long term predictive validity was produced. Also it was the first major non-military AC application, and the first to be designed on the basis of a systematic job analysis. Now nearly 40 years old, it must still rank as one of the most thorough and professional assessment

systems operated anywhere.

Assessment at AT&T

In Britain today virtually all branches of central government make use of some form of WOSB derived assessment in selecting those destined for middle and senior managerial levels, examples being CSSB, the Army's Regular Commissions Board, and the Admiralty Interview Board. However, where ACs are used in industry and commerce they generally derive from applications in the USA. Dulewicz, Fletcher and Wood (1983) point out that many British users of ACs are UK subsidiaries of US multinationals such as IBM and Rank Xerox, and this may in part explain the trend. The AC programme designed and evaluated by Douglas Bray and others as part of American Telephone and Telegraph's (AT&T) 'Management Progress Study' was the forerunner of ACs as practised in the USA today.

The Management Progress Study began in 1956 as 'a longitudinal study of the development of young men in a business management environment' (Bray and Grant, 1966; p.1). Assessment per se was only one of several research methods used. An important feature of the programme was the lack of:

> 'contamination by the assessment results of the subsequent criterion data. Along with all other information collected on the 422 subjects of the Study, the assessment data are being held in strict confidence. Thus the judgments of the assessment staff have had no influence on the careers of the men being studied' (Bray and Grant, 1966; p.1).

)n the basis of a literature review and the judgemen
:xperienced personnel staff within the company, a list
:haracteristics was derived which formed the basis for the
)f assessment techniques, and was used for rating candi
erformance.

andidates spent 3.5 days at the AC. Assessment methods wer
wo hour interview focusing on personal development, attit
alues, interests, interpersonal relationships etc.;
In-Basket' giving the candidate 3 hours to deal with a s
aterials a telephone manager might find in his in-tray fol
' an interview about his performance; a 'Manufacturing Pro
ıich was a group simulation of a small-business enterpris
roup discussion' in which candidates were assigned role
d to argue out a promotion decision; and a variety of obje
sts, projective tests and questionnaires.

oups of 12 subjects were assessed by teams of about 9 s
ray and Grant, 1966) who were mostly psychologists (Thor
d Byham, 1982). Interviews were on a one-to-one basis, and
)up exercises subjects were assessed in teams of 6 by 2 s
ıbers. Written reports were prepared for each assess
hod, and at the end of the 3.5 days all the staff assem
reviewed the results. Following presentation of the repo
h staff member independently rated each subject on the
racteristics. Each characteristic was then reviewed,
gements about managerial potential were made.

Evidence as to the predictive validity of the AT&T procedures comes from two follow-ups of an all-male sample of young managers assessed between 1956 and 1960 (Bray and Grant, 1966; Bray, Campbell and Grant, 1974). Approximately two thirds were college graduates assessed soon after employment; the remaining third had been employed initially in non-managerial positions and had advanced into management relatively early in their careers. By 1965 42% of the 103 predicted to 'make middle management' had done so compared with 7% of the 166 not predicted to do so. For 7 sub-samples, median correlation of AC overall rating with salary in 1965 was 0.43. Bray, Campbell and Grant (1974) continued the follow-up for the college graduates only, using the criterion of management level at re-assessment which was eight years after the initial assessment in each case. Of those remaining with AT&T, 39(64%) of the 61 predicted to reach middle management had done so compared with 20(32%) of those not predicted to do so.

One of the most important aspects of the AT&T study was that the AC was for research purposes only. The study provided a clear demonstration of the potential utility of a sophisticated extended assessment procedure for the identification of at least middle management potential. It is worth noting also that the college graduates had already been fairly intensively screened using traditional selection methods. To take the example of Michigan Bell (one of the AT&T companies) Bray, Campbell and Grant (1974) report that in 1956 offers were made to 79 out of

608 job applicants as a result of the college re programme. The recruitment procedures used were b application forms and interviews sometimes supplemented and opinions provided by colleges. Also the re judgement was exercised against a background of cons in-house research which had identified academic performa 'achievement' in wider campus activities as valid predi later success in the company. In a sense, then, t research can be seen as providing evidence of the diff validity of ACs over more traditional assessment method were not very different from those used by many employers

Another important feature of the AT&T AC was its u assessment rather than selection. The Management Progres in reality a set of interlinking studies with different as about the development of managers both as employees ndividuals. Within an interactive framework, career p as seen as a function both of indi bilities/characteristics and of organizational factors. pened the door to a broader view of the role of AC therto; the possibilities of ACs for areas such as traini reer development came more into focus. It should be p t, however, that Bray, Campbell and Grant (1974) reached ssimistic conclusions about the development of mana ills, e.g.:

> 'At least as far as the average recruit is concerned, eight years of management experience had not improved his administrative skills' (p. 133)
>
> 'The assessors [at reassessment] obviously saw the average recruit as performing less effectively interpersonally than he had done eight years previously' (p.134)

This may explain why Bray, Campbell and Grant (1974) focused exclusively on management selection in their conclusions as to the organizational implications of ACs, while pointing out that selection should go hand in hand with manpower planning and careful job placement to avoid the deleterious effects of recruits becoming de-motivated.

As already indicated, the AT&T research had considerable impact on the subsequent shape of ACs in the US. The process by which this occurred is described by Crooks (1977):

> 'After AT&T published favorable research results, visitors from other companies flocked to AT&T to observe their assessment centers and to ask for copies of their exercises, rating forms, manuals, and whatever else was available. Even today, in observing programs from company to company, the basic AT&T format ... is readily discernible' (p.71).

In particular it is situational tests that give ACs their individual characters (i.e. these are what people seem to remember), and those developed by AT&T have been widely adopted since.

This description of AT&T procedures brings the history of ACs more or less up to date. In the next chapter AC content will be considered in more detail, though it is fair to say that most present day ACs are primarily extensions of the WOSB, CSSB, and particularly AT&T models.

Defining 'Assessment Centre'

At this point it seems appropriate to ask how ACs differ from other sorts of assessment systems. This is a difficult question to answer, and in the final analysis it is unlikely that any single definition would satisfy all those involved in the field. One attempt at an agreed definition has been made by practitioners in the United States (Task Force on Assessment Center Standards, 1980). According to this definition:

> 'The following are the essential elements which are necessary for a process to be considered an assessment center.
>
> Multiple assessment techniques must be used. At least one of these techniques must be a simulation. A simulation is an exercise or technique designed to elicit behaviors related to dimensions of performance on the job requiring the participants to respond behaviorally to situational stimuli. The stimuli present in a simulation parallel or resemble stimuli in the work situation ...
>
> Multiple assessors must be used. These assessors must receive thorough training prior to participation in a center.
>
> Judgments resulting in an outcome (i.e., recommendation for promotion, specific training or development) must be based on pooling information from assessors and techniques.
>
> An overall evaluation of behavior must be made by the assessors at a separate time from observation of behavior during the exercises.

> Simulation exercises are used. These exercises are developed to tap a variety of predetermined behaviors and have been pretested prior to use to ensure that the techniques provide reliable, objective and relevant behavioral information for the organization in question. The simulations must be job-related.
>
> The dimensions, attributes, characteristics, qualities, skills, abilities or knowledge evaluated by the assessment center are determined by an analysis of relevant job behaviors.
>
> The techniques used in the assessment center are designed to provide information which is used in evaluating the dimensions, attributes or qualities previously determined' (pp.35-36)

A problem with this definition is that it does not distinguish between essential and desirable attributes of ACs. The essential descriptive features referred to in the above definition would appear to be: use of multiple assessment techniques; use of simulations; use of multiple assessors; and pooling of information from assessors and techniques. Also the requirement that overall evaluation of behaviour should be separate from observation seems important in distinguishing ACs from sequential selection systems. However, training of assessors, pre-testing of simulations and the use of job analysis, while highly desirable, would not appear to be defining characteristics.

Taking stock of the 'essential' parts of the Task Force definition one might define an AC as **an assessment procedure involving multiple assessors, multiple techniques including at least one simulation, and suspension of overall judgement until the end of the procedure when a pooling of information takes place.** This replaces the working definition given at the beginning of the chapter. The word 'simulation' is used here and throughout the rest of the thesis in the broad sense of the Task

Force definition.

The writer does not suggest that this definition sh universally adopted; its point is merely to help est frame of reference for the thesis. In reality, th definition probably does not matter very much. It is ir to recognise that ACs have no intrinsic validity and tha is nothing unique about them. Each AC is only as goo relevance and validity of its constituent techniques specific organizational context, and the quality of its d making processes .

)ther terms for 'Assessment Centre'

n conclusion it is worth looking briefly at different hich are sometimes used instead of 'Assessment Centre'. nd Stewart (1981) argue that the use of this term, which S origin, is inappropriate:

'in the earliest programmes in the States, the proc happened in a separate building, which was known assessment centre; by a transfer of epithets, the name attached to the actual programme of events, so irrespective of whether it took place in a specially ded building the programme was known as an assessment centr prefer to talk of assessment programmes, first because th likely to be more accurate even for a one day programm secondly because an assessment programme could if nec last 18 months (using planned assignments, for instance the phrase then becomes ridiculous' (pp.62-63)

While this is a reasonable argument, the proposed alternative 'assessment programme' might be used to describe any multiple assessment procedure, e.g. a test battery plus interview, and does not hint at the distinguishing AC characteristics of multiple assessors and simulations. In any case the term 'Assessment Centre' has won broad acceptance and, however illogical, seems likely to remain in use.

Another term used to describe some British ACs, particularly CSSB and its derivatives, is 'Extended Interview'. The AC which is the subject of this thesis is referred to by this term. To some extent the term may reflect the large part interviews play in ACs so labelled.

Chapter 2

ASSESSMENT CENTRES - CONTENT, STRUCTURE AND DECISION MAKING

From the definition of ACs given in the last chapter it is clear that there are as many potential AC techniques as there are assessment methods. However, certain types of assessment method are used much more frequently than others. In particular ACs are characterized by the use of simulations, the only specific type of technique referred to in the Task Force definition of the AC. In the next section a look will be taken at the sorts of simulations used in ACs, and in the section after that other common types of AC technique will be described.

Simulations

The key AC technique is the simulation. A great variety of types of simulation have been used, though the majority of ACs seem to adhere to standard formats, such as the in-basket and the management game. Crooks (1977) has produced quite a useful classification of AC simulations, but from a US standpoint. The following classification attempts to be more representative of developments on both sides of the Atlantic:

Administrative Problems

The emphasis here is on the assessment of competence in h administrative-type tasks representative of a ma day-to-day work-load. The best known format is probab In-Basket or In-Tray exercise, illustrated by Stewart and (1981):

> 'Here the participants work individually, against pressure. They are given a brief which tells them th have just moved into a new job and that this is their morning in it; it is a Saturday morning and they have co the office for an hour and a half before leaving to c plane to somewhere where they will not be able to c their office work. There is no one else in the buildi the switchboard is shut down; they have a full in-tray i of them which they have to cope with as best they can time provided, knowing that they will not be back unt Wednesday. They are to attach notes to each item indi the action they are taking' (p.104).

he In-Basket is usually followed by an interview in whic ndividual's handling of the material is discussed and h nderstanding of problems in the In-Basket explored. Ratir hen made on relevant dimensions, e.g. planning, decisiv se of delegation (Crooks, 1977).

nother example of this category of simulation is the ervice Selection Board's Drafting Test in which the asses equired to 'draft an answer to a difficult letter from xample, an influential member of the public, which cal ict, judgement and a reasonable command of language' ervice Commission, 1977; p.10)

Complex Problems

This type of exercise is designed to assess capacity to deal with broad policy-type issues such as might be encountered in some managerial positions, and to give cogent expression to decisions/recommendations either in writing or orally. A good example is the Civil Service Selection Board's Written Appreciation in which the candidate is confronted 'with a file of papers - including committee minutes, imaginary excerpts from Hansard, statistical analyses and letters from members of the public - all describing a problem of the kind many civil servants might have to tackle in real life. Past themes have included the development of water resources, the siting and choice of fuel for a large power station ... and projects for overseas aid' (Civil Service Commission, 1977; p.9). The candidate is required to look at three or four possible solutions, to choose one, and to justify his/her choice.

A similar test is Stewart and Stewart's (1981) White Paper Exercise in which the 'participants are given a white paper, green paper or official publication of some kind and asked to write a report on it, detailing its implications for the company as a whole or perhaps for a specified part of the company' (p.113).

/ariants on this theme are Presentation and Fact F
:xercises. In Presentation Exercises assessees are requi
:ngage in the sort of task described above and to
›resentations before groups of peers, their superiors,
utside groups' (Crooks, 1977; p.75). In Fact Finding Exe
he 'assessee collects data on a problem verbally by a
uestions of a resource person then has to present the pr
nd his or her conclusions either verbally (during or after
:ssion he or she submits to questioning) or in writing' (Cr
977; p.76).

:aderless Group Exercises

iese fall into two main categories, the Leaderless (
scussion (LGD) and the Management Game (or 'Business Gam
lanufacturing Game'). LGDs may be classified as ha
n-assigned roles and assigned roles (Crooks, 1977). In
mplest, least structured form, a non-assigned role LGD 'm
tting the candidates down together, giving them a topi
scuss, and leaving them to get on with it while the asses
t outside the group and make notes on the ensuing interact
·e often, though, a greater degree of task structure
›vided.' (Fletcher, 1982; p.43). Crooks (1977) gives
ustration of the more structured type of exercise:

the group of participants ... is handed short case studies
anagement problems. As consultants, they are asked to reso
he problems and present a written recommendation. Probl
ealing with supervision, business judgment, conflicts betw
epartments and employees, job dissatisfaction, and setting
riorities among alternative actions are examples, depend
.. on important factors in job performance at the tar

level. Both quality of thinking and group process variables can be observed' (pp.74-75).

A key element of assigned role exercises is competition between candidates. Participants 'have individual aims which they are expected to pursue with vigour and they share a group objective which will be at variance with all the individual objectives present except, ultimately, one' (Stewart and Stewart, 1981; p.112). We refer again to Crooks (1977) for an illustration:

> 'Each of six assessees in a group is given a description of a fictitious subordinate he or she is recommending for promotion. The descriptions are formulated so that the candidates are about equally qualified. The assessees study their candidate descriptions and each is then allowed five minutes to make a pitch for the candidate the assessee is sponsoring. After all six assessees are heard, a period of free discussion is followed by a rank-ordering of the job candidates by the assessees from most deserving to least deserving. Assessors observing the group ... judge the assessees on ability to sell their candidates and what they have done to aid the group in reaching a decision ... [Individual] skills and group process variables can be observed' (p.75)

Management Games differ from LGDs in that participants:

> 'are formed into teams, each one representing either a firm or some particular section of the firm. They are then usually required to operate within a given market environment in such a way as to achieve maximum efficiency according to some criterion (often net profit). The team may operate in co-operation or, especially where they "represent" different companies, competitively. The organizer of the game generally needs access to a computer, since the decisions taken by each group of candidates have to be fed into the computer model and the results of those decisions on production, sales figures, and so on are given back to the syndicates as quickly as possible. The simulation may last some hours, with candidate teams continually analysing the data produced by the computer, taking decisions, looking at the outcome, and revising their strategy ... Assessors observe each team in operation and scrutinize the decisions made and their outcomes. The exercise can throw light on leadership ability, numeracy, business sense, and the capacity to organize and get on with other

people. If it is used with feedback from senior manag
afterwards, it can also clearly be a stimulating learn
experience for the participants.' (Fletcher, 1982; p.45).

signed Leader Group Exercises

.s type of exercise is characteristic of British-style A
: emphasis is on assessment of ability to take charge o
)up faced with finding a solution to some specified proble
t ACs run by the British armed services use physical/practic
nand tasks of the War Office Selection Board type.
iralty Interview Boards, for example, 'each member of a gro
'our or five candidates takes charge of the group and attemp
solve a physical problem (getting the group, with or without
len, across an obstacle)' (Jones, 1981; p.81). A sedentar
valent of this is the Committee Exercise developed by th
l Service Selection Board: each candidate in turn takes th
of Chairman of a committee meeting called to resolve
ific problem, e.g. an industrial relations or personne
.em, the placing of a contract, etc. When a candidate is i
hair he or she is judged on his/her ability 'to get the
em across to the group and to give them a lead, to run the
ssion, absorb its useful points and finally to present a
ion acceptable, it is hoped, to the committee as a whole'
l Service Commission, 1977; p.9). Candidates are also
sed on their contributions to discussion when others are
; the chair.

Dyadic Situations

These are two-person role-plays designed to represent situations in a target job. The most common type of exercise in this category is the Interview Simulation in which, for example, 'the assessee (playing the role of a manager), interviews a capable but troublesome employee, a standardized role played by a staff member' (McCormick and Ilgen, 1980; p.465). Another example is the 'situational testing' sometimes used in ACs for entry-level police selection in the US. An example is provided by Filer (1979):

> 'In the Ft. Collins and Colorado State University assessment centers the general procedure was for the applicant to be brought to the testing room and given a gun belt to wear. He or she was then briefly instructed in handcuffing and frisk procedure and was handed a card on which minimal instructions were typed. An example would be, "You are driving on patrol in the downtown area when you notice a young man ... prying at a parking meter with a screwdriver. It is 4.45 p.m. Do your duty."
>
> ... the applicant ... entered the room in which an irate citizen was kicking and prying at the parking meter. The "theme" that the confederate followed ... was that he had been looking for a parking space for 15 minutes and that, when he finally found one, his nickel jammed in the meter. The confederate was in a "big hurry," ...' (p.224).

Other Assessment Centre Techniques

Interviews of various kinds are commonly used in British ACs, though less frequently in the US. Interviews may focus on work-related areas e.g. work history, professional knowledge, career expectations, values and goals etc. and/or on more personal areas such as emotional stability, self-insight, and

:lationships. Another approach is to use the interview
:st, e.g. of an individual's ability to argue a case f
irticular point-of-view under cross-examination.

whole range of psychometric tests are used in ACs to a:
neral abilities, specific aptitudes (e.g. numerical), are:
owledge, personality traits, needs and values,
lf-description and/or biographical questionnaires may al:
cluded, often as a preparation for interview. In addition
sessments (or 'sociometric ratings') may be used. These
nge from simple nominations where, for example, the asse
ght be asked which of his/her colleagues would make the
i second best 'Senior Officers', to full sets of ratings
'formance in group simulations.

ıventional Wisdom on Assessment Centres

)ugh it would be difficult to produce a general statement
)od assessment centre practice' on which everyone would ag
tain important assumptions run through most of the writing
subject. The 'conventional wisdom' is that one should
analysis to determine job-relevant **assessment dimensi**
ch provide the basis for the **design and selection**
essment techniques. When the AC is in operation, assess
e assessees' performance over the set of dimensions and m
'all judgements/recommendations on the basis
lnical'/subjective combination of the resulting information

In the next section methods of job analysis will be compared, and in the section after that the use of assessment dimensions will be critically examined. In the final section the 'clinical' method of pooling AC information will be compared with mechanical/statistical alternatives.

Job Analysis

Statements to the effect that systematic job analysis (JA) is essential to programme design are routine in the AC literature. JA is seen as a prerequisite to the choice and design of tests and exercises and to the choice of dimensions by which to assess behaviour (where dimensions are used). In the particular case of AC simulations, content validity is frequently claimed on the grounds that the simulations parallel situations in specified target jobs; clearly JA is of crucial importance here in identifying the relevant job content domains and their relationship to overall job effectiveness.

In view of the importance attached to it, it is surprising that the details of JA methods used in AC design are frequently not at all clear in published accounts. Here are two typical descriptions:

> 'Intensive discussions with key administrative personnel identified the characteristics thought to be indicative of successful performance in both the mental ability and personality areas' (Ginsburg and Silverman, 1972; p.664).

'The initial steps in the job analysis procedures involv interviewing of a representative sample of members wit given target position. In addition to the interviews, a of job observations were performed in order to clarify job elements identified through the interview p (Magaldi, Mendoza, Stafford and Frank, 1984; p.11).

Why so little attention should be paid to the description is hard to understand. Perhaps it is seen as a rout: standard operation which requires no detailed comment; as (1977) points out, JA is often discussed in such a way suggest that 'any fool can do it' (p.167). In reality m re not at a level of standardization or sophisticat ustify this sort of complacency.

here are, nevertheless, a large number of systematic appr o JA. Two techniques frequently used in AC design (Je 977) are Hemphill's (1960) Executive Position Descr uestionnaire (EPDQ) and Flanagan's (1949,1954) Critical In echnique. Stewart and Stewart's (1981) use of Repertory lus questionnaires is another technique which has stim ecent interest (e.g. Dulewicz, Fletcher and Wood, 1 scussion of these three techniques should serve to brir me of the advantages and disadvantages inherent in diff pes of approach.

mphill's EPDQ is a standard questionnaire which aims mprehensive description of executive jobs in terms of comp b behaviours. Hemphill's original research form of the ntained 575 items or 'positions elements' and was admini 93 executives working in a variety of business functi

'Each respondent utilized a seven-point response scale to describe his position in terms of the degree to which the element was a "part of the position"' (Hemphill, 1960; p.xiii). Factor analysis of the questionnaire answers identified 10 orthogonal factors. Subsequent researchers have produced various modified questionnaires and factor structures for the EPDQ (reviewed by Prien and Ronan, 1971).

The main advantage of this type of approach is that it can provide a standardized, systematic and comprehensive way of analysing the content of any individual job, and of assessing the extent of similarity/overlap among different jobs. The main disadvantages, as identified by Hemphill (1960), are the lack of specificity of standard questionnaires, and their static nature:

> 'It has been shown that groups of executive positions have common denominators which can be measured as dimensions of the positions. Undoubtedly, these common dimensions do not completely cover any particular position. There will remain parts of a position that are outside the range of the dimensions and that may be relatively unique to the particular position, the particular company, or the particular business situation at a given point in time ... [Positions] will change their characteristics from time to time, both within and without the framework provided by the ... dimensions' (p.63).

The Critical Incidents Technique (CIT), the principles of which were first outlined by Flanagan (1949,1954), is a very different type of approach. It produces selective description of job behaviours using open-ended questioning. The criterion for description is criticality, i.e. effect on overall success of job performance. One application of CIT is described by Dunnette (1976):

'Essentially, the methodology involves a series of four to
/orkshop sessions of about two or three hours each with pers
/ho are very familiar with the job being studied. The prim
)urpose of these sessions is to elicit stories or anecdo
iescribing critical incidents that the participants h
)bserved ... Participants actually write down their stories
'orms ... [on which more details of the incident
'equested. A] diligent effort is made to avoid obtain
mployee attributes or traits in favor of obtaining detai
iescriptions of job behaviors - descriptions which, because
he focus on eliciting "successful" and "unsuccessf
ehaviors, form the basis for categories reflecting
ehavioral requirements - desirable and undesirable behavior
f the jobs being studied ...

ext steps involve editing slightly the incidents (be:
areful to retain the essence of each), forming prelimin
erformance dimensions, and utilizing later workshop sessi
o work out final specifications and definitions for
ategories which cover sufficiently the total range
ehavioral requirements for the job or jobs being studi
pp.490-492).

dvantage of CIT is flexibility; its applicability is
trained by a predetermined set of job element
dvantages are that: much depends on the skill of t
-gatherer in eliciting information; interpretation rests
ent analysis; and, because the approach is selective, the
a danger that important elements might be overlooked (e.
nce graduates in positions involving routine but essenti
utational work might not identify this work as critical
success).

art and Stewart's (1981) Repertory Grid plus questionnai
nique represents something of a cross between open-ended a
structured approaches to JA. The focus is on job occupant
'effective' and 'ineffective' workers (normally managers) ar
ared. The process is in two stages. First, the clinica

Repertory Grid technique has been adapted for use in interviews with people familiar with the target job(s). The interviewer requests examples of tasks/behaviours (e.g. 'something you do which is very important', 'something you do frequently', etc.) or job occupants (e.g. 'someone you know who is not performing well', 'someone who is happy in his work' etc.) which are written down on cards. Sets of 9 cards (e.g representing 9 job occupants) are used to elicit constructs in the traditional way, i.e. presenting 3 cards and asking how two might be similar but dissimilar from the third.

In the second stage, constructs thus derived are used in workshop 'brainstorming' sessions to generate a number of bipolar (5-point scale) items in performance questionnaire format. Stewart and Stewart's example is the construct 'done with people - paperwork' from which scales such as 'he is better with people - he is better with paperwork' and 'he judges others on their ability to deal with people - he judges others on their ability to deal with paperwork and so on' might be derived (p.92).

> 'We draw up a questionnaire containing between 80 to 120 such items, trying to cover all appropriate areas. Then we issue this questionnaire to **second line managers** (managers of the position in question) and ask them to think of their most effective subordinate first line manager and to fill in the questionnaire with this manager in mind, using the five point scale to indicate degrees of strength or frequency of the behaviour ...
>
> ... when all these first questionnaires have been returned, we send out a second batch, identical with the first, except that this time we ask them to bear in mind the most ineffective first line manager subordinate they have and describe him' (Stewart and Stewart, 1981; p.79; bold substituted for italics).

this the basic procedure is to determine which items
icantly discriminate between 'effective' and 'ineffective'
rs and then to group these items (content analysis) in
ion format to give a person-specification. The description
h dimension is a composite of statements drawn from the
nt items, for example:

king in groups Works better in groups than alone. Invites those who need to attend. Most effective as chairman. not insist on rank/seniority prevailing. States his tion openly; would rather co-operate than compete.' wart and Stewart, 1981; p.83; bold substituted for ics).

t and Stewart's technique shares an advantage and a antage of CIT in that it is a highly adaptable/non-specific que but the inherent selectivity of description means that ant job elements could be overlooked. Perhaps another antage of the technique is its focus on job occupants. y JA technique, there is a danger that one is initiating a erpetuating process. McLeod (1982) points to the essential arity of the job analysis, person specification, selection ire, selection-validation cycle. 'Maybe what job-analysis one is what one **wants** to keep the organisation going, and alid selection goes on to provide it; but maybe it fails l one what one **needs** to make the organisation develop' l, 1982; p.13). This danger may be accentuated where JA on job occupants. JA techniques which stay close to nd task related behaviours may exclude some correlates of uccess' which are not specifically task related, for

example, characteristics reflecting identification with organizational policy or culture. In Stewart and Stewart's method there seems to be no attempt to make this sort of discrimination. The aim is to identify typical characteristics of effective job occupants. But questions remain as to whether those characteristics are: intrinsically necessary for effective job performance; indicators of the degree to which the job occupant conforms to prevailing norms/expectations; or by-products of the job occupant's self-perception of his/her success or lack of it. While it is clearly important to know about aspects of a job which are not specifically task related, it seems a mistake to confound different types of information. This is especially true in the context of organizational change. For example, changes in organizational policy might be partly effected through a change in recruitment policy: the more global the picture of job performance the less informed will be decisions as to the sorts of recruits required.

Taking stock of the JA techniques discussed, it seems clear that the advantages and disadvantages of different techniques are to a large extent complementary. In particular, some techniques are more suited to the purpose of detailed and comprehensive job description, while others may be useful in identifying the relatively unique and/or critical demands made by a particular job. An exclusive focus on characteristics of job occupants, however, may be undesirable. Anything approaching a complete analysis of any given job will probably require the use of

:tiple methods, as is frequently recommended (e.g. Pri
'7).

essment Dimensions

t ACs today make use of ratings of assessees' behaviour
ensions thought to be relevant to target job performanc
e Home Office Police ACs to be described in chapter 6 do r
olve the use of such ratings, and in this respect a
pical). As an example, the list of dimensions used in an
select San Fransisco police captains (Hurley et al., 198
: communication skills; problem-solving ability; planni
.ity; emotional control; interpersonal skills; supervisc
ls; organizational skills; and public relations skills. Th
about as short a list as one normally finds; Crooks (197
mates that numbers of dimensions used in ACs range 'from
as seven or eight to 26 or more' (p.72). Finkle (197
ts out that such variation 'suggests either fundamenta
erences in the choice of characteristics, in the semanti
ls of specificity, or in both' (p.871). While it is clea
ACs do to some extent attempt to assess differen
icteristics, the general similarities are also very evident
ling (1977), for example, compared dimensions used in si
Two aspects of performance - 'oral communications' an
rpersonal influence' - were common to all six ACs while
er six aspects - 'decision making, planning, organizing,
energy, personal likeability/acceptability, stress

tolerance' (p.192) - were included in four ACs. These ACs, like most, were designed for managerial assessment, and one reason for similarity may be, as Byham (1970) suggests, that certain key aspects of performance are important to many organizations. Another factor might be the perceptual set of those involved in formulating dimensions. Relatively unstructured job analysis procedures, illustrated in the earlier quote from Ginsburg and Silverman (1972), would seem to offer considerable scope for expectations to influence outcomes.

One might ask what the point of job analysis is when the result so often seems to be the derivation of predictable lists of dimensions which are then used as the rationale for the selection of fairly standard sets of 'off-the-shelf' tests and exercises. It was noted earlier that the amount of attention paid to job analysis by AC practitioners often does not appear to match its presumed importance. Paradoxically it may be the very predictability of job analysis output which is to blame for this. In this writer's view, predictability should be taken as an indication of a need for scrutiny of the way in which job analysis information is obtained and utilized, rather than as a rationale for downgrading its importance.

Assessment dimensions, once derived, are used as the basis for choosing appropriate AC tests/exercises though, as Finkle (1976) indicates, there is a certain amount of circularity in the process; the choice of variables is 'clearly a function of the methods and techniques of generating data input. For example,

In-Basket produces input particularly appropriate f
gments of administrative skill ...' (p.873).

the AC is in operation the usual procedure is to rate ea
ssee's performance in each subjectively marked exercise usi
bset of relevant dimensions. So, for example, from a list
dimensions Dulewicz, Fletcher and Wood (1983) used a 'Lett
ing Exercise' to assess 'written communication', 'socia
l' and 'relations with subordinates'. At the end of the A
ngs and other measures are normally combined, usually in
ical/judgemental way, to give an assessee profile across th
e set of dimensions. This is then used as the basis fo
ll rating and decision making.

process of dimension derivation, test/exercise selection,
nsion rating, information combination and decision maki
n has been described appears, on the face of things, t
esent a logical and consistent approach to assessment
ver, certain important assumptions are involved. I
nsions are viewed as traits and tests/exercises as methods o
measurement, each dimension rating becomes, in Campbell an
's (1959) terms, a **trait-method unit**, 'a union of
cular trait content with measurement procedures not specifi
at content' (p.81). In ACs the 'trait content' of dimensio
gs should be sufficient to satisfy at least two practical
ds. Firstly, ratings of given dimensions in different
ises should show consistency such that reliable overall
sments may be made of each individual's performance on each

dimension. This corresponds to Campbell and Fiske's requirement that traits should (on construct validity grounds) have **convergent validity**; independent measures of the same trait/dimension should converge/correlate closely. Secondly, dimension ratings should be relatively independent of the methods/exercises from which they were obtained; the final dimension profile of any individual assessee should provide a basis for fine discriminations among aspects of his/her overall performance, and this will not be the case if dimension ratings are primarily global measures of performance on specific exercises. One way in which exercise independence may be demonstrated empirically is by showing that ratings of one dimension agree more closely with ratings of the same dimension in other exercises than with ratings of other dimensions in the same exercise. This is one of a set of tests by which the **discriminant validity** of a trait may be judged (Campbell and Fiske, 1959); for a trait to have discriminant validity it must not correlate too highly with other traits from which it was intended to differ.

The reliability and convergent validity of AC dimension ratings may be statistically assessed with coefficient alpha, a measure of 'internal consistency' which has traditionally been used in objective test development. Drawing an analogy with test development, each exercise-specific rating is viewed as a single test item and a hypothetical composite score on the dimension concerned (formed by adding together exercise-specific ratings)

ed as the test; coefficient alpha indicates the
ty of the test/dimension. Two studies which have
this approach are those by Hinrichs and Haanpera (1976)
tt and Dreher (1984). Hinrichs and Haanpera assembled
369 participants in similar ACs run by one company in
ferent countries. Performance in six simulations was
oss a total of 14 dimensions. Coefficient alpha ranged
of -0.04 ('administrative ability') to a high of 0.73
mmunications') with a mean of 0.49. Sackett and Dreher
a reanalysis of earlier data (Sackett and Dreher, 1982)
coefficient alpha for seven dimensions rated in six
s (N=86). Alphas ranged from -0.06 ('written
ion') to 0.65 ('oral comunication'). Alphas of this
gnitude are well below those traditionally required for
paper ability tests (e.g. see Nunnally, 1970).

ings could, as Sackett and Dreher (1984) comment, be
support for measuring at least some of the intended
/constructs and 'by adding additional exercises these
y estimates could be increased even further' (p.189).
, however, that some dimensions show no potential for
assessment, and in practice the costs of assessing a
of dimensions more reliably might be prohibitive.
bly assessed dimensions would have convergent validity
cessarily discriminant validity. For example six
and reliable dimensions might measure the same

Sackett and Dreher (1984) report an additional finding which puts dimension reliability estimates in perspective. They computed 'alpha for exercises in the same way as for dimensions. The mean alpha for exercises was .90, indicating ... the predominance of situational variance over dimension variance' (p.189). So it seems likely that the reason for low reliability of trait ratings is their dependence on the exercises from which they are drawn.

More direct statistical tests of the method independence and discriminant validity of dimension ratings have used correlation, factor analysis and analysis of variance techniques. Sackett and Dreher (1982) showed that for ACs in three organizations within-exercise ratings for different dimensions correlated more highly than across-exercise ratings for specific dimensions. Similar findings from two previous studies (Archambeau, 1979; Neidig, Martin and Yates, 1979) were cited in support. Thus dimension ratings fail one of Campbell and Fiske's stated tests of discriminant validity.

Sackett and Dreher (1982), following Smith (1976), went on to argue that factor analysis provides a better way of examining such data 'since it is less susceptible to the effects of small fluctuations in the size of the correlation coefficients than the Campbell and Fiske approach is. In factor analytic terms, the question is very clear and to the point: Do the factors underlying these judgments represent dimensions or exercises?' (p.402). When Sackett and Dreher factor analyzed their dimension ratings the factors which emerged clearly represented exercises

'ather than dimensions for each of the three ACs studied.

'urnage and Muchinsky (1982) approached the same basic ques y subjecting data on 2056 AC candidates to an analys ariance. They extracted a main effect for 'per orresponding to convergent validity, a person-trait intera ffect corresponding to discriminant validity, and erson-situation interaction effect indicating the exter .tuational specificity of ratings. Large person rson-situation effects were found plus a person-trait ef ich, though statistically significant, was 'so weak as to actically nonexistent' (p.187). Conclusions were that igh degree of convergent validity and associated lack scriminant validity across traits indicated that assessees v aluated globally rather than differentially' (p.188). A ere was strong evidence for the situational specificity iaviour. It should be pointed out that the measure vergent validity in this study was an absolute rather tha ative one (i.e. testing the existence of the effect) and es little guide to the practical question of dimens: iability.

the results of these various analyses are pieced together, erent picture of AC dimension rating presents itself: t rcorrelations of ratings on specific dimensions acro cises tend to be insufficient for high reliability, thou some dimensions convergent validity, and hence a potenti reliable assessment, is indicated; this potential reliabili

reflects the extent to which dimensions relate to global evaluations of overall AC performance rather than to the constructs they purport to measure; AC dimension ratings are predominantly a joint function of the situations in which assessees are rated and of assessors' global evaluations of assessees; and on the basis of low convergent validity and negligible discriminant validity, AC dimensions cannot be described as construct valid. Given the available evidence one must conclude that the conventional rationale for using dimensions to rate AC performance is unsound.

In the light of the above it is appropriate to ask whether AC dimensions serve any useful purpose. Zedeck and Cascio (1984) argue for their continued use, at least in the short term:

> 'We suggest that the dimensions be given a secondary role and that assessment centers be driven by tasks and behaviors that are representative of the position for which one is interested in selecting or developing managers. This is not to suggest that dimensional considerations be ignored or eliminated. The overall rating in assessment centers is often cast in terms of "potential" or "likelihood" of success, but it can be viewed as an evaluation of success in the whole exercise process. Over time, these ratings have been demonstrated to be valid ... Dimension and exercise ratings and discussions may serve as a cognitive means for the assessor to structure observations ... If the global rating works, don't fix the process, but continue to research the process to get a better understanding of it.' (p.482).

This leads into consideration of various aspects of the validation process, which will be left till chapters 4 and 5. However a comment on Zedeck and Cascio's view seems appropriate at this point. It may be that dimension ratings provide a structure for observation and discussion, but the argument that

the use of invalid dimensions is justified by the validity
final global rating seems to be taking empiricism to an e
A point in favour of dimensions is that their use may enco
comprehensive sifting and weighing of evidence as a precur
global judgements. But, against this, consideration of n
constructs may sidetrack assessors from the main is:
evaluating overall performance. While a strong case can b
for imposing some sort of structure on the assessment pr
there seems no particular reason why this structure shc
rait based. An alternative might, for example, involve an
f performance in terms of contributions to spe
bjectives/sub-objectives etc., perhaps modelled hierarchi
oubtless there are other possibilities.

ecision Making: Clinical or Mechanical?

ie AC overall ratings (OARs) on which decisions tend to be
e generally arrived at in clinical/judgemental fashion.
ternative is some sort of mechanical/statistical combinati
formation. Which approach is chosen has implication:
ficiency and validity. Combining data by means of a formul
early less time consuming and hence less costly than a prc
judgement/discussion. However the real issue is the valid
ernal and/or external, of the mechanical approach. Appl
internal validity framework to the AC consensus pro
kett and Wilson (1982) showed that, for two ACs, one sir
lsion rule could be used to predict final dimension rati

from assessors' pre-discussion ratings with 94.5% and 96.5% accuracy. A cross-validated multiple correlation of 0.89 was obtained using final dimension ratings to predict OAR. In similar vein, Herriot and Wingrove (1983) reported a study of consensus decision making for 6,000 naval officer applicants from which it emerged that in 99% of cases where assessors were agreed on which side of the pass/fail line an applicant should be rated, no ammendment to the final decision was made. It was suggested that extensive discussions were unnecessary in cases of initial concordance.

It would appear, then, that mechanical processing of information can be used to streamline typical AC decision making without significantly affecting the decisions themselves. It is of interest, however, to consider what happens when clinical and mechanical decisions differ; is there any systematic difference in the quality of those decisions? In other words, how does the external/criterion-related validity of the two approaches compare? Surprisingly very few AC studies have looked at this question. Sawyer (1966) carried out a very thorough and comprehensive review of clinical vs. mechanical/statistical prediction in general. Comparisons were classified according to whether data had been collected clinically (i.e interview or other direct observation) or mechanically (e.g. objective test) or by both means. A clear finding from the 45 studies looked at was that 'the mechanical mode of combination [is] always equal or superior to the clinical mode; moreover, this is true whether

ta were collected clinically or mechanically' (p.192).

rlier review by Meehl (1954) had reached a similar conclusio

vyer makes an important point about the use of multi

gression as a basis for comparison with clinical judgement:

> Multiple Regression, more often than not in these studie
> mployed weights derived from the same sample on which t
> alidity was assessed; consequently, it overestimated t
> eplicable relation ... On the other hand, the vario
> onfigural methods mostly represent a priori combinations, a
> hus they avoid the substantial overfitting that results wh
> hese methods are applied a posteriori' (p.190).

Sawyer's review, mechanical prediction, even when based on

ori configurations such as weighting all scores equally

ged as equal to or better than clinical prediction overall.

AC studies which have made clinical-mechanical comparison

reported findings in line with Sawyer's conclusions

owick and McNamara (1969) found that AC OAR for 94 lower an

le managers correlated 0.37 with increase in management

onsibility. However this was bettered by multiple Rs base

pencil and paper tests alone (R=0.45), exercises alone

.39) and characteristics alone (R=0.41). The multiple R

ed using all three sorts of information was 0.62. Tziner

Dolan (1982) found that AC OAR correlated 0.38 with

rmance in training for 193 female officer trainees compared

a multiple R using the same assessment data of 0.47; scores

one verbal intelligence test correlated 0.39 with the

rion, and the multiple R derived from ratings on 5

ations was 0.36. Mitchel (1975) followed up three

sub-samples of AC participants (managers; total N=254) after 1, 3, and 5 years using salary growth criteria. Assessors' overall rating of 'potential' had an average validity of 0.22 compared with an average multiple R of 0.42 based on dimension ratings and scores on pencil and paper tests. However the average generalized correlation, that is the average correlation with criteria when regression equations were cross-validated across sub-samples, was 0.28. This demonstates the need for cross-validation when making clinical vs. multiple regression comparisons. But even after cross-validation, multiple regression was superior. Borman (1982) reported a study of an AC for 57 trainee US army recruiters. OAR correlated 0.38 and 0.27 with two training criteria. A mechanical composite (unit weighting ratings on each exercise and pooling ratings across exercises) correlated 0.48 and 0.35 respectively with the same criteria. Wingrove, Jones and Herriot (1985) have recently addressed a somewhat more specific issue in a study of an AC for Naval Officer Selection. OAR was calculated as a mean of assessors' individual post-discussion ratings. Wingrove et al. found that where mean pre- and post- discussion ratings differed (by at least one standard deviation), validities found for a composite training criterion did not differ (r=0.45, after correction for range restriction, for both pre- and post-discussion ratings; N=387). The implication is that simple mechanical pooling of ratings would not only save time but should also have no detrimental effect on AC validity.

ɔ although there has not been a great deal of research
echanical means of processing AC information, when AC rese
indings are set against the wider research background
ıdications are that mechanical processing could both stream
ecision making and increase (or at least not decrease)
lidity of the decisions themselves. A strong a priori case
made for the power of the clinical/judgemental approach (
ample see the Davies Report on the Civil Service Select
ard, 1969). But, given the available evidence, it seems t
nething more than cogent argument is required. In view of
gh cost of assessor time involved in typical processes
scussion and judgement it would seem that the onus should be
ɔse advocating the retention of such procedures to prod
dence to back their case.

Chapter 3 - THE CONTEXT OF ASSESSMENT CENTRES

In this chapter some of the wider issues surrounding the use of ACs will be considered. Whatever an AC's purpose - selection/promotion and/or placement and/or training/development - its efficient use will involve alignment with organizational objectives and integration with other parts of the personnel function. There is, however, a danger that the AC may aid the process whereby organizational norms and values are maintained regardless of whether those norms/values are conducive to organizational effectiveness. The use of ACs may also bring into focus questions about the relationship of the organization and the individual employee. Finally, as part of an organization's personnel policy, the standing of ACs with regard to racial and sexual discrimination legislation needs consideration.

Assessment Centres and Organizational Systems

As organizations grow a predictable consequence is functional specialization. Berrien (1976) cites Haire (1959) as finding that a specialized personnel officer was established in four firms when they grew to a size of 177, 152, 138, and 248 employees. The similarity in size 'suggests that the accumulation of special kinds of issues pertaining to the recruitment, hiring, promotion, and morale of people reaches a critical point where the full-time attention of a specialized

erson is required' (Berrien, 1976; p.53). In one s
unctional specialization within organizations is no more th
xtension of the economic principle of 'division of la
dentified by Adam Smith in 'The Wealth of Nations'; improve
n efficiency can result from breaking down complex tasks
mponents and assigning individuals to those components.
nother sense, organizational specialization is a reflection
rld in which increasingly diverse areas of detailed knowl
late to the solutions to particular problems. Much of
ape of manufacturing industry, for example, seems to de
om the 'need to divide and subdivide tasks and from the fur
ed to bring knowledge to bear on these fractions and from
nal need to combine the finished elements of the task into
nished product as a whole' (Galbraith, 1972; p.32).

le specialization serves various purposes, in Berrien's
'is one of the lamented features of large organizations t
many specialities have evolved. Too few persons
ilable who possess the global view which holds the speciali
their proper perspective' (p.53). Galbraith (1972) arg
t there is a positive side to this process in terms of
ividual's adaptation to the organization. Adaptation is

> ... reinforced by the nearly invariable tendency
> ndividuals to narrow the universe so that it is coterminu
> ith their own horizons. This is most important.
> chool teacher's world is the school ... The world of
> ureaucrat is his unit, section, branch or bureau ... To
> esire of the individual to mould the world to his goals,
> houghtful Providence has added the illusion of a great abili
> do so. This is accomplished by reducing each individual
> rld to manageable size. Adaption, as a motive, is mu

strengthened as a result' (p.165).

But whether or not taking a narrow view of the organization is consistent with individual adaptation, there is little doubt that a global view is consistent with efficiency. Returning to the specific issue of ACs, they must clearly be seen in the context of organizational objectives.

> 'First we must consider whether a selection approach is appropriate to solve the organization's problem. The selection model assumes we can improve organizational effectiveness by better choice of people. A number of other approaches to performance improvement may be equally effective, including improved performance appraisal and feedback, individual career planning, supervisory and management development, organization team building, and organization restructuring. What is critical here is the systematic diagnosis of the organization problem, the clear statement of organizational objectives, and the careful consideration of alternative personnel programs before the unquestioning adoption of a specific technique such as an assessment center' (Thornton and Byham, 1982; pp. 400-401).

The decision as to whether to use ACs should also take into account features of the wider personnel system of which the AC will form a part, e.g. job advertising, pre-selection, other selection/promotion systems within the organization, training, career development and remuneration. In addition the functions that each part of the system is meant to perform should be clearly specified, for example, a statement of the skills and aptitudes expected of new recruits, together with a plan for building upon those skills and aptitudes in training and early work experience.

A coherent personnel policy should also include some attem
'human resource planning' (HRP) which Zedeck and Cascio
lefine as 'an effort to **anticipate** future business
nvironmental demands on an organization and to mee
ersonnel requirements dictated by those conditions' (p.464
ubstituted for italics). A number of conceptual models
RP process exist, but much of the literature is
rescriptive and exhortatory. Research attention has
ocused on forecasting human resource supply and de
erminations and retirements, and on management succe
lanning. While various sophisticated HRP models have
eveloped, to date HRP seems to be more of an ideal t
eality (Zedeck and Cascio, 1984).

othwell (1984) uses the term 'employment policy' a
escription of attempts to integrate an organization's pers
d human resource planning activities and shape them to
ganizational objectives. She notes that, while such pol
e recognized as important, they tend to be under-emphasise
lation to other policies, e.g financial and marketing. To
tent this may reflect the potential complexities of employ
licy. As Rothwell points out:

'The rapidity of change, and its increasingly drastic imp
makes policy planning more essential, while making it
straightforward' (p.31).

The logic of integrating ACs with related personnel activities, and personnel administration with human resource planning, all under the umbrella of an employment policy with organizationally defined objectives, is manifest. Putting theory into practice seems more problematic. Perhaps this is not surprising given the complexity of the undertaking and the dynamic and uncertain nature of the environments in which most organizations operate.

Organizational Norms

Despite changing conditions some organizations appear to maintain quite consistent norms and values over time. Harries-Jenkins (1980) interprets this, in the British context, as a reflection of the way in which 'the dominant elite' promotes the growth of public and private sector bureaucracy so as to be 'able to rationalize decisions, particularly those which are unpalatable, as the inevitable result of bureaucratization' (p.319). One of the ways in which control is maintained is through organizations' selection of staff:

> '... in highly bureaucratized organizations, the selection process is a means whereby power groups at the center are able to maintain control within the organization. Although such selection is in reality the expression of individual preferences based on a complex set of internalized attitudes and values, the process and its effects are rationalized by the elite as indicative of a search for technically competent staff' (p.323).

One of the examples Harries-Jenkins cites is the Bri

which

> 'maintains its preference for selection based traditional military ideology of "leadership quali its emphasis on the diffuse superiority of the leader preference negates pressures for selection based ideology which emphasizes the qualities of tech professional expertise. Even though the latter is fu related to the requirements of a technological organization, ... traditional patterns and criteria recruitment still prevail' (p.322).

Such criteria favour individuals from prestigious e institutions who are ideologically committed to t organizational practices.

It is not necessary to go along with the 'conspiracy element in Harries-Jenkins' argument to recognize th have a valid point. There probably is a tende organizations to select in their own image even when the outdated, and this process is likely to be rationaliz search for competent staff. Taking the argument a stage success in an organization may reflect to some degree th of the 'fit' of the individual and the prevailing cultu tendency for an organization to select in its own i result in validity coefficients which further substant selection rationale, and hence the status quo.

The possibility of such circularity is something of which those using ACs and other selection methods ought to be aware, though the practical implications of this awareness are not altogether clear. To some extent the issue relates back to questions discussed in the context of job analysis in Chapter 2; when looking for correlates of effective job performance it is important to try to distinguish task-related from culture- and policy-related determinants.

Organization and Employee

It is sometimes claimed that ACs are a stimulus to organizational change. For example:

> 'Since an assessment program yields powerful information about the strengths and weaknesses of assessees, it has high potential utility for improving management development efforts. By forcing attention to participant strengths and weaknesses, an assessment center program encourages career planning, attention to morale considerations, and more intelligent efforts at organizational change and development. The assessment center does not directly do this, but it acts as a stimulant to bring it about. Although it is possible to install a management assessment center without changing other personnel programs, the natural history of many assessment programs indicates that this is not likely to happen. If it did, the yield would be only a small part of the benefit such a program can provide in a total personnel management system' (Boche, 1977; p.247).

This sort of claim is difficult to put to empirical test, but when ACs are used to select from within an organization it seems reasonable to expect that some spin-offs of the kind Boche describes will occur. The emphasis, however is very much on the individual rather than the organization to change. As Mant

(1974) points out, the AC seems a particularly on
activity; he likens it to a 'cattle-market' with 'pre-man
bloodstock being appraised by a line of judges and ..
through the hoops' (p.37). Mant goes on to comment
practice by some organizations of scheduling tin
'development' at the end of the AC:

> 'This assumes an **individualistic** learning process, as each man enters alone to fight for survival and must p what scraps of learning he can on the way out.
>
> Such an approach denies the rich **institutional** le possibilities of so many people, from so many levels hierarchy of the same organisation, resident together single hotel. In other words, the institution inspec individual and may even develop him as an individual, bu individual never gets to inspect the institution and it assumed he could develop it in any way' (p.37; bold subst for italics).

his issue is probably more relevant to ACs than to
election methods. The high cost of ACs will be justif
erms of benefits to the organization, and allocatio
esources in this rather than in other areas may sugge
elief that the organization's structure is sound while
uality of the workforce leaves something to be des
rthermore the thorough-going nature of the AC itself may
is impression.

e way around the problem may be for the organization which
s to be seen to be acting on other fronts as well, for exa
involving its workforce in real organizational change.
re specific level, real-life organizational problems mig
cluded as simulations in the AC. Stewart and Stewart (

describe this type of simulation:

> '... a senior line manager not otherwise involved with the programme ... comes prepared to present the problem and to support it with sufficient facts and figures ... After his presentation he is available for further relevant factual information, but the participants are told firmly that they should not expect him to volunteer information after his presentation, nor should they expect him to chair or take any further part in the discussion except by invitation and then minimally. The participants are finally told that, since this is a real problem, any solution which they arrive at which looks workable will be seriously considered by whoever is actually responsible for solving it ...' (p.117)

This last point is particularly important if candidates are to believe that their views are being consulted. A drawback with this approach, however, is that the AC itself may not be conducive to outspoken criticism of organizational policy. Also minor modifications to the AC could be seen as little more than 'tinkering with the system'. Wider policy decisions are probably necessary if the employee is to play, and to see himself as playing, a real part in organizational change.

Acceptability of Assessment Centres

The acceptability or face validity of ACs for those assessed is generally reported to be high. For example, in a study of an AC run by a multinational organization Kraut (Ungerson, 1974) found highly favourable assessee reactions in seven countries. And Teel and Dubois (1983) found that while high scoring assessees reacted more positively to an AC than low scoring assessees, in both groups a clear majority of overall reactions were favourable. Thornton and Byham (1982), following on from Dodd

(1977), have drawn together much of the attitude resea
produced summary statistics. Some of their conclusion
follows:

> 'From 58 to 93% of the participants believe that the measures important managerial qualities ... Part reactions are a function of level of assessment perfo quality of feedback on performance, the purpose of the p and subsequent use of the data in the organization. Up report that their performance in assessment was differe "real life" situations, but less than 15% report undue in the program.
>
> ... When asked whether assessment center results sho used for promotion decisions, 75-100% of the responden Yes. Larger portions say the results should be us identifying developmental needs' (p.82).

The face validity for assessors is also generally acknowle
be high, though this has been the subject of less research.
(1977) surveyed attitudes to the AC of 489 IBM managers, 1
whom had participated as assessors. Answers to a vari
questions were generally favourable, though former assessor
rather more positive than other managers.

Various explanations for high AC face validity have bee
forward. For example Finkle (1976) interprets the app
anagers in this way:

> 'Persons with any business or other managerial experience had to form judgments about work performance and potential several similar types of exposure: interviewing - as hiring or other job filling; record examination - as in re or written materials; observation of work or reports on observations by other supervisors; occasional observati individual contribution at meetings or conferences; c contact. By contrast, many managers have been quite skep of the value of tests and have been encouraged in skepticism by such eloquent, if not entirely construc appeals for common sense as presented in **The Organization**

(Whyte, 1956). The assessment center theme gives the manager a close approximation to all of his usual exposures (interviewing, work observation, background data, formal and informal observation of individual contribution), but under more standardized conditions and with the opportunity to share reflections and judgments with other managers given similar exposure. At the same time, most assessment programs play down the use of formal tests - the object of most managerial skepticism about formal assessment. The result is a strong appeal and a sharp growth in assessment center activity' (pp.864-865; bold substituted for italics).

But however one explains high AC face validity, the fact of it has important implications. On the positive side it is a help to psychologists in their efforts to convince organizations of the value of a relatively sophisticated and standardized approach to assessment/selection. On the negative side face validity may be too high. As Ungerson (1974) comments:

'Ever since the early days of WOSBs, it has been necessary to correct, with research results, the enthusiasm of assessors, candidates and spectators alike. It is important that we should use procedures which are felt to be fair and informative but this alone is not enough. We must beware of allowing our enthusiasm for a procedure which looks good to persuade us that it is infallible. The best assessment methods make many errors' (p.12).

Discrimination Issues

The legality of methods of selection with respect to discrimination is nothing like as important an issue in Britain as in the US. Nevertheless there are two pieces of legislation, the Sex Discrimination Act 1975 and the Race Relations Act 1976, of which account should be taken by selection practitioners. These Acts use similar definitions to classify discrimination into 'direct' and 'indirect' types, described in the Commission

· Racial Equality's Code of Practice (1983):

Direct discrimination consists of treating a person, on raci
rounds, less favourably than others are or would be treated
he same or similar circumstances.

.. Indirect discrimination consists of applying in a
ircumstances covered by the Act a requirement or conditi
hich, although applied equally to persons of all raci
roups, is such that a considerably smaller proportion of
articular racial group can comply with it and it cannot
nown to be justifiable on other than racial grounds' (p
-7).

:her related term is 'adverse impact', widely used in the US
:h can be defined as 'the disproportionate rejection rate c
sub-group by comparison with the rest of individuals bein
:ssed for selection' (Runneymede Trust and British Psychogica
ety, 1980; p.45). This concept is closely allied to that o
irect discrimination' but broader in that it does not includ
qualification of 'justifiable' requirements or conditions.

rehensive guidance as to how to comply with legislative
irements is provided in a joint report of the Runneymede
: and British Psychological Society (1980) entitled
:riminating Fairly: A Guide to Fair Selection'. Its wide
.ng recommendations make clear that that the demands of
'' selection overlap to a considerable degree with commonly
tenets of 'valid' and 'efficient' selection. There is
g emphasis on job analysis and employee specification, the
by trained staff of carefully designed or structured
tion methods, and validation of methods against follow-up
:ia.

The real impact of the discrimination legislation on British personnel selection practice is difficult to assess. The subject receives very little attention in the personnel literature. Guest (1984), however, sees some evidence that organizations 'are prepared to operate within the spirit of the law. Agencies that offer selection interviewer training and those that supply selection tests are reporting an increased interest in "fair selection"' (pp. 14-15).

Chapter 4 - RELIABILITY AND VALIDITY ISSUES

The purpose of this chapter and the next is to bring together theory and evidence concerning the validity and value of the AC approach. The 'four faces' of test validity - content, construct, concurrent and predictive - provide one framework for evaluating ACs. 'Each of these four terms refers to the process of investigation through which the accuracy of inferences to be derived from test scores may be evaluated' (Guion, 1976; p.785). Underpinning validity is reliability of measurement; if tests are unreliable, inferences will be inaccurate. In the next section the reliability of ACs and their component parts will be considered. This will be folowed by a discussion of approaches to AC validation.

Reliability of Assessment Centres

Reliability concerns precision of measurement (Nunnally, 1970), a corollary of error of measurement. Reliability may be thought of as a component of validity. Validity concerns the accuracy of inferences from measures. To the extent that measures are in error, inferences will be inaccurate.

Precision of measurement' may in practice mean different tl
ccording to the different methods available for reliab:
stimation, e.g. test-retest, alternate-form and inte
nsistency. Different methods take into account diff
urces of error and are likely to give different estimates
y given case. Judgements about precision, then, are cle
lative to the types of error which are deemed to be import
ch relative judgements are accommodated in the idea
liability as **generalizability**. According to Campbell (19
the generalizability model the problem of reliabilit
ated as

'the extent to which scores on a sample of observations
generalize to the class (population) of observations to wl
hey belong. This forces the investigator to define
niverse he or she has in mind. If the universe of inter
ncludes observations made at different points in time, as w
veraging ratings obtained from different raters on differ
lays, then the sample must be representative of the time poi
omprising the population. If the universe includes seve
ontent factors then each of them must be represented in
ample ...' (pp.201-202)

model requires judgements of the representativeness
iability estimates in any given situation, and considerat
l shortly be given to what constitute representative estima
the AC context. As a forward to that two questions will
essed. First, given the multivariate nature of the AC, wl
tly is the measure whose reliability is to be assessed? A
nd, given the generic nature of the term 'assessment centre
it possible to make worthwhile generalizations abc
ability?

Reliability of What?

The information gathered in the course of an AC - exercise ratings, dimension ratings, peer ratings, interview ratings, test scores, etc. - is typically input into a process of human judgement resulting in final overall ratings (OARs). It is normally on the basis of these OARs that decisions are taken. Consequently in assessing the reliability of the AC the primary focus should be on the OAR. This said, it would seem a reasonable assumption that the reliability of the input will to some extent influence the reliability of the output. So it is helpful to know something about the reliability of AC 'components' as well.

Can One Generalize?

Tenopyr and Oeltjen (1982) point out that generalization about methods called 'assessment centres' is difficult because of the variety of techniques and assortment of constructs involved. Probably the most that can be achieved by AC reliability research is to indicate a general range within which the reliabilities of ACs constructed along broadly similar lines may be expected to fall.

What Constitute Representative Estimates of AC Reliability?

Before answering this directly, it is instructive to note the limitations of research which has been done on AC reliability. Most of this research is, in the writer's opinion, of very little value for four main reasons. Firstly, the primary focus has been on the reliability of dimension ratings. One of the things

erging from chapter two is the general lack of constr
lidity of AC dimensions. A focus on their reliability, th
/ distract attention from more important issues, particula
: reliability of overall exercise ratings and of the O/
:ondly, nearly all the research has centred on estimates
.er-rater reliability. This approach takes into account or
'or due to differences among assessors. It leaves out err
to variations in exercise content and to short te
ctuations in individual behaviour. Thirdly, in inter-rat
dies it is often unclear, as Jones (1981) points out, 'h
n pooling of judgements may have occurred and what the effe
discussion was on the final level of agreement' (p.81). Jon
id important pre- to post- discussion increases in inter-rat
iability. Fourthly, virtually all research has focused on tl
.abilities of AC 'components' - exercise and dimension ratin
ither than on the reliability of the OAR.

ably the most representative way to estimate AC reliabilit
d be to carry out 'alternate-form' studies in the same way a
ometimes done for pencil and paper tests. In outline th
oach would be to construct parallel versions of an AC an
assess the same individuals using the different ACs an
erent teams of assessors with a short period betwee
ssments. By correlating the parallel sets of score
ability estimates could be obtained for the OAR and also fo
.fic AC techniques - simulations, interviews etc. These
iates would take account of error due to differences betwee

assessor teams, differences in AC content, and short term fluctuations in assessees' performance. With a more sophisticated experiment designed along ANOVA lines it would be possible to isolate 'time', 'content', 'assessor teams' and 'assessees' as separate effects.

Reliability Studies

There are only three studies which approach this alternate-form model. One of these, referred to in chapter one, involved WOSBs in the early part of World War II. As described by Morris (1949)

> '... two batches of candidates were assessed by each of two Boards. One batch went first to Board A, then on to Board B. The second batch proceeded in the reverse direction. A double assessment of each batch was thus obtained, and allowances could be made for any "learning" effect. Significantly different acceptance rates were found. In 60 per cent. of cases there was agreement as to disposal. Disagreement on a major issue of disposal was found in 25 per cent. of the cases' (p.232).

Reliability as calculated by correlating the Final Grades (OARs) of the two boards for 116 candidates was 0.67 (Vernon and Parry, 1949). Though it is not clear how similar the content of the two ACs was, the general impression given by those who have written about WOSBs is that boards varied considerably during the early part of the war. So it would seem reasonable to view this experiment as an alternate-form study.

ris goes on to describe a subsequent WOSB reliabil
eriment undertaken in 1945:

The personnel used were the best and most experienc
vailable. They were given an initial period of comm
raining. Common forms of reporting were introduced and
tandard personality profile adopted. The basic design was
ollows: Two Boards X and Y were set up. They lived a
orked on the same premises but were sworn to have
ntercourse relevant to their selection tasks, during t
xperiment. Each Board simultaneously observed the sa
andidates performing the same tests.' (pp.232-233).

intercorrelation of the two boards' Final Grades for 1
idates was 0.80 (Vernon and Parry, 1949). Though th
riment was more sophisticated in terms of standardization
edures and training, it assessed only the reliability
ssor teams and cannot be seen as an alternate-form experime
the same way as the earlier one. The higher reliabili
re might be accounted for in these terms.

e these studies give useful indications of OAR reliability
reliability of scores on individual techniques is nc
rted.

s (1973) reports a study which, though not designed to asses
ability, may be construed in that way. An AC called th
/ Identification Assessment program (EIA) was set up at AT&T
one-day AC was modelled on the company's previousl
olished 2.5 day AC called the Personnel Assessment Progra
). 85 men and women were assessed at PAP shortly after EIA
OARs of the two ACs correlated 0.73. The media
correlation of seven dimensions rated at both ACs was 0.56

No information is given on the reliabilities of individual techniques.

Seen as reliability research, Moses' findings are to some extent confounded by AC length. For the PAP a fairer comparison would have been another 2.5 day AC. For the shorter EIA, however, the correlation with PAP may possibly give an over-estimate of alternate-form reliability.

A third relevant study is by McConnell and Parker (1972), though the sample size was small (N=21) and the research reported in little detail. Assessees underwent the same one-day AC on two separate occasions with different teams of assessors. It appears to have been a test-retest rather than alternate-form study, though the time between assessments is unspecified. Ratings of 'overall management ability' on the two occasions correlated 0.74.

Inter-Rater Reliability Studies

The second WOSB experiment provides a good illustration of what might be called 'inter-rater-team' reliability, and appears to be the only study of its kind. Many studies have looked at inter-rater 'within-team' reliability, but the evidence from studies where coefficients have been calculated only for post-discussion ratings, or where it is unclear whether discussion or pooling of judgements has taken place, seems open to doubt. The writer can find only two studies where the published accounts clearly indicate that assessments were

ıdependent. Schmitt (1977) looked at inter-rater reliability mension ratings of four AC assessors. Median reliabil ross 17 dimensions was 0.66. This rose to 0.84 af scussion, perhaps not surprisingly in view of the fact t sessors were instructed to produce ratings 'within 1 sc int of each other' (p.172). Jones (1981) found that the m er-rater reliabilities of pre-discussion summary evaluati two group exercises were 0.73 ('Command Task') and 0. 'opic Discussion'). After assessor discussion these figur e to 0.83 and 0.77 respectively.

iability of Specific Techniques

en the paucity of the reliability evidence emerging from earch, a logical next step is to turn to the literature cific assessment techniques, particularly the simulation sin s is a defining feature of the AC. As with ACs, go iability research on simulations is scarce. Bass (195 .ewed the test-retest reliability of ratings of leadership i lerless Group Discussions (eight studies). Test-rete rvals varied between three hours and one year, an ability coefficients ranged from 0.39 to 0.90, median 0.74 ver, in seven of the eight studies the same raters were use est and retest, so the independence of the assessments i tionable. In the only study to use different rater t-retest interval one year, N=172) a coefficient of 0.53 wa .ned.

Glaser, Schwarz and Flanagan (1958) found alternate form reliabilities of 0.74 for a leaderless group discussion and 0.34 for a dyadic role-play (N=80), though there was only one assessor per observation. Frederiksen (1961) investigated alternate form reliabilities of 34 performance dimensions (no overall rating) across 4 in-baskets. Median reliability was 0.31. This compared with a median 'odd-even' (split-half) reliability of 0.53.

Two other types of technique widely used in ACs are cognitive tests and interviews. Reliability is not really a problem with reputable cognitive tests as it has traditionally been a criterion governing test construction. Though the trend of recent years towards the use of latent trait models in item analysis makes the situation somewhat less clear, precision of measurement is likely to remain an important consideration. In fact the latent trait approach facilitates estimates of precision tailored to the individual case. Hambleton and Cook (1977) point out that the information obtained from scores on any particular test varies according to ability level:

> 'When information at an ability level is high, we have narrow confidence bands around our estimates. If information is low, we have wider confidence bands. Because the information function varies with ability level, it has been suggested that test information curves ought to replace the use of classical reliability estimates and standard errors of measurement in test score interpretations' (p.84).

In the first comprehensive review of the literature with the employment interview Wagner (1949) reliabilities of 174 sets of interview ratings ranged to 0.97 with a median r of 0.57. Subsequent reviews 1964; Ulrich and Trumbo, 1965) similarly noted reliability. However Arvey and Campion (1982) report research has been less pessimistic. It is clear depends on the objectives of the interview and th imposed. The earlier reviewers interpreted the resear as favouring structured over unstructured inte assessment centred on particular areas, such as int relations and career motivations, over assessment general suitability. Arvey and Campion draw out tw themes in the reliability and validity research which to be consistent with this picture.

'1. The use of board or panel interviews appears prc a means of improving the validity and reliabil interview. Perhaps sharing different perceptions different interviewers forces interviewers to become of irrelevant inferences made on non-job related vari

2. Use of directly related job analysis and information as a basis for interview questions i method of improving the accuracy of the interview' (p

Summarizing the Reliability Evidence

There is not much useful information available on AC rel Contrary to common opinion it is an extremely under-r area. The evidence relating to alternate-form reliat OARs, by far the most important single test of AC rel suggests coefficients of around 0.7. Jones' (1981

suggests inter-rater reliabilities for group exercises at around this same level. And the second WOSB experiment gives an estimate of 0.8 for inter-rater-team reliability of the AC OAR. Evidence as to the alternate-form reliability of simulations is scanty and inconsistent, but reports of coefficients only marginally above 0.3 for dyadic role-plays and in-baskets should at least be a stimulus to further research. The reliability of cognitive tests is likely to be satisfactory where those tests have been properly designed and applied, while the reliability of interviews appears to be strongly a function of the form the interviews take.

It is difficult to know what level of reliability should be expected of ACs. Validity, to which reliability contributes, is ultimately more important, and the two should really be viewed together. Nevertheless one comparison is with reliability of pencil and paper tests. Here, according to Nunnally (1970), one is in general 'suspicious of a test that has a coefficient under .80. Some of the better-standardized instruments have reliability coefficients over .90' (p.127).

Approaches to Assessment Centre Validity

> 'Prior to 1954 the notion of validity was in considerable disarray, and as everyone knows, a special APA committee on test validity and reliability was constituted to assist in instilling a modicum of order. The ensuing technical recommendations (American Psychological Association, 1954) collapsed the competing taxonomies into the now classic formulations of content, concurrent, predictive, and construct validity' (Campbell, 1976; p.202).

These formulations provide a useful framework for discuss complex validity issues, though compartmentalization m been carried to an extreme. The consensus now seems to b rigid categorization is wrong and that it is better to 'strategies' of validation rather than 'types' (Tenop Oeltjen, 1982). The essential unity of validity lies in t that each of the four validity terms

> 'refers to the process of investigation through whi accuracy of inferences to be derived from test scores evaluated ... That process is called validation, and it many forms; in the final analysis, the judgment that a t sufficiently "valid" for employment office use should be on a comprehensive and integrated set of investiga perhaps of all four aspects' (Guion, 1976; p.785)

/hile all four aspects of validity are relevant to ACs, the)n which most attention is centred are content, concurre redictive. The last two are frequently grouped together he heading 'criterion-related' validity.

ontent Validity Issues

-guments are often put forward for AC content validity, anc)proach sometimes appears to substitute for criterion-re :search. For example, in a survey of 115 ACs (includir)lice programmes) in US state and local government (Fitzg id Quaintance, 1982) a large number of validity studies ported but nearly all were content-orientated. Co lidity tends to be claimed for ACs on grounds such as llowing:

'Multiple exercises are included in an attempt to adequately sample the relevant content domain of incumbent behavior ... Properly designed assessment centers are carefully developed to provide observations of the participants' behaviors in a variety of contextually accurate job situations. The different situational exercises are designed to represent the various demands that confront incumbents in the target positions' (Neidig and Neidig, 1982; p.183).

The main problem with this sort of argument is that content validation of selection procedures involves much more than noting the resemblance of parts of the procedure to parts of a job. Guion (1976) indicates the complexity of the undertaking:

'In developing a test to measure a specific universe of tasks or observations, the test developer must define that universe with some care. In doing so, he specifies a "universe of admissible operations" ... and develops the test as a representative sample of that universe. It is evaluated by noting correspondence of test development procedures to the basic definition of the content area' (p.787).

In employment situations the definition of the content area can be problematic. As an illustration, suppose that in selecting for the job of supermarket cashier a work-sample test were devised consisting of the assessee operating a cash-till for 20 minutes with a queue of role-playing 'customers'. Such a test would have plausibility, but would probably not in itself constitute a content-valid selection procedure. Operating a cash-till might be one very important aspect of the job, alongside things such as reporting on time for work, handling money honestly, cooperating with other staff, etc. All of these various aspects would somehow have to be sampled for content-valid selection.

ıere ACs are concerned, arguments for validity based purely
ıntent evidence may be inappropriate for any of at least
asons. Firstly, as in the example of the cashier, simulat
sets of simulations are unlikely in themselves to pro
tally comprehensive representation of job content. Secon
simulations are clearly tests of 'maximum' rather thar
ypical' performance (Cronbach, 1970). Apparently contextua
curate job simulations may be contextually inaccurate to
ent that performance is directly observed and evaluat
rdly, ACs are frequently used to select for jobs for wh
ining will be given and then, as Sackett (1982) commen
le 'the assessment center may be a very useful select
ice, it is being used as an aptitude test rather than as
ple of necessary job skills' (p.142). Fourthly, claims for
tent validity tend to be based on the similarity of parts
AC to parts of a specific job on which it is targeted. Whi
ection for a target job may in some cases be the on
ective, in many ACs a major concern is long term potential.
ent valid test of potential would involve sampling t
ent of a range of jobs at different organizational level
hly, ACs are typically not designed as work-samples. Rathe
analysis is used to identify performance dimensions
tructs - which are used to select AC techniques. Techniqu
be chosen with job content in mind, but the approach
truct-driven. The inherent confusion is illustrated t
aker and Grant's (1982) model of AC content validation whic
ists of nine stages. The third of these stages is labelle

'Identification of content domain of knowledges, skills, and abilities' (p.2). The model is essentially a construct validity approach which makes reference to job content.

These arguments are not intended to suggest that content-orientated approaches are valueless, but merely that content validation on its own is unlikely to be enough. The emphasis implied by the term 'content validity' may be better described by the term 'content-orientated test development' (Dunnette and Borman, 1979). Use of the latter term reflects the usefulness of a content approach in AC construction, but recognizes that validity evidence should also be sought by other means.

Criterion-Related Validity Issues: Predictive versus Concurrent

These 'other means' are likely to be found in criterion-related strategies of AC validation. Predictive strategies, which by definition involve a time interval between the collection of test and criterion data, are commonly held to be more appropriate in selection studies than concurrent strategies.

> 'Scientific sampling requires that the sample represent the population to which the research is to generalize. The population of interest in employee selection is a population of potential, not actual, employees. In this sense, a concurrent validity design is working with what appears, on the face of it, to be an inappropriate population.
>
> However ... [if] the current employment status of the person has no effect on the data, then the use of the concurrent or present-employee sample is not a violation of scientific principle' (Guion and Cranny, 1982; p.243)

Judgements as to whether employment status materially affects the data are facilitated by a more differentiated classification of research designs than is expressed in the predictive-concurrent distinction. Guion and Cranny (1982) distinguish five principal types of predictive design:

> '1. Applicants are tested, and selection is random; test scores are correlated with subsequently collected criterion data (follow-up, random).
>
> 2. Applicants are tested, and selection is based on whatever selection procedures are already in effect; test scores are correlated with subsequently collected criterion data (follow-up, present system).
>
> 3. Applicants are tested and selected on the basis of the test scores; test scores are correlated with subsequently collected criterion data (select by test).
>
> 4. Applicants are hired and placed on the payroll; they are subsequently tested (e.g. during an orientation or training program), and the scores are correlated with criteria collected at a still later time (hire, then test).
>
> 5. Applicants are hired, and their personnel records contain references to test scores (or other predictors), which may or may not have influenced the hiring decision. At some subsequent time, when criterion data are available, the files are searched for information that **might** have been used and validated had it occurred to anyone earlier to do so (shelf research).' (p.240; bold substituted for italics).

While it is also possible to distinguish different types of concurrent design, most concurrent studies would appear to fall into Guion and Cranny's 'present-employee' category where a test is administered to existing employees and the results correlated with currently available criterion data.

These various designs will be differentially affected by factors such as range restriction, motivation and contaminants such as age and tenure. Three particular problems with present-employee designs are: (1) the difficulty of making appropriate corrections for range restriction (also a problem with some 'shelf research'); (2) motivational differences between employees and applicants which may affect the results of test procedures (e.g. personality inventories) which fall outside the cognitive test category (also a problem with 'hire then test' designs); and (3) the likelihood of contaminants (particularly age and job experience) for which correction is difficult.

Overall it is clear that the choice of design used in a selection study may have important implications for the data obtained and that typical concurrent designs may be inappropriate in many selection situations, but that the decision as to the appropriateness of any particular design needs to be made in the light of the specific characteristics of the situation under study. Clearly, however, practical considerations may dictate the use of less than optimal designs.

Range Restriction

Excepting situations where no selection has actually taken place or where selection is random, results of criterion-related research studies are virtually always affected by restriction of range; where the sample exhibits significantly less variance than the population to which it is desired to generalize, the validity

ıe population will be consistently underestimated (Campbell,
. Schmidt, Hunter and Urry (1976) have produced tables to
that in typical cases where direct range restriction occurs
where the predictor itself is used as the basis for
;ion) statistical power is much reduced. For example if
'alidity is 0.5, criterion reliability is 0.6, the selection
is 0.2 and a two-tailed test is used with $p<0.05$, a power
(i.e. giving a 9 out of 10 chance of discovering the
ty that is present) requires a sample size of 278 compared
6 in the equivalent unrestricted case. Recent research
tt and Wade, 1983; Raju, Edwards and LoVerde, 1985) has
ed the work of Schmidt et al. to include the case of
ct range restriction (i.e. where range restriction is due
ection on another predictor correlated with the predictor
;erest) and found a rather more optimistic picture: to
e equivalent statistical power to the comparable direct
restriction case much smaller sample sizes are necessary.
:al evidence for the differential effects of different
of range restriction comes from metaanalyses by Schmitt,
, Noe and Kirsch (1984). For 114 cases where direct range
tion occurred, average validity was 0.259 compared with an
validity of 0.296 for 99 cases where only indirect range
tion occurred.

Range restriction has two main practical consequences. Firstly, it may cause the 'true' validity of selection procedures to be underestimated. Secondly, and more crucially, it may reduce the probability of detecting validity at all, that is, increase the likelihood of Type II error.

Underestimation

One way of tackling the underestimation problem is by statistical correction (Thorndike, 1949) though this is somewhat controversial. Campbell (1976) reviews theoretical arguments and empirical evidence and concludes that the use of correction formulae is a

> 'rather risky business. In almost every real situation their estimate of the population parameter will most likely be biased in some respect, except in very clear and straightforward situations. The safest recourse is not to use them and to fall back on the collection of more data in hopes of accounting for the selection factor by more substantive means' (pp. 218-219).

Collection of the right sort of data, however, is not often practicable in selection studies. Range restriction is such an important factor that even somewhat inaccurate estimates of its effects would seem to be better than no estimates. Validity is about the accuracy of inferences, and an inference about a population which ignores range restriction in the sample will be inaccurate however sound the basic statistic. A reasonable policy would seem to be to present the uncorrected and corrected coefficients alongside one another; the former gives the sample statistic and is relatively reliable and the latter, though less reliable, gives a rough guide to the corresponding population

stic.

ple Increase in Type II Error

ncreased probability of Type II error brought about by range ction is harder to deal with. Account may be taken of this stimating statistical power as in the Schmidt et al. e. Such estimates indicate sample sizes necessary to e given probabilities of detecting effects of specified

However it is frequently the case that sample size is a on of practical constraints rather than a matter of choice.

, Valenzi and Silbey (1978,1980) have argued that one way crease effect size and hence reduce the possibility of Type or is to form tests into linear composites.

> assuming that multiple predictors are used in a validity and that each predictor accounts for some unique erion variance, the effect size of a linear composite of predictors is likely to be higher than the effect size of single predictor in the battery' (Cascio, Valenzi and ey, 1980; p.135)

composites formed with unit weights Cascio et al. rate the truth of this assertion in a variety of tical but realistic situations. The linear composite is shed as an efficient approach to combining predictors, and tes formed with unit weights perform particularly well superior to regression based composites in very many ons (Schmidt, 1971; Dawes and Corrigan, 1974). However a lty with the Cascio et al. approach is the common need to e not tests but decisions made on the basis of tests. It

is likely that the use of linear composites to make decisions would be appropriate in many selection situations, but the fact is that decision making strategies are frequently sub-optimal and researchers are left with the problem of attempting to validate; the decreased likelihood of range restriction using linear composites may not be very relevant in such situations.

Choosing Criteria: Goals and Achievement

While range restriction is an important technical problem with the criterion-related approach, more fundamental difficulties are posed by the search for adequate criteria. According to Smith (1976) a main requirement of a criterion is

> 'that it be relevant to some important goal of the individual, the organization, or society. Determination of relevance is, however, a matter of judgment. Some group or person must decide which activities are most relevant to success. Once these activities have been identified, efforts must then be directed toward developing psychometrically sound measures of these activities. The measure of a criterion should be neither contaminated with irrelevant variance nor deficient in terms of measuring the important objectives of the organization and of the people in it.
>
> ... [Relevancy] consists of two parts. One is the validity of the goal which is judged to be important. The second is the validity of the measure(s) of the goal achievement. This requirement is parallel to the requirement that a test be valid' (Smith, 1976; p.746).

In practice it may be difficult to obtain clear definitions of goals and agreed judgements as to their relative importance. Goals can be various and conflicting, particularly for organizations with a public service brief. For example, among the goals of a police force one might list 'detecting crime', 'preventing crime', 'maintaining public order' and 'efficient

ment of resources'. In evaluating the contribution of an dual policeman to achievement of organizational goals there problems of how much weight to attach to performance in different areas. Also goals may not always be acknowledged be understated. A school headteacher, for instance, might ledge something like 'maintenance of staff well-being' as a objective, but his/her judgement of its importance 're to, say, teaching goals might not accord with parent and perceptions.

.ly one should start with the process of goal 'ication and proceed, as Smith suggests, to the design and .on of criteria to provide valid measures of goal ment; validity in this context, can be seen as 'the degree h the result of the measurement process (the numbers) ctorily represent the various magnitudes of the intended te' (Guion, 1980; p.393). In practice, however, logic appears to be turned on its head with criterion ility largely determining what is measured. Also criteria s promotions and salary attained often seem to be accepted in themselves, rather than as imperfect measures of goal ment. This said, the pragmatic 'use what's available' n to criteria should not be criticized too severely. dentification of goals is a logical first step, in reality nay not be articulated by an organization or its ntatives in any very clear or consistent way. It may be e nearest many organizations come to an agreed statement

of goals, and of the way individual employees are expected to contribute to those goals, is in things like annual reports and decisions to promote. In other words, the processes of goal identification and measurement of goal achievement may be inextricable. Clearly, however, it is desirable, where possible, that ideas of what different criteria may measure should inform the processes of selecting, developing and using those criteria.

The criteria most frequently used in AC validation have been indices of achievement in training, supervisory ratings of performance and potential, and 'objective' indices of promotions and salary attained (Schmitt, Gooding, Noe and Kirsch, 1984; Thornton and Byham, 1982). Few studies have departed from these basic formats. In the criterion-related validity studies of Home Office police ACs - to be described from chapter 6 onwards - the criteria used are in line with this general trend: supervisory ratings, training grades, and rank attained. It seems appropriate, therefore, to look briefly at these three types of criteria.

Job Performance Ratings

Performance ratings provide information which is, on the face of things, more job-centred, explicit and differentiated than that obtainable from most other sorts of criteria. In theory, the goals which the organization sets for the individual can be translated into rating scales and goal achievement assessed directly. However there are numerous psychometric problems with rating scales, notably leniency/severity, distribution errors,

ılo and other intercorrelational errors (Cooper, 1981; Jacc
ıfry and Zedeck, 1980; Smith, 1976). Also studies which h
›oked at ratings in terms of a multitrait-multimethod matr
th 'rating sources' (e.g. supervisors, subordinates)
ethods', have typically found moderate levels of converg
lidity and, with one or two exceptions (Lawler, 1967; Mou
34), only limited evidence of discriminant validity (Dunne
i Borman, 1979). It is of relevance to note that most of the
.titrait-multimethod studies have used ANOVA as opposed
'relational analysis, and the original Campbell and Fis
59) formulation of convergent validity as the extent to whi
ependent measures of the same trait agree has been translat
o convergent validity as a unidirectional main effect f
rsons' across sets of trait ratings. So typical findings
erate convergent validity appear to reflect a pervasive effe
global evaluation rather than trait rating reliability, whi
erally low discriminant validity suggests that raters tend nc
liscriminate reliably among different aspects of performance
picture fits in with the findings of Thomson (1970) wh
ed at inter-rater reliabilities of supervisors' ratings. H
d a median inter-rater reliability of only 0.52 for 1
ormance dimensions but a reliability of 0.85 for ratings o
ential', suggesting that raters 'were unable to arrive at
lse and common understanding of the meaning of the differen
es, and responded to some generalized notion of the goodnes
ıdness of the assessee' (p.501).

Many different formats for performance rating have been tried in an attempt to improve psychometric properties, for example forced choice (Sisson, 1948), mixed standard scales (Blanz and Ghiselli, 1972) and behaviourally anchored rating scales (Smith and Kendall, 1963). There seems, however, to be no consistent evidence for the superiority of one type of format (Tenopyr and Oeltjen, 1982).

In the absence of clear evidence for discriminant validity of performance ratings, and with no significant prospect of improvements in rating scale technology, it may be that the main use for supervisory ratings in selection research is in the derivation of global assessments of job performance and potential. Though this falls some way short of the ideal of measuring specific 'goal achievement', such global assessments are still useful, particularly if they can be shown to be reliable as in Thomson's (1970) study. However supervisors' ratings may provide only a part of the total picture of job performance. Supervisors frequently observe narrow and unrepresentative samples of employee behaviour (Mount, 1984), and it may be that different types of raters (supervisors, peers, subordinates, self) are needed for a more complete assessment. Mount (1984) reviews research which indicates limited agreement of supervisor and peer ratings, while self ratings show very little agreement with other rating sources. Subordinate ratings have been subjected to less research, though Mount found them more similar to supervisor ratings than to self ratings. The

esearch, then, would appear to support the notion that differ
ıting sources tap different aspects of performance.

ere seems to be something of a contradiction between resea
ndings which indicate that the main information conveyed
tings is of a global kind and the observation (Lawler, 19(
at factor analyses of performance ratings have typica
elded between three and five factors. The question that ari
this: How can trait ratings which typically show
scriminant validity form the basis of quite differentiat
ctor solutions? Part of the answer may lie in the somewh
ıitrary nature of factor analytical techniques; the variety
ision rules for numbers of factors to extract allied to thin
h as discretion over whether or not to allow correlat
tors means that 'number of factors' is not a particular
ective indicator. Another explanation could be that fact
lyses of performance ratings indicate more about the cogniti
ucture of the raters than the behaviour patterns of t
ees. Landy and Farr (1980) argue that a certain amount
earch evidence points in this direction, though the case
proven.

ıotions

ıotions represent a fairly long term and global criteric
e validity

s limited by the fact that many factors other tha
rformance may affect promotions, such as politica
pediency, organizational structure, and labor marke
nditions. Nevertheless, promotions represent

> chips-on-the-board decision concerning the value of a person to the organization ... [Promotions] are frequently not based on performance evaluation, but rather word-of-mouth and other informal evaluations, and hence, may reflect many situational factors ... Nevertheless, job level has been used by several investigators ... with success' (Smith, 1976; p.756).

Promotions, and the closely related criterion of salary increase, have been heavily relied on in AC predictive validation, with some quite favourable results. As will be seen, however, some criticisms of the AC predictive validity evidence centre on doubts about what it is that these sorts of criteria measure.

Training Criteria

Criteria such as performance during training and whether or not training is completed are of interest in their own right, particularly where training is expensive and/or where wastage rates during training are high. There are instances where training is such a crucial career stage that prediction of training criteria takes precedence over prediction of later job performance. Also metaanalyses (Pearlman, Schmidt and Hunter, 1980) have indicated that, at least for cognitive tests and clerical occupations, validities obtained using training and job proficiency criteria are highly correlated (r=0.77). The earlier metaanalytical conclusion of Ghiselli (1966) that aptitude test validities for training and job proficiency criteria correlated on average only 0.14 for all occupations appears to have been erroneous due to statistical artifact. Re-analysis of Ghiselli's data using broader test categories increases the correlation for the two criterion types to 0.82 (Pearlman et al., 1980).

low Many Criteria?

Selection theory has shifted over time away from a focus o
criterion towards a recognition of the need for mu
criteria. While the essential multidimensionality of
erformance - that is, the fact that jobs are composites of
hich may demand different abilities and skills - has long
cknowledged, early efforts to identify the single
riterion, involving a variety of methods for combining mea
see Smith, 1976), seem to have been motivated by the ne
ake unitary decisions about individuals. In recent years
as been a recognition of the loss of information which c
hen essentially heterogeneous data - from different t
ubgroups, and criteria - are lumped together. Dunnette (
ote of the possibility of improving the 'batting average
election research by the adoption of more complex predi
odels. Similarly the logic of sacrificing multidimension
the cause of decision making has been challenged o
ounds that it is 'more rational to give different weight
fferent predictions at that final point of decision tha
mbine functionally independent elements at the outset of
search' (Guion, 1976; p.794).

nahan and Muchinsky (1983), analysing trends in selec
search from 1950 to 1979, noted that researchers had not re
sponded to this change in theoretical perspective;
lection studies in the latter years still involved the us
ly one criterion.

A Ceiling to Prediction?

Even with the most sophisticated criterion-related validity strategies the extent to which tests can be expected to predict job criteria may be inherently limited. This is partly due to factors like predictor and criterion unreliability and range restriction which will depress the level of obtained coefficients. However the effects of range restriction can be estimated and where criterion reliability is known (which it seldom is) statistical correction for attenuation (Thorndike, 1949) is legitimate.

Perhaps more fundamental is the inability of the predictor-criterion paradigm to take account of the 'open systems' nature of individuals and organizations (Rundquist, 1969). A

> 'major characteristic of interest in an open system is the dynamic interrelationship of its components. This means that the same output can be the result of a different combination of the components or, to put it another way, that the same behavior can be the result of different antecedent conditions' (Rundquist, 1969; p.111).

Equivalent levels of test and criterion performance may, in theory at least, be attained in different ways. In Rundquist's words, 'phenotypic score similarities are no guarantee of similarities in the basic processes involved' (p.111). Whether or not this really matters may depend on whether one views tests and criteria as signs or samples (Wernimont and Campbell, 1968). If tests are seen as 'signs' or trait measures and selection decisions are based on a standard trait profile then Rundquist's

criticism is valid in that individual process differences being taken full account of. If on the other hand t viewed as samples of behaviour and criterion-related vali a function of the overlap of test and criterion content one has a framework which allows for the fact that given of test and criterion performance may be achieved in a va different ways. The samples approach can be aplied to not only about simulations but also pencil and paper Cronbach (1980), argues that test interpretations which close to actual performance, for example, 'performa reasoning with words, numbers, and diagrams' (pp. 39-40) more defensible than those which relate to general qualiti as 'intelligence'.

The view of **organizations** as open systems perhaps has implications for criterion-related validity. Rundquist comments that

> 'Just as the same test score can be obtained with dif mediators as antecedent conditions, so can similar out industrial systems be obtained by a number of combinatic selection, training, supervision, and production techr Changing any one of these may produce a marked effect c output or criterion' (p.113).

t was pointed out in chapter 3 that selection in organiz eeds to be seen as part of the wider personnel system de meet specified organizational objectives. For the sel searcher, however, factors other than individual differ ich affect criterion performance may emerge only as 'error e predictor-criterion relationship; it is virtually impos

to control for the effects of things like 'job placement' and 'relationship with supervisor' in the prediction model.

In considering factors such as these the very crude nature of the criterion-related approach to validation becomes apparent. It may be unwise to rely on this strategy alone. If one can argue that selection tests are 'job-related' one has a basis on which to expect them to correlate with relevant job criteria; Guion (1976) calls this the 'rational foundation for predictive validity' (p.802). Job-relatedness overlaps to a considerable degree with 'construct validity', a concept which has been defined less rigidly in recent years than formerly. According to Cronbach (1980): 'Construct validation is nothing more than argument that combines data and accepted beliefs to bridge over uncertainties and reach a persuasive prediction' (p.44). Within this sort of framework the role of criterion-related validation becomes essentially (dis)confirmatory. If the results of a criterion-related study conform to expectations, an additional piece of evidence has been obtained in support of the basic rationale. If not, there is a case for questioning both the validity of the rationale and the quality of the criterion-related evidence.

Criterion Contamination

Another reason why criterion-related evidence should be treated with caution in the evaluation of selection procedures is the possibility of 'criterion contamination', a term used to refer to two basic types of systematic error. One type of contamination

:urs where both selection and criterion measures are partly
ıction of common factors which are not strict
formance-related. For example, if those in an organizati
make selection decisions and those who make promoti
isions both favour individuals from a 'public scho
kground', and if such a background is unrelated to 'tru
formance, correlations obtained between selection assessmen
promotions, when interpreted within a predictive validi
ıework, may be spuriously high. Where selection procedur
not clearly job-related there is likely to be considerabl
ɔe for the imputation of factors of this kind in interpretir
ɛrion-related validity evidence. As will be seen, son
ıicisms of AC criterion-related evidence have been based o
type of contamination argument.

econd type of contamination may occur when selectio
ssments directly influence criterion measures. This i
/ant where selection assessments are fed into the personne
gement system with the possibility that selection assessment
hemselves influence later assessments. In most AC validity
es selection information has not been kept secret, and it is
imes argued (e.g. Sackett, 1982) that the interpretation of
findings of such studies is as a result problematic. Some
ical evidence on this issue comes from Huck and Bray (1976).
compared means and standard deviations of supervisory
gs, and also the correlations of those ratings with earlier
ɛings, for groups whose supervisors knew and did not know of

subjects' assessments. Similar statistics were obtained for both groups suggesting that criterion contamination was not an important factor in the ratings.

Chapter 5 - EVALUATING ASSESSMENT CENTRES

In this chapter the criterion-related validity evidence for ACs will be examined, and criticisms of that evidence discussed. Consideration will also be given to 'utility'; to justify the use of costly ACs it is necessary to make some assessment of the 'payoff' relative to other cheaper methods. Utility may be seen as one aspect of the broader construct of 'usefulness'; a full assessment of usefulness will also take into account the system-wide repercussions of ACs, and these will be discussed. Also considered is the overall efficiency of the AC, which affects both validity and utility; several possible types of inefficiency are identified.

In the next section, the criterion-related validity evidence for ACs will be looked at. Following that will be discussions of the efficiency and usefulness/utility of ACs.

Assessment Centre Criterion-Related Validity - Evidence

From Chapter 1 it is apparent that since World War II ACs in Britain and the US have tended to develop along somewhat independent lines, the former being characterized by the War Office Selection Board (WOSB) and Civil Service Selection Board (CSSB) tradition, and the latter owing much to work at American Telephone and Telegraph (AT&T). Techniques naturally tend to cross national boundaries, and the terms 'British-style' and

-style' will be used to refer to an historical rather than t
eographical distinction. There are now many more US-styl
1 British-style ACs in operation, a situation which ma
.ain and/or be explained by a far greater volume o:
erion-related validity evidence for ACs of the former type
validity evidence for US-style ACs is documented i1
iderable detail in several published reviews, and detaile
tition here seems unnecessary. However the criterion-relate
dity evidence for British-style ACs seems never to have bee
n together in the same comprehensive way, and a closer loo
be taken at this area.

tyle Assessment Centres - Criterion-Related Validity Evidence

ies of US-style AC valididty reported in the literature have
the most part yielded favourable results. Finkle (1976)
ews a number of the major studies, and a detailed and
ustive review, including several unpublished studies, is
ided by Thornton and Byham (1982). No attempt is made here
replicate Thornton and Byham's review, but mention will be
of one or two of the more important and/or more original
les before going on to look at metaanalytical findings.

inal US-style AC validity study is the original AT&T
stigation (Bray and Grant, 1966; Bray, Campbell and Grant,
in which, as described in Chapter 1, AC OAR was shown to be
y predictive of management progress. Howard (1981; quoted
hornton and Byham, 1982) reported that the validity
icient for both the college and non-college groups reached

0.46 in the early years, declining by the sixteenth year to 0.33 for the college group and to about 0.40 for the non-college group. Since all those assessed were already employed by AT&T and no use was made of the assessment results, there was no direct range restriction. However the college graduates had already come through a fairly intensive traditional selection system so it is likely that some indirect range restriction would have occurred. Seen in this light, the obtained validity coefficients may be somewhat conservative.

In another important AT&T study (Bray and Campbell, 1968) 78 men who had already been selected for the job of 'communications consultant' (salesman) underwent an AC consisting of an interview, biographical data blank, three cognitive tests, a 'contemporary affairs test' and three simulations: 'leaderless group discussion', 'oral fact finding exercise', and 'consulting case'. Assessment findings were not released. Criterion ratings were provided by trainers, supervisors, and a special review team who observed firsthand 'actual behavior in sales contacts 6 mo., on the average, after assessment' (Bray and Campbell, 1968; p.38). There was no significant relationship between trainers' or supervisors' ratings and the field performance criterion. AC OAR did not correlate significantly with supervisors' and trainers' ratings, but correlated 0.51 with the review teams' ratings.

Of at least as much interest as this basic validity coef quite high by normal standards, is the apparent weaknes supervisory ratings as a criterion. As Bray and Campbell comment 'the assessment center might have been conside sufficiently accurate for use if supervisory judgment ha relied upon as the sole criterion' (p.40). But the field criterion should not be accepted without question as a b against which to assess the validity of supervisory rating might ask, for example, whether field reviewers were ob 'maximum' or 'typical' performance and whether their eval were of 'true' job effectiveness or of the degree to salesmen's behaviour conformed to company norms and expecta It is noteworthy that actual sales were judged an unsu criterion.

The use of supervisory ratings as criteria can resu respectable validity coefficients as is exemplified in a st Thomson (1970). He investigated an AC used by the Standar Company (Ohio) which consisted of 'objective and proj ests, participating in simulated management exercises nterviews, and making a prepared oral presentation' (p.498 anagers and psychologists as assessors. AC dimension ra ere found to predict corresponding supervisory ratings in onth to two year three month follow-up of 71 professi echnical and managerial personnel with median correlatic .42 for psychologists and 0.38 for managers. Rating otential at the AC correlated 0.64 with supervisors' ratin

potential for both psychologists and managers.

Another validity study of a US-style AC (Tziner and Dolan, 1982) is of interest for two reasons. Firstly there appears to have been little if any range restriction, and secondly the AC took place in an environment very different from the US, male-dominated, business-managerial context in which such ACs have typically been operated. Israeli women soldiers (N=193), aged 18 to 19, who had expressed an interest in attending officers' school were all admitted to training following an AC consisting of five simulations: in-basket, leaderless group discussion, individual presentation, 'field command game' and role-playing game. Predictive validity of the AC was compared with 'traditional' selection methods, namely superior evaluation, interview, cognitive and projective tests. AC OAR correlated 0.38 with a rating of training performance, though this was bettered by the correlation of a single verbal ability test with the same criterion (r= 0.39).

A study by Schmitt, Noe, Meritt, and Fitzgerald (1984) is noteworthy for its use of innovative criteria. These investigators researched an AC for school administrators consisting of 'two in-baskets, a semi-structured interview, a fact-finding and decision-making simulation with an oral presentation, and an analysis and group discussion of a case study' (p.209). AC OAR correlated significantly with overall performance ratings of supervisors (r=0.25, N=118) and of teachers (r=0.29, N=119), but not with overall ratings of support

(r=0.09, N=116) and not with student ratings of school
:e (median r= -0.05, N=68). This is an important study
AC validity obtained with a conventional supervisor rating
ion was only partially supported when less conventional but
y relevant work performance criteria were used. The
gs would seem to support those critics who have argued that
imited range of criteria typically employed in AC validity
ch calls into question the interpretations put on the
ce. This issue will be returned to in a later section.

ul way to take stock of the criterion-related validity
ce for US-style ACs is by referring to metaanalyses by
t, Gooding, Noe and Kirsch (1984) of studies published
1 1964 and 1982. The sources were the Journal of Applied
.ogy and Personnel Psychology, so the ACs included may be
l to be of the US type. The average of 21 validity
:ients (not corrected for range restriction) was 0.407.
down according to type of criterion, average coefficients
'performance ratings' 0.428 (6 validities);
ement/grades' 0.312 (3 validities); 'status change' 0.412
dities); and 'wages' 0.237 (4 validities). While the
s on status and salary criteria is clear, it appears that
about equally valid for performance criteria.

British-style ACs - Criterion-Related Validity Evidence

British-style ACs are today much less common than their US counterparts. While the WOSB tradition is still strong in the British public sector, recent applications in the British private sector (e.g. Dulewicz and Fletcher, 1982) have tended to be along US lines. Nevertheless there have been several criterion-related studies of British-style ACs which merit attention.

Validity evidence for WOSB and CSSB procedures was discussed in chapter 1. To recap: Vernon and Parry (1949) reported a validity coefficient of 0.165 (0.35 after correction for range restriction) for wartime WOSB OARs with later field performance assessments as the criterion (N=500). For post-war WOSBs Reeve (1971) found validity coefficients, using training performance criteria, of 0.217 (N=3965), 0.280 (N=649), and 0.153 (N=684). Vernon (1950) found validity coefficients for CSSB OAR of 0.422 (N=106) with a rating of potential following training, 0.254 (0.509 after correction for range restriction; N=147) with a general performance rating for administrators after one year, 0.164 (0.505 after correction; N=202) with performance of administrators after two years, and 0.215 (0.499 after correction; N=123) with performance of Foreign Service staff after one year. Anstey (1977) found a validity coefficient for the CSSB-based 'Final Selection Board' mark of 0.354 (0.660 after correction; N=301) with rank attained by administrators with 30 years' service.

)47 the Royal Navy set up the Admiralty Interview Board
', an AC for officer selection designed along WOSB lines. In
)riginal AIB candidates were assessed by teams of seven
:sors using pencil and paper ability and aptitude tests,
aphical questionnaires, 'individual command and group
', group discussions (general topics), short talks, and
idual interviews (Gardner and Williams, 1973). The AIB is
in operation, though it has undergone considerable changes.
s been the subject of several mostly unpublished predictive
ity studies, summarized by Jones (1984). For the period
1960 the emphasis of validation research was on the
:tion of 'officer like qualities' (OLQ) as assessed during
ing, and validity coefficients of 0.36 (N=214; 0.47 after
:tion for range restriction), 0.34 (N=40), and 0.19 (N=74)
'ound for AIB OAR with the OLQ criterion. Gardner and
ms (1973) further report that for the first officers to
the navy via AIB (1947-1949), OAR correlated -0.22 with
taken to reach commander, the career grade. Following
s in the use made of AIB assessments in 1960, overall
ng performance became a more relevant criterion than OLQs.
tudy in the early 1960's OAR was found to correlate 0.32
; 0.52 after correction) with this broader criterion. A
r study for the years 1981-1983 produced a correlation of
ith total training marks of 0.36 (N=565; 0.52 after
:ion).

Though ACs now in use in the British private sector tend to be of the US type, in the 1940s and 1950s there were several industrial applications of WOSB- and CSSB- type methods. A small criterion-related validity study of such an application was reported by Castle and Garforth (1951) and Handyside and Duncan (1954). A procedure which would now be called an AC and which included biographical history sheets, intelligence tests, interviews and group discussion sessions was used to select for promotion to supervisory positions in a Scottish engineering works. Castle and Garforth (1951) reported a correlation of OAR with a performance rating criterion obtained nearly two years after selection of 0.68 (N=44). Handyside and Duncan (1954) continued the follow-up to four years and found that OAR correlated 0.65 (0.72 after correction for range restriction) with a composite criterion of performance ratings and promotions. Estimating the reliabilities of the predictor and criterion at 0.95 and 0.65 respectively, Handyside and Duncan went on to correct for attenuation and arrived at a coefficient of 0.92 commenting that 'the experimental procedure was predicting so well that little improvement could be made except by improving the reliability of the criterion' (p.20). In the light of subsequent research findings the uncorrected validity coefficients produced by this investigation seem extraordinarily high.

validity study of similar sorts of methods used in
election of administrative trainees for a large South Af
ndustrial corporation was reported by Arbous and Maree (1
19 candidates of whom 168 were existing employees were ass
n two group exercises: a leaderless group discu
iiscussions of general topics) and an 'assigned leadership t
committee chairmanship). An unspecified but 'fairly l
mber of the applicant population' (p.84) were available
llow-up. Average final ratings of three assessors over the
ercises were found to correlate 0.60 with a criterio
pervisors' ratings of potential administrative capacity a
e year. However the extent to which selection assessm
fluenced the criterion is unclear. Follow-up data
ndidates who were already employees appears to have l
llected for both selected and unselected groups, which w
early have distorted the findings.

'en the small number of validations of British-style ACs
ir disparate natures, it would be misleading to attempt
raging of coefficients. Nevertheless these studies h
erally, with the possible exception of WOSB which was in
e highly experimental, found ACs to be reasonably predict
chosen criteria. To this extent the findings
tish-style ACs are in line with those for their
nterparts.

Assessment Centres are Valid: So What?

While criterion-related validity studies of ACs have generally yielded favourable results, no one need be surprised by this. Selection decisions resulting from lengthy and expensive ACs, which are merely composites of tried and tested selection techniques, really ought to relate to relevant training and job criteria unless something is seriously wrong. Simulations or work samples, the techniques central to ACs, have consistently been shown to be valid in their own right. In metaanalyses by Robertson and Kandola (1982) median validities were found for work samples categorized as 'individual, situational decision making' (including in-baskets) of 0.25 with training criteria (9 coefficients), 0.28 with job progress criteria (11 coefficients), and 0.28 with job performance criteria (26 coefficients). Median validities for work samples classified as 'group discussions/decision making' were 0.33 with job progress criteria (16 coefficients) and 0.35 with job performance criteria (10 studies). With validity coefficients of this order for single work samples of the AC type, the somewhat higher validities achieved by complete ACs, which may include several work samples alongside other types of technique, should not be unexpected.

What do Assessment Centres Predict?

Though the results of AC criterion-related validity studies have generally been favourable, job criteria have been overwhelmingly measures of status/salary change and supervisors' ratings of performance and potential, with status change by far the most

ular single criterion (see Thornton and Byham, 1982). Th
es it possible to question whether ACs predict more th
ewhat narrow aspects of job success, especially when t
ited range of criteria is viewed in the light of the findin
a review by Cohen, Moses and Byham in 1974 (cited by Klimosk
Strickland, 1977) which indicated generally higher validiti
re ratings of potential rather than performance were t
teria. Such considerations led Klimoski and Strickland (1977
suggest that ACs may be 'prescient' rather than valid:

> Could it be that assessment center staff members are able t
> evaluate a candidate using all the data obtained **from the poin
> of view of the organization's decision makers?** In a sense th
> well-trained staff member is making a judgment of potentia
> based on his or her knowledge of the organization in which th
> incumbent must operate, but also on knowledge of th
> proclivities, propensities, and preferences of those highe
> level managers who must ultimately make promotion decisions
> Herein lies a possibility when advancement data are used fo
> criteria: Assessment centers may work because assessmen
> center staff are able to anticipate or predict how (and on wha
> biases) operating managers will make their decisions in the are
> promotions. Thus, what we may have is a special and subtl
> kind of "criterion contamination," or at best, anothe
> demonstration of policy capturing ... But do we have
> validity?' (p.358; bold substituted for italics).

e the Cohen, Moses and Byham review, a more extensive revie
Thornton and Byham (1982) has indicated no significant
erences in the AC's ability to predict several different
s of criteria. It seems that significant results are just as
y to be obtained with performance ratings as criteria as
potential ratings and progress/salary indices. This to some
t negates Klimoski and Stricklands' arguments, though there
till an important underlying point. Criteria used to

validate ACs, whether they be assessments of performance/potential or decisions made on the basis of such evaluations resulting in promotions and salary increases, have much in common. Wallace (1974) argues that the amount of chance variance associated with real-world performance, together with various practical considerations, discourages the use of objective criteria. Even where large differences in individual performance exist, objective measures of performance may be highly unreliable. As a result researchers are forced to turn to criteria which are 'deceptively predictable'.

> '... we can do a better job of predicting what people will say about an individual's performance than the performance itself. Not only that, but one of our best predictors of ratings of various performances has turned out to be peer ratings which, puzzlingly enough, seem to be as potent when based upon very limited exposures to the ratees as when based upon long and continued exposures. Is it possible that we have developed a system which measures and predicts some quality we might call the ability to make people say good things about oneself? Is this spuriously present in both our predictors and criteria?' (Wallace, 1974; pp.403-404)

Clearly the ability to predict how an individual will be judged by others is not unimportant, but equally clearly the reliance on this type of criterion which is a feature of selection research in general and of AC research in particular is undesirable. The study by Schmitt, Noe, Meritt and Fitzgerald (1984) described earlier in this chapter illustrates the more differentiated picture of performanance that can be obtained when a wider range of criteria are used.

ssessment Centres Efficient?

·iterion-related validity evidence for ACs indicates that ıpproach is a potentially useful one in the selection ct. But this is not to say that ACs are necessarily ent selection methods. It may be that less costly methods be employed to similar effect or that greater benefits be obtained by improvements in AC design. The question: s efficient? is clearly meaningless in the abstract, the being: It all depends on the AC. To gauge the efficiency particular AC one needs answers to three more specific ons: Firstly, is the information gathered during the of the AC efficiently processed? If not, it is of st to know why not. This leads into the investigation of al validity which generally translates, in the AC context, investigation of assessors' use of information. Secondly, re redundancy in the information generated by the AC? And /, could any or all of the information generated by the AC ıined more economically by other means? The extent to research studies have attempted to answer these questions w be considered.

ssment Centre Information Efficiently Processed?

ıestion has been addressed by two types of research son. In one, assessors' judgements have been compared ose made by assessees themselves (peer evaluations) at the the AC. In the other, clinical/judgemental methods of ng AC information have been compared with

mechanical/statistical alternatives.

Three studies have compared the predictive validities of assessor and peer evaluations. Mitchel (1975) found that for assessments of 'potential' peer ratings were just as predictive as assessors' ratings of a salary growth criterion. Vernon (1950), however, found that while peer nominations correlated positively with a performance criterion they were less predictive than OAR. Similarly Turnage and Muchinsky (1984) found that peer nominations predicted supervisors' potential ratings, but were less predictive than OAR. While these studies would seem to suggest that peer evaluations are on the whole a less efficient way of processing AC information than assessor evaluations, it is worth noting that the two studies which found assessors' ratings to be superior predictors were correlating managers' assessments at the AC with later managerial assessments. It would be of interest to compare validities using peer assessments as criteria.

As regards clinical/judgemental versus mechanical/statistical processing of AC information, the evidence reviewed in chapter two suggests a general superiority of the latter approach. Judgemental combination, which is almost universal, appears to be a source of AC inefficiency though why this should be the case is largely unclear. There has been very little published research into assessors' use of the information provided by the AC's various components.

nal Validity or Assessors' Use of Information

pparently straightforward way to investigate assessors' nation use is with regression analysis. If measures derived some of the AC's components are found to correlate at a gible level with AC OAR then one has evidence that these es are little used in decision making. Gardner and ms (1973) found just this for several pencil and paper which in fact proved to be predictive of training and job ia. More usually, however, AC component measures correlate oderate level with OAR and with each other (e.g. Bray and 1966; Borman, 1982; Tziner and Dolan, 1982), making it ult to identify causal relationships. A problem with using le regression in such circumstances (i.e where there is colinearity') is that estimates of regression coefficients uctuate from sample to sample (Kim and Kohout, 1975). As a uence when multiple regression is used to try to identify AC techniques which contribute most to OAR, conclusions may ccurate.

fundamental problem with regression-based investigation of on making was identified by Vernon (1950) who, in the of reporting his research into early CSSB procedures, ed: 'Successive gradings were not intended to be ive, yet each set was ... inevitably affected by previous ions, hence there is no way of assessing the value of the exercises independently' (p.82). This point would seem valid for any AC where assessors are involved in scoring

more than one part of the procedure. The problem is particularly acute where interviews form part of the AC and where these are tailored to the individual as previously assessed.

Vernon (1950) identified a method which partly meets these difficulties. The approach is to build up regression equations by adding in scores from individual techniques in the order in which they are obtained. In this way it is possible, for any given order of techniques, to model, albeit crudely, the contribution of each additional piece of information to decision making. Vernon (1950) applied the method in a fairly broad way, adding in groups of scores to the regression equation rather than individual scores. His findings indicated that different types of technique - cognitive tests, examinations, background data, group exercises and interviews - each made a useful contribution to final assessments. The writer is not aware of any other use of this type of approach, but has applied it in research into Home Office ACs described in chapter 7.

Is There Redundancy in the Information Generated by the AC?

Clearly information will be redundant if it is not used in decision making, but redundancy may also occur through sheer duplication of information provided by different tests. Redundancy in this sense is a function of the collection as opposed to the combination of information.

research studies have looked at the extent to which iques of different types provide different sorts of ıation. Bray and Grant (1966) showed that correlations of .mension ratings and a salary progress criterion were `icantly reduced but remained substantial when correlations pencil and paper ability tests were partialled out, ıting that scores on pencil and paper tests were useful :tors but that additional criterion variance was accounted ıy the remaining techniques: interviews, simulations, :tive tests and personality, attitude and biographical onnaires. Tziner and Dolan (1982) found that a multiple R 400 for pencil and paper ability tests with a training ion was increased to 0.499 with the addition of scores from consisting of five simulations, though evaluations from a ured interview and ratings of candidates (female soldiers) uperior officers added virtually no unique variance. ski, McCloskey and Bourgeois (cited by Thornton and Byham, found correlations for pencil and paper tests and tions with a criterion of increase in management level of and 0.258 respectively compared with an overall multiple R 02 for the two types of technique together.

findings suggest that simulations and pencil and paper may each make independent contributions to the prediction terion variance, though the situation as regards other of AC technique is less clear. Much more research is ed; the aim should be to build up a differentiated picture

of the degree of overlap of different types of AC technique in the prediction of different types of criteria.

It is also important to ascertain the degree of overlap amongst techniques of a similar type, in particular simulations which are by definition universal to ACs and which tend to be lengthy and expensive relative to other types of technique. Sets of AC simulations often appear to make similar demands - for example 'working in groups', 'use of social skills' - and the potential for redundancy would appear to be great. What relevant research there is does indeed suggest redudancy. Tziner and Dolan (1982) found a multiple R for five simulations with a training criterion of 0.36 compared with a correlation of overall rating from a single simulation ('role playing') with the same criterion of 0.34. Wollowick and McNamara (1969) conducted a stepwise multiple regression analysis of scores from six simulations with a promotional criterion and found that only two (an in-basket and a manufacturing game) contributed to the overall R of 0.39. And Borman (1982) found that a unit weighted composite of scores from five simulations correlated at 0.48 and 0.35 with two training criteria compared with correlations of 0.41 and 0.36 respectively for the most predictive single simulation in each case. Redundancy, then, would appear to be a consequence of the use of multiple simulations, though more research is needed.

Could Information be Obtained More Economically?

While the focus so far has been on internal efficiency, the question: Could any or all of the information generated by the AC be obtained more economically by other means? concerns external efficiency. What is at issue is whether techniques which do not typically figure in ACs could usefully substitute for the AC or for some of its components. Two main types of technique to consider in this respect are biodata and peer and supervisory evaluations. Some evidence as to their predictive validity relative to ACs comes from metaanalytical comparisons by Schmitt, Gooding, Noe and Kirsch (1984) and Reilly and Chao (1982). Their findings are shown in Table 1.

Table 1 - Average Validity Coefficients for Various Predictor-Criterion Combinations

PREDICTORS	PERFORMANCE RATINGS	STATUS CHANGE	WAGES/ SALARY	TENURE/ TURNOVER	PRODUCTIVITY
Assessment Centre	0.43 (6)	0.41 (8)	0.24 (4)	--	--
Biodata	0.32 (29)	0.33 (6)	0.53 (7)	0.21 (28)	0.20 (19)
Supevisor/Peer Evaluations	0.32 (12)	0.51 (9)	0.21 (4)	--	--
Biodata	0.36 (15)	--	0.34 (7)	0.32 (13)	0.46 (6)
Peer Evaluations	0.37 (18)	0.51 (5)	--	--	--

Note (1) Figures from Schmitt et al. (1984) above the dotted line with those from Reilly and Chao (1982) below; data bases used by these two sets of authors overlap to some extent.
(2) Figures in brackets are numbers of coefficients averaged.

The figures suggest validities for biodata and peer and supervisory evaluations of a comparable order with those for ACs over a range of criteria. It is also worth noting that biodata were found to correlate usefully with 'tenure' and 'productivity', measures rarely used as criteria for ACs. Metaanalyses, however, can provide only a very general guide to relative validities since the sets of studies from which average coefficients are obtained may differ in important respects. Also such analyses give little indication of the extent to which different techniques may supplement or substitute for one another. What is needed are studies which specifically compare ACs and non-AC techniques for the same samples.

A few studies have made straight comparisons of the predictive validity of ACs, biodata, and peer and supervisory evaluations. Turnage and Muchinsky (1984) found that one composite biodata measure ('career objective - level desired') predicted three ratings of potential (median r=0.28) at a higher level than AC OAR (median r=0.23). The same biodata measure also predicted 'transfers and reductions' from the target job (r=-0.22) whereas OAR did not (r=0.01). However, there was no attempt at cross-validation in this study. Somewhat similar results were obtained by Drakely (1984). He found that biodata (cross-validated Weighted Application Blanks) correlated 0.50 and 0.17 with two training criteria compared with 0.62 and 0.46 respectively for AC OAR (coefficients corrected for range restriction). However biodata also correlated 0.24 with

commitment related turnover' (voluntary withdrawals
ompulsory terminations attributed to lack of motivation) whe
AR did not (r=0.00). The two studies together seem to indi
hat respectable predictions of criteria traditionally use
alidate ACs can be made by biodata alone, but the dispari
hich emerge in relation to other sorts of criteria suggest t
asures derived from ACs and biodata are of different kir
is conclusion is borne out by the findings of Slivins
Closkey and Bourgeois (cited by Thornton and Byham, 1982).
ltiple R of 0.402 for simulations and pencil and paper te
th a status change criterion was significantly increased
483 by the inclusion of biodata.

e picture as regards supervisory and peer evaluations is l
ear. Tziner and Dolan (1982) found that superior office
aluation correlated 0.16 with performance in training compa
h a correlation of 0.38 for AC OAR with the same criteri
erior evaluation and a selection interview combi
tributed virtually no unique variance to multiple Rs based
ulations and pencil and paper tests. Though superi
luation here emerges as a weak predictor in relation to t
type measures, it would be of interest to make the comparis
relation to job as well as training criteria.

Hinrichs (1969,1978) compared AC predictions with a 'naturalistic' management evaluation which bears some similarity to supervisory evaluations. The naturalistic measure consisted of the combined ratings of two managers who had access to all the information available in assessees' personnel files. The correlation of AC OAR with management level eight years later, partialling out management level at time of assessment, was 0.40, compared with 0.49 for the naturalistic evaluation (N=30). Another finding of interest was that management level after eight years correlated 0.46 with AC OAR and 0.55 with the naturalistic evaluation, whereas a multiple R for the two together was only 0.58. Taken together, these findings suggest considerable overlap in the predictive power of the two types of assessment. However, in view of the small sample size and the wide confidence intervals which should be applied to the correlation coefficients obtained (e.g the 95% confidence limits for a 0.49 coefficient with N=30 are 0.19 and 0.72), replication seems desirable.

There is insufficient research on which to base any firm conlusions as to the relative efficacy and overlap of peer/supervisory evaluations and ACs as predictors. Hinrichs' research is suggestive in relation to supervisory evaluations but replication is needed. There appears, however, to have been no research at all comparing ACs and independent peer evaluations.

:ions about the internal and external efficiency of ACs are .al to AC evaluation, but taking this section as a whole it)parent that relevant research evidence comes from only a 'ul of published studies. It is perhaps a measure of AC face ity that large resources continue to go into the design and mentation of programmes while very basic and quite ghtforward research questions about efficiency remain wered.

ity and Usefulness

nental to rational selection is some attempt to estimate / benefits in relation to costs of alternative selection :gies, though costs and benefits will not necessarily be :fiable in monetary terms. When formalized, the process of ing at such estimates may be called 'utility analysis' :d by Cascio (1982) as 'the determination of institutional)r loss (outcomes) anticipated from various courses of ' (p.127). This definition reflects the reality that ion is normally carried out on behalf of organizations. y utility will vary according to the values assigned to es, with the result that individual and public assessments lity may differ markedly from organizational ones.

Validity may in some circumstances suffice as a measure of 'relative utility'. If it is found that criterion-related validity of one selection procedure is matched or bettered by the validity of a cheaper procedure with no adverse 'side effects' (e.g. deterring good applicants), then the cheaper procedure clearly has greater utility for the chosen criteria. However when the question is of the form: Are the benefits of a more expensive procedure relative to a cheaper procedure worth the additional cost? validity coefficients in themselves are of little assistance.

Cascio (1982) summarizes models available for utility estimation, the principal ones being the Taylor-Russell, Naylor-Shine, and Brogden-Cronbach-Gleser models. An important way in which they differ is in the units used to express utility. The Taylor-Russell model makes possible estimates of changes in 'success'/'failure' rates for given combinations of validity, selection cut-off score, and base rate (the proportion of candidates who would be successful without the selection measure). The Naylor-Shine model differs in that it produces linear rather than dichotomous estimates and does not require classification of employees into satisfactory and unsatifactory groups. It enables estimates of the increase in average criterion score for a given validity and selection ratio. Taking the process a stage further, the Brogden-Cronbach-Gleser model enables estimates of the average productivity gain (in money) per person selected as a function of the validity coefficient, the

core on the predictor of those selected, and the deviation of job performance (in money). Hence the riterion'.

tary 'payoff' from selection procedures provides a t comparison with selection costs, but there is an question as to how it should be calculated. Cronbach represents traditional thinking in arguing that tion of payoff comes down to 'tracing effects on cost, work output, spoilage, turnover, and other that have balance-sheet consequences' (p.39). But the is fraught with difficulties as exemplified in a study (1965). He carried out a cost-benefit analysis for pencil and paper tests and questionnaires being d for the selection of Radial Drill Operators. The andard deviation of job performance (SD$) was calculated payoff for each operator, that is, profit as a function ity and quantity of production. Yet, despite the quantifiable nature of the outcomes, the attempt at 'off measurement met with considerable accounting And it is clear that in business settings exact ll become progressively more difficult to determine the ne moves away from point of monetary transaction for example, functions like product development and elations. In the public sector the difficulties are ar greater as payoffs, for example in a police force, seen in monetary terms at all. Even in business the

attempt to analyse all activities in terms of profit and loss may be misdirected. The implied goal of 'profit maximization' may be a considerable over-simplification of organizational objectives, as demonstrated by Seashore and Yuchtman (1967).

Non-Traditional Payoff Estimation

As a reaction to the obstacles in the way of traditional payoff estimation other methods have been developed. Cascio (1982) outlines two recent approaches to calculating SD$. One, developed by Schmidt, Hunter, McKenzie and Muldrow (1979), involves direct expert judgements of differences in dollar value of output of individuals performing one standard deviation apart (i.e. at the 50th and 85th percentiles). The second approach, developed by Cascio and Ramos, involves the assumption that the economic value of an employee's labour is best reflected in his/her salary. Principal job activities are rated in terms of time/frequency, importance, consequence of error and level of difficulty. These ratings are then translated into dollar values such that the sum of values for an individual job equals the average salary. Individuals are then rated on the dollar-valued activities and a payoff for their performance is derived. The mean and standard deviation of the payoffs for all employees can then be calculated. Another method of calculating SD$ (Eaton, Wing and Mitchell, 1985) is the Salary Percentage Technique suggested by Hunter and Schmidt (1983). On the basis that SD$ has typically been found to fall between 40% and 70% of annual salary, the method is simply to use 40% as a quick, inexpensive

and approximate estimate. Eaton, Wing and Mitchell suggest two further approaches which appear to have applicability in the public sector. One - the 'Su Equivalents Technique' - involves assigning a value to a performance and estimating the number of above average percentile) performers needed to match the performance of a number of average (50th percentile) performers. The other ystem Effectiveness Technique - involves the translatic ssessed performance differences into estimates of resu ifferences in the effectiveness of system units/compor hese differences in effectiveness are then valued accordi ne costs of the units/components.

hese different approaches have been found to result in w ffering estimates of SD$ in given situations (Eaton et 985; Weekley, Frank, O'Connor and Peters, 1985) though it m at different techniques are appropriate in diff rcumstances; there are indications, for example, that hmidt et al. (1979) method of directly estimating the mone lue of performance differences works best where there levant reference points for judgements, for example objec les data or costs of contracting out work.

general, however, two important problems with these sorts chnique may be identified. Firstly they tend to place g iance on the accuracy of supervisory-type ratings - a reli ch seems difficult to justify in the light of the kn eliability and lack of discriminant validity of such rati

and the fact that they provide only one, perhaps quite limited, perspective on performance. Secondly, and more fundamentally, they are a long way from the traditional idea of dollar-utility as 'balance-sheet consequences'. The status of such figures as data on which to base decisions might be questioned. Nevertheless payoff estimates of this kind may be better than the likely alternative of no estimates.

That this type of approach can be useful in AC evaluation has been demonstrated by Cascio and Silbey (1979). As their starting point they used the Schmidt et al. (1979) method to estimate SD$ for second-level sales managers selected without an AC. Then they systematically varied all relevant parameters as a means of discovering their relative effects. Over a hypothetical five year period AC cost was found to have a relatively small impact on dollar utility compared with selection ratio, criterion standard deviation and validity. It is clear from this study that for most selection situations the cost of ACs should not of itself deter potential users.

Other Aspects of Usefulness

Finally, standard utility estimates deal only with one aspect of 'usefulness', that is usefulness as a function of predictive accuracy. For a full appreciation of usefulness the system-wide impact of selection procedures should be considered. Selection has repercussions, particularly in terms of the perceptions of those affected. A good example is wartime WOSBs where, as described in chapter 1, the AC was successful in winning

dence at all levels in the army, with beneficial effects on cation rates and recruitment. Almost regardless of validity can be judged to have been 'useful' in these terms. on (1985) has drawn attention to similar processes in ion to modern day in-company ACs. Well run and well ned ACs can inspire confidence in the same way that poorly ived ACs can have deleterious effects on factors such as e and turnover.

ptions of assessees external to the organization are also tant in terms of the extent to which self selection is itated. Herriot (1984) and Williams (1984) have both drawn tion to the AC as a realistic job and/or organizational ew. Robertson and Kandola (1982) review some evidence ating that when people are given an opportunity, via work es, to find out how they will fare in a job they are likely elf select, with beneficial effects on eventual turnover . It would be reasonable to expect this sort of effect to for ACs, provided simulations are an accurate sentation of job tasks and not simply 'off-the-shelf' .ses. A parallel sort of effect might be expected to occur opportunities are provided for aspects of organizational e to be experienced, as, for example, where ACs are held in vice training centres (Williams, 1984).

Summary of AC Evaluation

A metaanalysis of studies of US-style ACs (Schmitt, Gooding, Noe and Kirsch, 1984) has produced an average criterion-related validity of 0.407 (not corrected for range restriction), with ACs about equally valid for job performance as for status and salary criteria. Though there have been insufficient studies of British-style ACs for meaningful use of the metaanalytical approach, these ACs have generally been found moderately valid. But ACs are lengthy composites of tried and tested assessment techniques, particularly simulations which have proven validity in their own right; that ACs prove valid should come as no surprise. Also valid ACs are not necessarily efficient. Assessors' information processing appears to be at least as efficient as that of assessees, but less efficient than mechanical/statistical processing. While the indications are that simulations and pencil and paper ability tests each make unique contributions to the prediction of criterion variance, redundancy seems a probable consequence of the use of multiple simulations. Techniques which might be considered as substitutes for ACs are biodata and peer and supervisory evaluations. The evidence is that respectable predictions of criteria traditionally used to validate ACs can be achieved with biodata alone, but that the measures obtained from biodata and ACs are of different kinds; biodata should supplement not substitute ACs. With regard to peer and supervisory evaluations, there is insufficient research on which to base conclusions about the relative efficacy and overlap of these methods and ACs as

ictors.

oski and Strickland (1977) suggested that criterion amination could have contributed to favourable AC validity ings. Though their specific argument is not supported by nt evidence, it is clear that criteria used in AC validation to fall within a narrow range. Questions remain about what e criteria measure.

g utility analysis Cascio and Silbey (1979) have shown that, nost selection situations, AC cost should not of itself deter ntial users. But one should consider also the wider impact election procedures, particularly on the perceptions of those cted. These perceptions may relate to things like morale, ver and self-selection.

Chapter 6

POLICE ASSESSMENT CENTRES AND HOME OFFICE RESEARCH NEEDS

In this chapter the thesis will be brought into focus with a short history and description of the AC under consideration - used for police selection and selection for training/promotion - and the identification of research questions associated with that AC. A look will also be taken at the use of ACs in other police contexts and at issues which may distinguish police use of ACs from other types of AC application. Finally detailed research questions to be addressed by this thesis will be outlined.

The Police Special Course and Assessment Centre Selection

ACs, described in the Home Office as 'Extended Interviews', have been used for selection to the Police 'Special Course' since it began. The course was recommended by the White Paper 'Police Training in England and Wales' presented to parliament in 1961. Designed for Constables the course was to be of twelve months duration combining training in aspects of police duty with a wider educational content as a 'stimulus to acquiring a broad and liberal outlook' (Home Office, 1961; p.3). On successful completion of the course, automatic promotion to Sergeant would follow. A course along these lines came into being in 1962 at the Police Staff College, Bramshill.

The course may, in one sense, be seen as a delayed consec actions initiated by the Police Post War Committee repc 1946 and 1947. In what Stead (1973) has called an 'act (the committee recommended both one-tier entry into the pc the setting up of a national police college. Before Worl the majority of senior police appointments were made fron the service, particularly from previously commissioned in the armed services (Skitt, 1982). A corollary of entry was the need to develop talent from within and to better recruits. The Police College, open in 1948, made with courses for ranks of Sergeant and above. But by 196 clear that more was needed. The document in which t decision to set up the Special Course was embodied make the underlying concerns:

> '... we have given consideration to the question wheth is more that needs to be done by way of training, be mind the need to ensure an adequate supply of officers right qualities and adequately trained to fill the vaca the future. We are all agreed that in order to attr develop the right type of man the Service must not only rewards commensurate with what he might expect to rec other fields of the public service which are comparal must encourage him to feel that he will be giv opportunity to use his talents to the best advantage are left with the impression, however, that good n deterred from joining the service not because th unwilling to serve in rank of constable but because the that, whatever their qualities, they will, unless th exceptional good fortune, have to spend long years in th without any prospect of advancement. We have therefc considering whether there is any way in which the arrangements could be altered so as to make it clear promising entrant that if he has the qualities he can ri higher rank by merit without any more delay than is ne to ensure that he has a proper grounding and experience work of the service' ('Report of the Informal Comm para.5; quoted in the Report of the Working Party Special Course, 1974)

The Special Course has operated until the present day. Two important changes came about in 1968 on the recommendations of the 1967 Working Party on the Recruitment of People with Higher Educational Qualifications into the Police Service (the 'Taverne Working Party'). One change was a second promotion to Inspector for those who had successfully completed the Special Course and a year's satisfactory service as Sergeant on return to their forces; this was implemented with retrospective effect. The second change was the introduction of the 'Graduate Entry Scheme' (GES), linked with the Special Course, as a means of attracting more graduate recruits (see Report of the Working Party on the Special Course, 1974).

The GES offers graduates or those in their final year at University the chance of being assessed for the Special Course before joining the Police. That a measure of this kind was seen as necessary in 1967 is understandable in view of the fact that of the half million graduating from Universities between 1945 and 1965 only 25 joined the police (Skitt, 1982). The scheme has effectively operated on two levels, as Skitt (1982) illustrates:

> 'What ... appears to happen is that the scheme acts as the carrot to encourage applicants and whilst very few (27 in 1982 [out of 1,366]) are chosen, a substantial number continue to join the service in normal fashion. To quote an example ... the West Midlands Police attracted 55 applicants for the Graduate Entry Scheme in 1981-82 of whom only 1 was successful. Of the remainder, 27 were rejected as unsuitable for entry to the service by normal criteria, 7 subsequently withdrew their applications and the remaining 20 have continued their applications and been accepted for entry' (pp.13-14).

mpossible to say how many graduates would join the Police GES did not exist, but a very high proportion now join pplying through the GES and this would seem to support the at the scheme faciliates initial contacts.

cial Course Itself

cial Course has recently been extensively revised with from 1985. But from 1962 to 1984, the period within which rent research is located, the general aims and format of ourse remained substantially the same. The emphasis out has kept close to the initial intent of combining in aspects of police duty with a wider educational The specific aims of the course as stated in 1974 and ed until 1984 were 'to provide selected officers with

prepare them for the operational role of sergeant at the the course;

o prepare them for the operational role of inspector they will hold after twelve months' satisfactory duty as nt; and

to give them an insight into the roles and sibilities of senior officers' (Working Party on the l Course; Appendix 4)

se itself consisted of three phases:

Police Studies, following closely the syllabus for the ng examination for promotion to Inspector. Students were in the subjects taught but did not sit the qualifying ion as such since the course earned them exemption.

Phase II: Academic Studies, principally in economics, political science, sociology and social administration. Students were examined and required to submit a dissertation.

Phase III: Training in operational police studies and man-management to prepare students in the practical skills needed as Sergeants and Inspectors.

At the end of the course students received three grades for professional studies, academic studies, and overall performance.

Special Course Eligibility and Selection

In the early years of the course the field of eligibility was steadily widened and the initial stages of the selection process underwent some change. In every year since 1963, however, Extended Interview (EI) has been the final hurdle for all candidates. From 1969 to 1984 eligibility requirements and pre-selection arrangements remained constant, and these are the subject of this section. EI will be described in the next section.

Those eligible to apply for the course (excluding, for the moment, the Graduate Entry Scheme) were:

1. Constables who had passed the qualifying examination to Sergeant and were not over 30 (exceptionally extended to 35) with not more than 10 years' service; and

2. Sergeants who had been promoted too soon after qualifyi to be eligible as Constables.

all forces except the Metropolitan Police candidates we eened in two stages:

1. Force Board. Each police force was allocated a quot calculated in relation to its number of eligib Constables, of candidates for further consideration. Force Board, presided over by the Chief Constable or h representative and incuding representatives of the Poli Federation and Superintendents' Association, consider all applications on the basis of written reports and/ interviews. Candidates who had passed within the top 2(nationally at the preceeding qualifying examination t Sergeant were automatically entitled to an interview.

2. Central Selection Board (CSB). Candidates nominated b Force Boards were interviewed by the CSB consisting c several panels under the direction of one of H Inspectors of Constabulary. Each panel was presided ove by a representative of the Association of Chief Polic Officers and included representatives of the Polic Federation and Superintendents' Association and Non-Service Member (NSM). NSMs were mostly retire senior public servants with experience of interviewing Each candidate was interviewed by the three polic assessors and the NSM. On the basis of CSB grade

candidates were called forward to EI.

The Metropolitan Police operated its own separate screening procedure called the Metropolitan Police Central Selection Board (MPCSB). The format of MPCSB was virtually identical to CSB, except that the Chairman was a Deputy Assistant Commissioner from the Met. Invitations to attend MPCSB were on the basis of either a paper sift or performance in the most recent qualifying examination to Sergeant - those within the first 100 on the competitive list who satisfyied age and service conditions were entitled to appear.

The Graduate Entry Scheme was an entirely separate selection procedure with EIs again as the final stage. The GES was open to graduates or final year undergraduates under the age of 30. 'Outstanding' candidates could be considered above that age provided they would not be over 35 when due to attend the Special Course. Applicants specified force preferences, and the Home Office routed applications to specific forces accordingly. Chief Officers made arrangements for medical examination, obtaining references/reports and interviewing, and on this basis awarded grades with the effect that candidates would be:

1. accepted for normal entry and considered for EI;

2. accepted for normal entry, but not considered for EI; or

3. rejected for normal entry.

isions to call forward to Graduate Entry EI were based on th
cific grade awarded by the force, the distribution of grade
tionally and the number of EI places. Those successful at E
ned their forces as Constables, served the normal two years
ation, and were required to take the qualifying examinatic
promotion to Sergeant at the first or second opportunit
e. in their third or fourth year) and to pass it in al
jects at the first attempt. On fulfilling these condition
would be interviewed by the Director and Co-Director o
ce Extended Interviews who would decide whether or not polic
rience had confirmed the initial assessment. If thi
ssment was satisfactory (as was almost invariably the case
Graduate Entrant could attend the Special Course. After tha
same conditions would apply as for those who had reached th
ial Course via the internal selection system.

nded Interviews

the Special Course began EIs were undertaken by the Civi
ice Selection Board (described in Chapter 1) using th
edures developed originally for selection of Civil Servic
nistrators. In 1970 responsibility for the provision of E
s and exercises moved to a new unit within the Home Offic
ed the Home Office Unit at Civil Service Selection Board.
format of EIs, however, remained substantially the same, an

Police EIs today are still very similar to their CSSB counterparts.

EIs are overseen by the EI Director - a Chief Constable - and Co-Director - the head of the Home Office Unit at CSSB. Groups of five or six candidates are assessed by teams of three assessors. The assessors are a **Chairman** and a **Service Member** who are both high ranking police officers and a **Non-Service Member** (NSM) who is someone from outside the police, often a retired senior public servant. Assessment procedures consist of:

1. Group Exercises in which the group of candidates is directly observed by the three assessors. The exercises are the

   . **Group Discussion,** a leaderless exercise in which the group discuss several general interest topics chosen by the assesssors, and the

   . **Committee Exercise** in which the group is put into a number of hypothetical decision making situations with each candidate in turn taking the chair and running the discussion on his or her problem.

2. Written Exercises. These are the:

Written Appreciation in which the candidate reads through a file of papers - including such things as committee minutes, letters, excerpts from white papers and press cuttings - relating to a fairly complex problem and is required to analyse three or four options, recommend one and justify his/her choice; and the

Drafting Test in which the candidate is required to write a tactful and persuasive letter to deal with a difficult situation.

ncil and Paper Tests. From 1969 to 1982 three pencil d paper tests were included in EI. These were the n-verbal **Raven's Advanced Progressive Matrices**, a mixed rbal and non-verbal test called **Abstractions** which had en designed in-house, and the **General Information Test** T), a broad measure of general knowledge periodically ised. In 1981 a numerical aptitude test was added and 1982 Abstractions was replaced with a verbal ability sure.

erviews. The candidate is interviewed by the Chairman Service Member together (the 'Service Interview'), individually by the NSM. The Service Interview aims determine what experience (particularly job and police erience) the candidate has had and what use he/she has e of it. The NSM's interview focuses on the

candidate's personal and intellectual development taking account of family background and educational opportunities.

5. Peer Nominations. At the end of the EI programme each candidate nominates, from the other members of the group, first and second choices as best **'Senior Officer'** and as preferred **'Holiday Companion'**. Sets of nominations are later scored to produce two rank orders of the group members.

Assessors award an overall mark on an eleven point scale for performance in each of the group and written exercises. In the case of the Committee Exercise marks are awarded for performance as Chairman and as Member. After discussion of each others' marks individual assesssors are allowed to modify their original marks. Marks are then recorded and an average mark for each exercise (two average marks for the Committee Exercise) is calculated. Interview performance is marked on the same eleven point scale. In the case of joint interviews marks may be modified after discussion, but no averaging of marks takes place. Scores on the pencil and paper tests are converted to a normative seven point scale. At the end of the EI the three assessors meet (the 'Group Final Conference') and agree an overall **Final Mark** for each candidate on a seven point scale which has four pass marks and three fail marks. Several assessment groups are run concurrently and after Group Final Conferences have been held assessors from the different groups all meet together with the

Director and Co-Director at the Joint Final Conference where reports and marks are discussed. This meeting is designed to ensure uniformity of standards and has the power to alter Final Marks. In practice, however, changes of mark at this stage very rarely occur.

The New Special Course

Research to be reported in this thesis should be seen in the context of the Special Course as it continues into the future, and it is of relevance to describe the recent changes. These changes, while important, do not reflect any fundamental shift in the rationale for the course, and the final EI stage of selection is unaltered. A synopsis of the changes and the reasons for them has been produced by Chief Superintendent John Linnane for the Police Extended Interview Office (at the Home Office).

'The shortcomings of the previous format of the Special Course can be summarised as follows

1) Its failure to attract a sufficient number of officers of the right calibre. The Special Course was intended to provide up to 60 places at the Police Staff College. In addition there are 20-25 places for officers who join the Service under the Graduate Entry Scheme. [Excluding Graduate Entrants, the numbers attending the course in 1982, 1983 and 1984 were 18, 15 and 14 respectively.] ...

2) The perceived lack of credibility of some newly promoted Inspectors who were products of the Special Course. Many people in the Service, including some who had attended the Special Course felt that one years operational duty as a Sergeant was insufficient to equip them properly to assume the role and responsibilities of Inspector rank.

3) The Course was considered to be too long. Some potential applicants were deterred from applying for the Course, firstly because of domestic reasons ... Secondly a years absence might cause some problems on re-entry to practical Police work.

4) It was thought that many prospective candidates considered the Course too academically/intellectually orientated and something of a graduate preserve.

5) Special Course students were exempt from the requirement to sit the national promotion examination to Inspector although they did sit an examination which was set and marked by the Directing Staff at the Police Staff College. This point caused some resentment amongst non-special course officers. It was also felt that the time spent at the Staff College studying to pass this examination could be better spent.

The new style Special Course sets out to remedy those defects and the principal changes are:

i) The Course is now a sandwich type course.

ii) It is open also to Sergeants who fulfil age and service requirements.

iii) Automatic promotion to the substantive rank of Sergeant follows selection and automatic promotion to Inspector on successful completion of Part II. The hope is expressed that Special Course officers should be ready and considered for promotion to Chief Inspector after 2-3 years in the rank of Inspector. This should make the Course attractive to ambitious young officers with potential for advancement.

iv) Special Course officers are now required to pass the national qualifying examination for promotion to Inspector, although two attempts are permitted.

v) Although the Course is more practically orientated there is no reduction in the academic or intellectual abilities looked for in successful candidates' (Police Extended Interview Office, 1985; pp.4-5)

Arrangements for Graduate Entrants remain much as before, except that sitting the promotion examination to Sergeant may now be delayed beyond the fourth year and the application to attend the the Special Course may be delayed beyond the point at which the Graduate Entrant becomes eligible. These changes are designed to give forces and individuals greater flexibility as to how much experience as Constable, or in specialized functions such as CID,

is appropriate before the course is attended.

A Tabular Summary of Special Course History

For reference, the principal events in the history of the Special Course and Extended Interview selection which have been described are noted in Table 2.

Extended Interviews out of Context?

A principal aim of this thesis is to assess predictive validity for Special Course and Graduate Entry EIs. It should be said, however, that a priori one would not necessarily expect such validity to be demonstrated. While ACs have been found valid across a wide variety of situations, the Home Office use of EIs violates one of the recognized tenets of good AC practice: that the choice of tests and exercises should be grounded in analyses of the target job or jobs. EIs were originally designed on the basis of analyses of the work of Assistant Secretaries in the Civil Service resulting in the inclusion of administrative-type simulations such as dealing with information in dossier form, working in committee, and letter writing. That these tests were appropriate in the Civil Service context seems to have been demonstrated by the research of Anstey (1966,1977) and Vernon (1950) which has been described. But there seems no particular reason to expect validity to generalize to the police context. Although management theories (see Stewart, 1979) and some empirical research (Stewart and Stewart, 1978) highlight elements common to a variety of managerial jobs, there are also very

important differences due to factors such as function, level, nature of contacts/relationships and degree of fragmentation of work (Stewart, 1983). Stewart (1983) has commented that no adequate selection can take place unless these differences are appreciated.

Taking the argument a step further, it is instructive to look at job analysis information which is available for Special Course 'target jobs'. Up until 1984 the only rank for which the course

Table 2 - Summary of Main Events in the History of the Special Course and Extended Interview Selection

YEAR	EVENT
1946-7	Police Post War Committee recommends: (1) one tier entry into the Police Service; and (2) setting up a national Police College.
1948	Police College opens.
1961	White Paper recommends introduction of Police Special Course, offering training and promotion to Sergeant for selected serving Constables.
1962	Special Course starts at Police College. Extended Interviews introduced for selection to the course.
1963	Extended Interview becomes final selection hurdle for all Special Course candidates.
1967	Taverne Working Party recommends: (1) accelerated promotion to Inspector for ex-Special Course students; and (2) introduction of the Graduate Entry Scheme.
1968	Taverne recommendations implemented.
1970	Responsibility for provision of Extended Interviews moves from the Civil Service Selection Board to the Home Office, though the format of EIs remains the same.
1984	Changes made to format of and eligibility for the Special Course.

specifically trained was Inspector. However the target was somewhat blurred in that all trainees had to perform satisfactorily as Sergeant before reaching Inspector, and Graduate Entrants also had to work as Constables between EI and attending the course. And in the other direction there was a tacit expectation that most Special Course graduates would achieve Chief Inspector in due course, and that over the longer term many would reach the highest ranks in the police service. Nevertheless, the target was technically Inspector. Those successfully completing the course who later performed satisfactorily at this rank were technically successes, while those who failed to do so clearly were not. Unfortunately there is no very up to date analysis of Inspector jobs. A large scale nation-wide investigation is underway but has yet to report. However this position description for a Metropolitan Police Inspector resulting from research by PA Management Consultants (1968) at least gives a flavour of the work:

'1. Directs and controls a "Relief" (1 SPS [Station Police Sergeant], 4 Sergeants and 44 PCs) during his tour of duty.

2. Acts as Duty Officer responsible for efficient policing of Sub-Division.

3. Approves or amends duty roster and allocation of duties prepared by Station Sergeant.

4. Checks that Station Sergeant carries out duties efficiently and gives him advice or instructions accordingly.

5. Tours area by car or on foot to ensure PCs working as instructed.

6. Checks that reports and records, i.e. crimes, accidents, property found, etc., are correctly kept and action taken.

7. Decides deployment, assistance required, etc. to deal with any incident reported to him, e.g. street fight.

8. Takes command at any serious incident, e.g. fire.

9. Decides whether arrested person should be charged, nature of charge and whether bail should be granted.

10. Informs Superintendent of major incident, e.g. major fire with loss of life.

11. Prepares reports on serious occurences, e.g. fatal accident.

12. Ensures "minor" crimes are fully and efficiently investigated by PCs to whom they are assigned.

13. Ensures Probationers gain practical experience of all aspects of police duty and reports on their performances bi-monthly.

14. Instructs subordinates at monthly meetings on current legislation, new policies, etc.

15. Has contact with general public, e.g. at incident, callers at station on important matter such as complaint.

16. Contact from time to time with local authority officials, e.g. roadworks, insane persons, etc.

17. Works an 8-hour shift, plus reporting on time and including meal break of 3/4 hour. Basic 42-hour week increased by compulsory overtime to minimum of 48 hours. Shift changes every 3 weeks.' (Appendix XV-4)

There is not much suggestion here of the analytical/conceptual demands imposed by an EI Written Appreciation or of situations analagous to the EI Committee Exercise or Group Discussion. The Inspector's job seems to consist of a reasonable administrative load combined with the need to deal with people on a one-to-one basis, talk to groups of subordinates, and take command in emergencies. Exercises such as in-baskets, role-play interviews,

lecturettes and outdoor command exercises come to mind as possibly more appropriate than some of those currently in use.

Extended Interview Research and Research Needs

Although the rationale for using EIs to select for the Special Course appears somewhat weak, the fact remains that EIs have been used for this purpose since 1962 and continue to be used. It would clearly be of interest, therefore, to know something about EI efficiency and validity. There have been two main attempts at evaluation. The first consists of regular collection and collation of information on ex-Special Course students. This work has been carried out over a number of years by Chief Superintendent John Linnane of the Police Extended Interview Office. What his summaries clearly show is that those selected for the Special Course have made much better progress overall, in terms of rank, than those seen at EI and rejected. This is, however, to be expected since the Special Course is specifically designed to accelerate promotion. A finding which is, perhaps, of more interest is the overall level of success/progress achieved by those attending Special Course EIs even when unsuccessful. The picture for the fourth Special Course in 1965 is fairly typical of that for other years. By 1984, of those rejected at EI 90% had made Inspector, 67% Chief Inspector, 42% Superintendent, 16% Chief Superintendent, and 2% [N=2] Assistant Chief Constable or equivalent. While the figures for successful candidates are better (94% Chief Inspector, 36% Chief

Superintendent, 15% Assistant Chief Constable and 3% [N=1] Chief Constable) it is clear that overall those who have reached EI through the internal pre-selection system are a very high potential group. Other figures produced by John Linnane reinforce this impression. In 1980 he estimated that at any one time there were 6,000 Constables meeting eligibility requirements (including passing the qualifying examination to Sergeant) of whom less than 100 would be seen at EI. Reaching EI is in itself a considerable achievement (Police Extended Interview Office, 1980).

The other main evaluation of EIs was carried out by Mays (1972). For the 42 officers attending the fourth Special Course (in 1965) he found significant relationships between EI Final Mark and ratings of overall performance and potential made by tutors at the end of the course (r=0.287 with both criteria). However there were no significant relationships between Final Mark and supervisory ratings of overall performance (r=-0.202) and potential (r=-0.020) five years after the course (N=35). These negative findings seem to have had little impact on Special Course selection policy. In fact the report of the Working Party on the Special Course in 1974 made no reference at all to Mays' work. The small sample sizes used by Mays may be one reason for the lack of notice taken of his findings. In the light of the fact that selection for the course has continued to operate virtually unaltered for 13 years since Mays' study, there seems to be a case for replicating and extending his research with the

benefit of additional data that has accrued since that time.

Other Police Assessment Centres

Assessment Centres have been widely used for police selection/promotion in the US as a survey by Fitzgerald and Quaintance (1982) shows. As a result some quite original types of simulation have been developed, particularly for recruit selection. For example Dunnette and Motowildo (1976) describe an AC in which assessees receive brief training and are then put through a series of simulations such as 'traffic stop' in which 'the candidate assumes the role of a patrol officer about to issue a citation to a driver, role-played by the assessor, for failing to stop at a stop sign' (p.63). In Britain a recent survey of police forces (Linnane, 1985) shows that 21 out of 43 forces (excluding Scotland and the Metropolitan police) are now using ACs for basic recruitment. None of these ACs pre-dates 1979, which suggests a considerable recent growth of interest in the area. However the writer is aware of only one force (Surrey) which has published a description of its AC (Hayes, 1984). This AC consisted of written and verbal 'autobiographies', a pencil and paper general knowledge test, a drafting test and group discussion similar to those used in Home Office EIs, a lecturette, a physical fitness test and a final interview. A novel feature of the AC was that Police Constables (PCs), one per candidate, served as assessors. Each PC was assigned to a candidate and most of the candidate's non-test time was spent

with this PC - in the social club, watching a training film, visiting the PC's normal duty station, perhaps meeting the PC's colleagues or family. A principal aim was familiarization with the police job/environment in order to reduce wastage rates. Despite a high pass rate (approaching 70%), zero wastage for the 57 who had taken up appointment within the first year of AC operation compared very favourably with wastage in previous years.

But despite the volume of work in the police AC area there has been very little in the way of criterion-related validation. In fact the only relevant published studies appear to be those reported by Mills (1976) and Ross (1980). Mills (1976) describes some validations of ACs used by the Cincinnati Police for recruit selection. The largest of these was a follow-up of 122 patrolmen who had served from one to nine years in that position (no mean or median time reported) who had been selected using an AC consisting of one-to-one 'on street' role-play simulations, the MMPI, 'occasional interviews', background investigations and polygraph examinations. AC OAR was found to correlate significantly with police academy score (r=0.314) and with most recent efficiency rating (r=0.183). Other similar follow-ups with much smaller samples yielded comparable findings, that is moderate correlations of AC OAR with academy performance but low/non-significant relationships with later supervisory ratings. It is also noteworthy that AC OAR proved to be a much less good predictor of academy performance than conventional pencil and

paper testing (Army General Classification Test).

Ross (1980) has reported a validity study for 49 police managers (in five US police departments) all of whom had been assessed for promotion using ACs which consisted of 'two leaderless-group problem-solving discussions, an oral presentation by the participants on their background, a written exercise, a personal interview, a leaderless group discussion - assigned role, and two paper and pencil personality inventories' (p.90). AC OAR correlated 0.47 with overall job performance as rated 1.4 to 3.3 years later, though Ross suggests that there may have been some criterion contamination resulting from knowledge by raters of assessment results. Ross' findings suggest that conventional types of AC may prove valid in the police managerial context, though there is clearly a need for further research in the area.

Special Issues in Police Assessment

Two factors which are sometimes claimed to distinguish police from other occupational groups are job related stress and 'police personality'. Though they are not central to the present study, it is worth looking briefly at both these areas since such differences between police and other groups might have implications for police assessment. A problem, however, is that most research on police stress and police personality has been done in the US; generalizations to the British context must necessarily be tentative.

Police Stress

It is frequently stated that policing is a high stress occupation. The evidence to support this contention, however, is not strong. In a recent review, Malloy and Mays (1984) have suggested that the answer to the question: 'Is police work stressful'? is

> 'quite likely yes; however, there is a growing awareness that all occupations are stressful ... A major problem with the police stress hypothesis stems from a priori assumptions regarding the stressors inherent in police work. The anecdotal literature suggests that the impending threat of physical harm or death and participation in violence are the major police stressors. However, in studies of police officers, these activities fail to emerge as major stressors ... While law enforcement is likely a stressful occupation, it is probably stressful for reasons quite different from those typically presented in the literature. Judging from the strongest research in this area, it seems that helplessness and feelings of uncontrollability in the work environment may be a major source of stress for police officers. Beyond this, little can be safely concluded. Quite clearly, studies are needed that differentiate those stressors peculiar to police work from those of other occupations' (p.207).

Malloy and Mays go on to look at the few available comparisons of stress in police work and control occupations and conclude that there is little support for the hypothesis that police work is more stressful than other occupations; more well-controlled comparisons are needed. Also attention should be paid to variables which may mediate psychobiosocial dysfunction in response to stress and to variations in police stress as a function of occupational role; intra- as well as inter-group research is necessary.

A British example of an intra-group study with the focus on occupational role comes from Gudjonsson and Adlam (1985). They administered a 45-item self-report stress inventory to three police groups: probationary Constables, Sergeants, and 'senior officers'. Twenty-five items were found to represent at least mild stress for upwards of 50% of one or more of the groups. The list included such things as long hours, giving evidence in court, role ambiguity, dealing with a messy car accident, negative community attitude, and so on. These results are perhaps put into perspective, however, by the finding that by far the most stressful thing reported by Constables and the second most stressful thing reported by Sergeants was 'having to pass exams'. The only item to emerge as more stressful than this was 'job overload' for the Sergeant group. In general senior officers reported much less stress than the other two groups, though 'job overload' was again the single most stressful item. 'Dangerous or violent confrontation' also emerged as at least moderately stressful for all three groups; it was the second most stressful item for the senior officers and third most stressful for Sergeants.

Though of interest in their own right, findings such as these are difficult to interpret without a background of inter-occcupational research. It is not known, for example, how 'job overload' for Sergeants compares with that for junior managers in other occupations. In the absence of such information, there seems little reason at present to view stress

as any more of an issue in police assessment than it is in assessment of other occupational groups. However, where ACs are constructed on the basis of job analysis, it is to be hoped that some of the more stressful aspects of police jobs will be mirrored in simulations of critical tasks.

Police Personality

Lefkowitz (1975) has pointed out that the question of whether there is such a thing as police personality comes down to asking 'what is the evidence for describing policemen as a somewhat homogeneous group, differing psychologically from the general population and/or other occupational groups?' (p.4). In reviewing research, much of it methodologically deficient, into a number of traits/syndromes linked to the police occupation, Lefkowitz (1975,1977) concludes that there is evidence that the personalities of policemen do differ systematically from the rest of the population, but in an evaluatively neutral sense. There is some evidence of tendencies towards conservatism and distrust of non-white ethnic minorities, and a syndrome of isolation, secrecy, defensiveness and suspiciousness. On the other hand it appears that police as a group have a low incidence of psychopathology and are not particularly authoritarian or dogmatic.

It seems possible therefore to identify a 'modal police personality', but very little is known about its aetiology. Lefkowitz (1977) lists seven possible determinants: (1) role-specific behaviours - entirely bounded by the spatial and temporal limits of the job and not introjected by the policeman; (2) occupational socialization; (3) selective attrition; (4) organizational selection; (5) recruitment from a restricted population - disproportionate selection from class and ethnic subgroups who may share some characteristics of the 'modal personality'; (6) self-selection from a restricted population; and (7) self-selection from predisposed personality types. There is no data bearing on 1 and 3, but some evidence for 2,4, and 6. Considerable evidence exists for 5, policemen tending to be working class and lower middle class whites, segments of society sharing many of the personality characteristics, norms, values etc. typically found among the police. As regards 7, research is inadequate in that there 'are no studies that have controlled for socioeconomic class, differential rates of application from socioeconomic classes, and organizational selection effects' (Lefkowitz, 1977; p.358). Finally Lefkowitz concludes that the only way to assess the relative contributions of different 'sources of variance' to police personality is by way of adequately controlled longitudinal research, yet to be carried out.

Lefkowitz' conclusion that differences found between police and other groups are, overall, evaluatively neutral, and that those differences may be strongly a function of the sub-populations from which most recruitment takes place, would seem to suggest that personality, like stress, is not an issue of unusual importance in police assessment. Even in the event of undesirable aspects of police personality being clearly identified and linked to self-selection, there appears to be no selection technology that would enable valid screening-out. The problems of personality test faking are well known, and the general invalidity of personality tests in selection contexts is well established (e.g. Schmitt, Gooding, Noe and Kirsch, 1984). And in the specific police context Sherrid (1979) has concluded that no effective method of screening out undesirable recruits has yet been demonstrated.

Research Questions

With reference to the specific topic of Home Office EIs used to select for the police Special Course three main questions, forming the basis of the research to be reported in the next chapter, may now be stated. Taken together these questions are designed to pinpoint issues of straightforward practical importance as well as issues of importance in the wider AC context, with considerable overlap. The questions are:

1. Do EIs have external predictive validity? In other words, do EI measures relate to later training and job criteria? EIs are now in their 24th year of operation without a satisfactory answer to this question. The only methodologically adequate investigation so far (Mays, 1972) is inconclusive due to the very small sample size involved. The validity of Home Office police EIs is also of wider interest given the general dearth of criterion-related research into police ACs. It would, however, be difficult to suggest a directional hypothesis as to EI validity, since the validity demonstrated for ACs across a variety of contexts has to be set against the lack of job-targeted design in the setting up of this particular AC and Mays' somewhat negative findings. Validity will be assessed in terms of EI overall Final Mark and marks/scores from the component tests and exercises. A related issue to be addressed is the efficiency with which external criteria are predicted. An attempt will be made to determine whether optimum validity could be achieved with some subset of EI measures. Given the lengthy and costly nature of EIs, this issue is clearly of practical significance. It is also of importance in the wider AC field since, as shown in chapter 5, AC efficiency is very much an under-researched area. The available research evidence suggests that redundancy may possibly be considerable in many ACs, and this seems a reasonable hypothesis for the

AC under examination.

2. How internally valid are EIs? In other words, how much use appears to be made of information from the various component assessment techniques in final EI decision making? Also how does this information use relate to the validity of the component techniques as found for external criteria? A main issue of practical significance here is the efficiency with which expensively obtained information is processed. The investigation of decision making is also of considerable importance in the wider AC context since, as far as this writer is aware, virtually nothing is known about the bases underlying subjective combination of information derived from component AC techniques. Given this lack of knowledge it is not really possible to state any specific hypotheses.

3. How does validity of the clinical/judgemental method of combining EI information compare with validity of a mechanical alternative? As shown in chapter two, despite the virtually universal use of the clinical approach in ACs, very few comparisons with mechanical/statistical alternatives have been undertaken; more are needed. The few comparisons which have been carried out, when set against a wider background of clinical versus mechanical/statistical research, would appear to suggest the hypothesis that mechanical combination of EI data

should prove as or more valid than the clinical/judgemental method in use.

Chapter 7 - THE PRESENT RESEARCH

The three research questions identified at the end of the last chapter have been addressed in three overlapping studies reported in this chapter. Analyses have been done separately for the two groups for whom the Special Course is designed: (1) internal Police candidates, called 'Special Course candidates'; and (2) Graduate Entry candidates.

A description of the general research method, including the measures and sampling strategy employed, is followed by a study by study account of the analyses and results.

METHOD

To recap, the research questions addressed were:

(1) Do EIs have external predictive validity? This question was investigated by **Study 1.**

(2) How internally valid are EIs? This question was investigated by **Study 2.**

(3) How does validity of the clinical/judgemental method of combining EI information compare with validity of a mechanical alternative? This question was investigated by **Study 3.**

Study 1: EI predictive validity was investigated with regressions of three groups of criteria - training grades awarded by course tutors, job performance as assessed by supervisory ratings, and advancement as measured by rank attained - on measures obtained at EI. As a preliminary to this, two other analyses were performed. Firstly, supervisory ratings were factor analysed in order to condense the information they contained and to make some assessment of any underlying dimensionality in the data. And secondly, criterion measures were intercorrelated to discover whether there was any redundancy among them and to explore their interrelationships.

Study 2: EI internal validity was investigated with regressions of EI Final Mark on the scores obtained from EI component measures - tests, exercises, interviews, and peer nominations.

Study 3: Judgemental versus mechanical combination of EI information was investigated by comparing the predictive validities obtained for EI Final mark with those obtained for unit-weighted composites of EI component measures.

Measures

Measures fall into four categories: Extended Interview, training, rank and supervisory ratings. All measures are on at least ordinal scales.

EXTENDED INTERVIEW

The following measures are all on an eleven point scale: Group Discussion average mark, Committee Exercise (Chairman) average mark, Committee Exercise (Member) average mark, Written Appreciation average mark, Drafting Test average mark, Chairman's Interview mark, Service Member's Interview mark and Non-Service Member's Interview mark. In each case the average mark is the mean of the three assessors' marks. Raven's Advanced Progressive Matrices is scored 0 to 36. The Abstractions and General Information Test (GIT), while retaining the same basic format throughout the period under investigation, underwent some revision with the result that raw scores for different years are not comparable. For purposes of analysis, the normative conversions of the raw scores on standard seven point scales have been used. The peer nomination measures, Senior Officer and Holiday Companion, have been converted to a numeric scale with one as the lowest rank and 6 as the highest. These scores are clearly subject to fluctuations in group composition, which will tend to reduce reliability. The Final Mark is on a seven point scale, with four possible marks for successful candidates. In 1974 there was a change in the wording used to describe some of the scale points (though the alphanumeric notation remained the same), and the 1972-1973 standard deviation for successful candidates of 0.75 compares with a 1974-1977 standard deviation of 0.47. For this reason where correlations are quoted for Final Mark which include pre- and post- 1974 data, the figure given is an average coefficient for the two periods, computed as described

by Guilford (1956). Similarly where corrected coefficients involving Final Mark are quoted, the corrections been have performed separately for the two periods and the resulting coefficients averaged.

TRAINING

Measures are Professional Studies grade, Academic Studies grade and Overall Performance grade. Professional and Academic grades, which reflect performance in course work and examinations, are on five-point scales. Overall Performance, which is a general assessment arrived at by college tutors, is graded on a four point scale for those successfully completing training. The few who may be counted as failures during the period under investigation all left the course before its completion, thus not receiving an overall grade. For purposes of analysis an additional scale point has been added for these failures, giving a five-point scale.

RANK

The measure Rank is rank attained at June 1984. Ranks from Constable to Chief Constable or equivalent have been coded 1 to 9. Clearly an important determinant of rank is length of service, and some of the samples in this study span several years. If correlations of EI and training measures with rank were to be computed without taking account of this, the coefficients would be deflated. To counter this, where correlation coefficients involving rank and earlier measures are quoted, they are in every case averages of coefficients computed

separately for each set of EIs (e.g. Graduate Entry 1972; Special Course, 1976). Average coefficients were computed as described by Guilford (1956).

SUPERVISORY RATINGS

All British police forces operate separate performance appraisal systems, so, in order to obtain supervisory ratings in a common format, it was necessary to construct a performance appraisal measure. A letter (dated 28th January 1983) was sent to all 43 police forces in England and Wales asking for copies of performance appraisal forms for ranks of Inspector and above. All 43 forces eventually replied, though three replies were too late to be included in subsequent analysis. Forms from all but three forces included rating scales relating to sets of performance dimensions. Some forces used different forms for different ranks, but typically these different forms included a core subset of dimensions supplemented with one or two dimensions used for specific ranks or groups of ranks. A policy of combining dimensions for different ranks to produce one research form therefore seemed justifiable. The strategy adopted was to make a list of all dimensions included in the forms produced by any of the forces. Where different forms of words were used to label what appeared on investigation to be substantially the same dimension, only one note was made. Dimensions of the following types were then eliminated:

- Dimensions which appeared diffuse, i.e. dimensions spanning a group of more specific and conceptually separate characteristics which could most usefully be included in their own right. Examples are 'intelligence', 'personal relations', and 'social attributes'.

- Dimensions which lacked an adequate definition, i.e. were stated without clarification on appraisal forms and for which no adequate definition was apparent, e.g. 'personal capacity for sound discrimination', 'crime intelligence'.

This reduced the number of dimensions to 29. In order to arrive at a subset which would be representative of police appraisal practice whilst sufficiently small to facilitate practical research use, a points system was used. Appraisal forms for each force were analysed individually in order to determine which of the 29 dimensions were referred to. If a clear reference was made to a particular dimension, that dimension was given a score of 1. If a particular dimension was implied rather than directly referred to, it was given a score of 0.5. Thus each dimension obtained a score of 1, 0.5, or 0 from each force. Points were then totalled for each dimension. The choice of dimensions then became a matter of applying a cut-off score above which all dimensions would be included. A cut-off of 12 points was chosen. This had two consequences. Firstly it left 17 dimensions for inclusion. An anlysis of the numbers of dimensions included in

the appraisal forms of individual forces indicated a range of between 4 and 23 dimensions, with a median of 14. Thus a form with 17 dimensions was at the upper end of normal expectations. Secondly the two dimensions immediately below the cut-off were 'loyalty' and 'interest in work', dimensions which do not directly describe performance. The 17 dimensions remaining formed the basis of the research form. Scales for rating overall performance and potential were added. Five point scales for performance and a four point scale for potential were adopted as being in line with police practice and with research evidence which suggests little improvement in scale reliability where more than five scale points are used, while reliability drops with three categories or less (Landy and Farr, 1980). A mixture of descriptive and percentage anchors was used. The final research form is included as the Appendix. The form was sent to forces to obtain follow-up information on police officers who had been successful EI candidates some years previously.

Samples and Sampling Strategy

Seven samples of candidates were used in this research, and details of each sample are given in Table 3. The reasons for choosing the various samples are now outlined. After this there is a note on wastage from follow-up samples.

SAMPLE 1

Detailed records of performance at Extended Interview go back to 1972. For the years prior to that, only EI Final Mark is recorded. Full details of the scales on which Final Marks were based prior to 1965 are not available. But for the period 1965

Table 3 - Research Samples - Description and Measures

	SAMPLES 1-7						
	1	2	3	4	5	6	7
Description							
. Candidate group: Special Course=SC or Graduate Entry=GE	SC	SC	GE	SC	SC	GE	GE
. All candidates (All) or successfuls only (Succ)	Succ	Succ	Succ	All	All	All	All
. Years of EIs	1965-1970	1972-1977	1972-1976	1972	1976	1972	1976
. Number in sample	223	157	86	126	73	86	109
. Sex (M and F)	215M 8F	146M 11F	68M 18F	122M 4F	71M 2F	79M 7F	94M 15F
. Age range	21-35	21-33	19-30	22-30	20-31	20-29	20-30
. Mean age	25.5	25.7	22.7	24.8	26.0	22.6	22.5
Measures							
. EI component measures	No	Yes	Yes	Yes	Yes	Yes	Yes
. EI Final Mark	Yes	Yes	Yes	Yes	Yes	Yes	Yes
. Supervisory Ratings	No	Yes	Yes	No	No	No	No
. Training grades	No	Yes	Yes	No	No	No	No
. Rank at June 1984	Yes	Yes	Yes	No	No	No	No

Note. Samples 2 and 3 include successful candidates from samples 4, 5, 6, and 7.

to 1970 (when the scale was changed) Final Mark was assessed on a constant scale. Although the paucity of data limits possibilities of analysis, it was felt that a sample from these years would be useful to test Final Mark validity for the rank criterion; major differences in rank take some years to emerge. Sample 1 consists of all successful Special Course candidates between 1965 and 1970 who were still serving police officers in June 1984.

SAMPLES 2 & 3

Samples 2 and 3 were central to testing EI external predictive validity. They span the years 1972 (from when full EI data is available) to 1976 (Sample 3) and 1977 (Sample 2). 1976/7 was chosen to allow a reasonable time between EI and the collection of follow-up data, such that some differences in the Rank criterion would have emerged. The one year difference in the termination of the sampling periods is to reflect the fact that Graduate Entrants (Sample 3) had to serve as Constables prior to attending the Special Course; consequently their progress was slower than that of Special Course candidates (Sample 2).

SAMPLES 4,5,6, & 7

These samples of successful and unsuccessful candidates were used to assess EI internal validity. The years 1972 and 1976 were chosen so that internal validity findings could be set in the context of external validity evidence. Four samples were used so as to give a comparison between Special Course and Graduate Entry candidates for each year.

WASTAGE FROM FOLLOW-UP SAMPLES

For the detailed follow-up of samples 2 and 3, a count was made of wastage and types of wastage:

Sample 2

By 1984, 12 of the original 157 successful Special Course candidates had voluntarily resigned from the police service or had retired due to ill health, 4 had been forced to resign or had resigned following failure on part of the Special Course scheme, and one had been dismissed. Of the remaining 140 who were still serving police officers, one had failed to complete the Special Course, three had failed to obtain the normal promotion to Inspector, and one had been demoted from Inspector to Constable.

Sample 3

By 1984, 22 of the original 86 Graduate Entrants had voluntarily resigned or had retired due to ill health, including 9 of the 18 women, and 4 had been forced to resign or had resigned following failure on part of the Special Course scheme. Of the remaining 60 who were still serving police officers, one had voluntarily withdrawn from the Special Course scheme and 6 had dropped out of the scheme through failing the promotion exam to Sergeant.

ANALYSES AND RESULTS

Analyses and results are now reported for each of the three studies which have been outlined.

Study 1: EI Predictive Validity

The general plan was to correlate EI measures with training and job criteria. Data for the two candidate groups - Graduate Entrants and Special Course candidates - were analysed separately. As a preliminary to this, two analyses of different kinds were undertaken. Firstly, supervisory ratings were factor analysed in order to condense the information they contained and to make some assessment of any underlying dimensionality. And secondly, criterion measures were intercorrelated to discover whether there was any redundancy among them and to explore their interrelationships. Since these two analyses concerned criteria rather than candidates, it was seen as legitimate to combine data for Graduate Entrants and Special Course candidates.

Figure 1 - Factor Matrix of Supervisory Ratings (N=180)

PERFORMANCE DIMENSION	FACTOR 1	FACTOR 2
Organization of work	0.78	-0.15
Judgement	0.77	0.04
Foresight	0.76	-0.05
Oral expression	0.58	0.08
Written work	0.59	0.03
Professional knowledge	0.61	-0.15
Application of professional knowledge	0.78	0.00
Manner with the public	0.60	0.38
Assessment of people	0.61	0.28
Relationships with colleagues	0.66	0.25
Reliability under pressure	0.76	-0.27
Initiative	0.77	-0.19
Effort and vitality	0.65	-0.35
Acceptance of responsibility	0.74	-0.24
Turn-out (i.e. general appearance)	0.42	0.27
Leadership	0.81	0.12
Supervisory ability	0.78	0.16

Note Full definitions of the dimensions are in the Appendix.

FACTOR ANALYSIS OF SUPERVISORY RATINGS

The 17 performance dimensions from the supervisory rating form (but not the ratings of Overall Performance and Potential) were factor analysed using Statistical Package for the Social Sciences (Nie et al., 1975). Subjects were all those in samples 2 and 3 for whom forces returned supervisory ratings and who were serving at the rank of Inspector or above (N=180). Principal factoring with iteration yielded two factors with eigenvalues greater than one. These factors accounted for 51% and 7% of the variance respectively. The unrotated factor matrix is shown in figure 1. Factor 1 is clearly a general performance or global evaluation factor, with no loading less than 0.4 and loadings of 0.6 or above on 14 of the 17 dimensions. Looking at factor 2 and considering loadings of above 0.2, what emerges is a contrast between positively loaded dimensions with an interpersonal skills content (manner with the public, assessment of people, relationships with colleagues, turn-out) and negatively loaded dimensions which seem to focus around task motivation (reliability under pressure, effort and vitality and acceptance of responsibility). So a tentative label for factor 2 is "interpersonal skills vs. task motivation". In subsequent analyses, scores on these two factors were substituted for the original 17 performance dimensions leaving a total of four measures from the supervisory rating form: Factor 1, Factor 2, Potential and Overall Performance. Eight policemen for whom rating forms were returned but who were serving at a rank below Inspector were excluded from the factor analysis and from

subsequent analyses using performance criteria. The reason for this was that the Special Course is designed to train for Inspector level and above and performance at this level seems the most appropriate criterion for a validity investigation. Also the form was designed for ranks of Inspector and above, and some of those completing the form for lower ranks experienced difficulty with some scales.

Figure 2 - Pearson Correlation Coefficients - Training and Job Criteria

	1	2	3	4	5	6	7
Training Criteria							
1. Academic studies							
2. Professional studies	0.27** N=193						
3. Overall performance	0.53** N=204	0.43** N=193					
Job Criteria							
4. Rank	0.16* N=187	0.19* N=176	0.29** N=190				
5. Factor 1	0.06 N=173	0.13 N=164	0.19* N=175	0.24** N=178			
6. Factor 2	-0.14 N=173	-0.12 N=164	-0.08 N=175	-0.12 N=178	-0.02 N=180		
7. Potential	0.07 N=172	0.16* N=163	0.24** N=174	0.18* N=177	0.66** N=179	-0.05 N=179	
8. Overall performance	0.13 N=172	0.14 N=163	0.23** N=174	0.18* N=177	0.84** N=179	-0.06 N=179	0.63** N=178

* - $p>0.05$; ** - $p>0.01$ (two-tailed)

INTERCORRELATIONS OF CRITERIA

Next training and job criteria were intercorrelated for samples 2 and 3 combined, and the results are shown in figure 2. The time gap between training (1973 to 1981) and job (1984) measures ranges from three to eleven years. A correlation of 0.27 suggests that ratings of performance on the Academic and Professional Studies parts of the Special Course are relatively independent, though, as might be expected, both are moderately correlated with the final rating of Overall Performance in training. Overall Performance (training) emerges as moderately predictive of later Rank attained (r=0.29). This training measure also significantly predicts later job performance ratings - Factor 1, Overall Performance (job), and Potential - though at a relatively low level. Factor 1 and Overall Performance (job) appear to be very similar measures as judged by their patterns of correlations with other measures and their intercorrelation of 0.84. Consideration was given to dropping one of them from subsequent analyses. However, there are arguments for the use of each measure, Factor 1 being derived from a number of systematic and relatively specific ratings while Overall Performance represents a direct rather than inferred global assessment. It was therefore decided to proceed with both. The Potential rating's pattern of correlations with other measures is similar to the patterns for Factor 1 and the Overall Performance (job) rating, though correlations of 0.66 with Overall Performance (indicating 44% of common variance) and 0.63 with Factor 1 (indicating 40% of common variance) demonstrate a degree of

independence. Rank attained is significantly related to the performance/potential ratings, though at a relatively low level. The correlations of Factor 2 with the other measures are all non-significant, which may be partly explained by the fact that Factor 2 is orthogonal to global performance as represented by Factor 1. It appears to be qualitatively different from the other criterion measures.

CORRELATIONS OF EI MEASURES AND TRAINING CRITERIA

Correlations of EI measures with training criteria for samples 2 (Special Course successful candidates) and 3 (Graduate Entrants)

Figure 3 - Pearson Correlation Coefficients - EI Measures with Training Criteria - Sample 2 (Special Course)

	Academic Studies (N=147-151)	Professional Studies (N=147-151)	Overall Performance (N=152-156)
Group Exercises			
. Group Discussion	-0.01	-0.10	0.18*
. Committee Chairman	0.02	0.06	0.11
. Committee Member	0.14	-0.02	0.22**
Written Exercises			
. Written Appreciation	0.11	0.05	0.01
. Drafting Test	-0.05	0.01	0.07
Pencil and Paper Tests			
. Matrices	0.05	-0.11	-0.02
. Abstractions	0.11	0.13	-0.08
. GIT	0.14	0.07	0.09
Peer Nominations			
. Senior Officer	0.00	0.02	0.17*
. Holiday Companion	0.00	0.01	-0.07
Interviews			
. Chairman	-0.01	-0.02	0.09
. Service Member	-0.06	-0.04	0.03
. Non-Service Member	-0.03	-0.02	-0.04
Final Mark	0.00	0.04	0.11

* - $p<0.05$; ** - $p<0.01$ (two-tailed)

are shown in figures 3 and 4 respectively. For neither sample do EI measures significantly predict performance on the academic and professional studies parts of the course, though correlations with Overall Performance are somewhat higher. For Sample 2, three measures: Group Discussion, Committee Member and Senior Officer correlate significantly with Overall Performance, though Final Mark does not. For Sample 3 only one measure, Senior Officer, correlates significantly with Overall Performance, and this is also the only measure which consistently predicts training performance across both samples. Generally the EI measures appear not to be predictive of training performance for Sample 3. Even taking account of range restriction and the small sample size, the fact that approximately half (23 out of 42) of the coefficients are negative suggests a pattern of negligible relationships overall. However, successful Special Course candidates begin training within six months of EI assessment whereas Graduate Entrants normally attend the Special Course between three and five years later, so results for samples 2 and 3 are not directly comparable.

CORRELATIONS OF EI MEASURES AND JOB CRITERIA - SAMPLE 2

Correlations of EI measures with job criteria for sample 2 (Special Course successful candidates) are shown in figure 5. Two EI component measures (Committee Chairman and Drafting Test) are found to correlate significantly with job performance ratings (Factor 1 and Overall Performance) obtained 7 to 12 years later, while one further measure (Non-Service Member's Interview) also

correlates significantly with Overall Performance. This subset of EI component measures is different from the subset of measures which emerged as predictive of training performance. Final Mark correlates significantly at 0.18 (0.33 after correction for direct range restriction) with Overall Performance and at a comparable though non-significant level with Factor 1. The General Information Test correlates significantly and negatively with Potential, though this may, in part, be an artefact resulting from the positive correlation of age with GIT grades ($r=0.37$, $p<0.001$) and the negative though non-significant

Figure 4 - Pearson Correlation Coefficients - EI Measures with Training Criteria - Sample 3 (Graduate Entry)

	Academic Studies (N=51-53)	Professional Studies (N=40-42)	Overall Performance (N=52-54)
Group Exercises			
. Group Discussion	-0.03	-0.08	0.04
. Committee Chairman	-0.09	-0.08	-0.24
. Committee Member	-0.17	-0.22	-0.04
Written Exercises			
. Written Appreciation	-0.07	0.04	-0.06
. Drafting Test	0.06	0.03	0.14
Pencil and Paper Tests			
. Matrices	-0.15	0.00	-0.14
. Abstractions	-0.09	-0.02	0.07
. GIT	0.17	0.06	0.17
Peer Nominations			
. Senior Officer	0.15	0.07	0.28*
. Holiday Companion	0.12	0.09	0.20
Interviews			
. Chairman	0.10	-0.25	-0.02
. Service Member	0.08	-0.19	-0.10
. Non-Service Member	-0.05	-0.08	-0.09
Final Mark	-0.07	-0.20	0.07

* - $p<0.05$; ** - $p<0.01$ (two-tailed)

correlation of age and rated Potential (r=-0.13). The partial correlation of the GIT and Potential controlling for age is non-significant (r=-0.15). Potential does not correlate significantly with any of the other EI measures, though the general pattern of coefficients bears some similarity to those for Factor 1 and Overall Performance. Factor 2 correlates significantly with the Holiday Companion peer nomination (positively) and the Chairman's interview mark (negatively). Correlations of EI measures with Rank are for the most part negligible, though the Senior Officer peer nomination correlates

Figure 5 - Pearson Correlation Coefficients - EI Measures with Job Criteria - Sample 2 (Special Course)

	Rank (N=135 -139)	Fac. 1 (N=124 -128)	Fac. 2 (N=124 -128)	Potential (N=123 -127)	Overall Perf. (N=123 -127
Group Exercises					
. Group Discussion	0.03	0.07	-0.15	0.04	0.10
. Committee Chairman	-0.06	0.20*	0.05	0.17	0.19*
. Committee Member	-0.02	0.07	-0.02	0.11	0.14
Written Exercises					
. Written Appreciation	-0.03	-0.03	-0.05	-0.12	0.04
. Drafting Test	0.05	0.18*	0.05	0.08	0.20*
Pencil and Paper Tests					
. Matrices	-0.01	-0.05	0.10	0.06	0.01
. Abstractions	0.00	-0.01	0.08	-0.07	0.05
. GIT	-0.04	0.00	-0.13	-0.18*	-0.07
Peer Nominations					
. Senior Officer	0.30**	-0.03	-0.11	0.06	-0.01
. Holiday Companion	0.02	-0.10	0.19*	-0.11	-0.05
Interviews					
. Chairman	-0.06	0.01	-0.18*	0.01	0.02
. Service Member	-0.04	0.01	-0.14	0.00	0.05
. Non-Service Member	0.03	0.17	-0.12	0.08	0.20*
Final Mark	0.00	0.15	-0.17	0.11	0.18*

* - $p<0.05$; ** - $p<0.01$ (two-tailed)

strongly and highly significantly with this criterion.

CORRELATION OF EI FINAL MARK AND RANK - SAMPLE 1

It might be argued that the lack of relationship between EI measures and Rank for Sample 2 is partly a function of the time necessary for large differences in rank to emerge; no individual in this sample had progressed beyond Superintendent, which is two ranks above Inspector and four ranks below Chief Constable. As an additional test, Final Mark and Rank were correlated for Sample 1 (Special Course successful candidates 1965 to 1970) of whom 34% were above the rank of Superintendent at follow-up including three who had reached Chief Constable. The coefficient obtained was a non-significant 0.10 (N=223). This 14 to 19 year follow-up provides confirmation of the finding of negligible validity for the rank criterion.

MULTIPLE REGRESSION AND PREDICTIVE EFFICIENCY - SAMPLE 2

As a means of assessing, for Sample 2, the overall predictive efficiency of EI measures, multiple regression analyses were undertaken for the four main criteria which had been found to be predictable using bivariate correlations - Overall Performance (training), Overall Performance (job), Factor 1, and Rank. The significant predictors were entered stepwise into multiple regressions provided they contributed significantly ($p<0.05$) to the total variance explained. Since only Senior Officer significantly predicted Rank, its correlation of 0.30 with that criterion is taken as the R for the full set of EI measures. The results of the multiple regressions for the three other criteria

are shown in figure 6. It may be seen that R=0.28 for Overall Performance (training) and Overall Performance (job), and R=0.27 for Factor 1. So taking the findings together 7% to 9% of the variance in the four criterion measures is found to be predictable from EI measures, and all of this variance may be accounted for by a subset of four EI measures comprising Committee Chairman, Committee Member, Drafting Test and Senior Officer.

Figure 6 - Stepwise Multiple Regressions of Overall Training Performance, Overall Job Performance and Factor 1 on Significant EI Predictors - Sample 2

Overall Performance (Training)

	R
Committee Member	0.22
Senior Officer	0.28
Group Discussion*	

Overall Performance (Job)

	R
Drafting Test	0.20
Committee Chairman	0.28
NSM's Interview*	
Final Mark*	

Factor 1

	R
Committee Chairman	0.20
Drafting Test	0.27

* These measures were in the predictor sets but were not entered into the regression equations as they would not have contributed significantly ($p<0.05$) to the variance.

CORRELATIONS OF EI MEASURES AND JOB CRITERIA - SAMPLE 3

Correlations of EI measures and job criteria for Sample 3 (Graduate Entrants) are shown in figure 7. All the coefficients are non-significant. These negative findings should be interpreted cautiously, however, with respect to relatively small sample size (N=50-58) and the presence of range restriction.

Figure 7 - Pearson Correlation Coefficients - EI Measures with Job Criteria - Sample 3 (Graduate Entry)

	Rank (N=56 -58)	Fac. 1 (N=50 -52)	Fac. 2 (N=50 -52)	Potential (N=50 -52)	Overall Perf. (N=50 -52)
Group Exercises					
. Group Discussion	-0.01	-0.06	-0.08	-0.07	-0.02
. Committee Chairman	0.19	0.02	0.00	0.14	0.00
. Committee Member	0.04	0.23	-0.03	0.08	0.11
Written Exercises					
. Written Appreciation	-0.15	-0.12	-0.21	-0.07	-0.09
. Drafting Test	0.13	0.10	-0.13	0.16	0.25
Pencil and Paper Tests					
. Matrices	-0.09	0.16	-0.07	0.23	0.10
. Abstractions	-0.11	0.11	-0.17	0.10	0.10
. GIT	0.07	0.06	-0.16	0.16	0.04
Peer Nominations					
. Senior Officer	0.16	0.22	0.01	0.23	0.25
. Holiday Companion	0.13	0.04	0.19	0.13	0.08
Interviews					
. Chairman	-0.06	0.17	0.18	0.01	0.18
. Service Member	-0.05	0.00	0.17	-0.08	-0.01
. Non-Service Member	-0.02	-0.08	-0.05	-0.05	-0.04
Final Mark	0.05	0.07	0.00	0.21	0.14

* - $p<0.05$; ** - $p<0.01$ (two-tailed)

SUMMARY OF STUDY 1 RESULTS

The basic question addressed by Study 1: Do EIs have external predictive validity? may now be answered. For Special Course successful candidates, EI measures do have some predictive validity for training and later job performance ratings, though the overall Final Mark predicts only Overall Performance (job) and the subsets of EI component measures which predict training and job ratings are different. Excepting the Senior Officer nomination, EI measures generally do not seem to predict rank attained. Senior Officer nomination is moderately predictive of rank as well as of training performance though not job performance ratings. As regards Graduate Entry EIs, the findings generally do not provide any evidence for validity, though small sample size and restriction of range should be borne in mind in interpreting this. It is worth noting, however, that Senior Officer nomination again emerged as a significant predictor of training performance.

Study 2: EI Internal Validity

EI internal validity was investigated by attempting to answer the question: How much use appears to be made of information from the various component techniques in final EI decision making? This question was investigated with regressions of EI Final Mark on the scores obtained from EI component measures. Analyses and results are now described.

CORRELATIONS OF COMPONENT MEASURES WITH EI FINAL MARK

The use made of information from component techniques in final EI decision making was investigated with correlation and multiple regression for samples 4, 5, 6 and 7.

Correlations of all EI component measures with Final Mark for these four samples are shown in figure 8. It is clear that measures derived from group and written exercises and particularly from interviews are strongly predictive of EI Final Mark. For the Graduate Entry samples, single interview marks

Figure 8 - Pearson Correlation Coefficients - EI Component Measures with EI Final Mark - Samples 4 - 7

	Final Mark			
	Sample 4 (Special Course 1972 N=126)	Sample 5 (Special Course 1976 N=73)	Sample 6 (Graduate Entry 1972 N=86)	Sample 7 (Graduate Entry 1976 N=109)
Group Exercises				
. Group Discussion	0.45**	0.30**	0.52**	0.55**
. Committee Chairman	0.53**	0.52**	0.64**	0.57**
. Committee Member	0.56**	0.43**	0.71**	0.59**
Written Exercises				
. Written Appreciation	0.41**	0.34**	0.59**	0.40**
. Drafting Test	0.36**	0.55**	0.53**	0.28**
Pencil and Paper Tests				
. Matrices	0.23*	0.13	0.09	0.00
. Abstractions	0.18*	0.35**	0.11	0.03
. GIT	0.20*	0.39**	0.30**	0.19
Peer Nominations				
. Senior Officer	0.19*	0.03	0.36**	0.36**
. Holiday Companion	0.01	0.20	0.05	0.17
Interviews				
. Chairman	0.57**	0.65**	0.80**	0.77**
. Service Member	0.58**	0.68**	0.78**	0.77**
. Non-Service Member	0.57**	0.62**	0.84**	0.72**

* - $p<0.05$; ** - $p<0.01$ (two-tailed)

account for between 52% and 71% of the variance in Final Mark. Pencil and paper test measures and peer nominations are less clearly related to Final Mark, though Senior Officer and General Information Test do appear to predict it fairly consistently. Clearly, however, correlation does not prove causality. Another problem in interpreting this evidence is the likelihood that EI component measures are not completely independent. All the subjectively derived marks are awarded by the same team of assessors, and the possibility that marks awarded for performance in specific exercises are to some extent affected by previous observations of candidate behaviour cannot be ruled out. Interviews are particularly suspect as they are designed to add to the information already gathered about a candidate, with the result that EI performance up to the point of the interview is almost inevitably taken into account in arriving at an interview mark.

Figure 9 - Pearson Correlation Coefficients - Interviews with EI Final Mark - Samples 4 - 7

	Final Mark			
	Sample 4 (Special Course 1972 N=126)	Sample 5 (Special Course 1976 N=73)	Sample 6 (Graduate Entry 1972 N=86)	Sample 7 (Graduate Entry 1976 N=109)
First Interview	0.58	0.54	0.81	0.71
Second Interview	0.57	0.73	0.83	0.79

Note All coefficients significant at $p<0.001$ (two-tailed)

ANALYSES ALLOWING FOR INTERDEPENDENCE OF MEASURES

In an attempt to deal with this possible lack of independence, it was decided to carry out hierarchical multiple regression analyses, adding EI measures into the equations in the order in which assessors produced/received them. In this way it was hoped to make an assessment of the extent to which each measure in sequence makes an additional unique contribution to the prediction of the Final Mark, and consequently of the extent of information redundancy in the decision making process. The main sequence of EI procedures was the same for all candidates and for all four samples 4-7. However for half the candidates the NSM's interview preceded the service interview (Chairman and Service Member), while for the other half the order was reversed. In order to deal with this, two new measures were constructed - 'First Interview' and 'Second Interview' - consisting of the marks awarded on the basis of the first and second interview regardless of which assessor conducted the interview. For these measures to be compiled it was necessary first to derive a single composite mark for the service interview; this was done by averaging the Chairman's and Service Member's marks. Before going on to look at the results of multiple regression analyses, it is of interest first to compare the correlations of First Interview and Second Interview with Final Mark as an indication of whether interview marks are influenced by earlier parts of the procedure. The results of this analysis are shown in figure 9. The trend in the correlations, except for Sample 4 where the difference is negligible, is in the direction of higher

correlations for the later interviews. This would seem to support the notion that interview marks are to some extent summary evaluations of information obtained up to that point.

The results of hierarchical multiple regression analyses for samples 4-7 are shown in figure 10. The top to bottom order (Group Discussion to GIT) indicates the order in which assessors produced/received measures and the order in which measures were entered into the regression equations. The results clarify the picture emerging from the bivariate correlations presented in

Figure 10 - Hierarchical Multiple Regression of EI Final Mark on all EI Component Measures Entered in Sequence of EI Procedure - Samples 4 - 7

	R Square Change			
	Sample 4 (Special Course 1972 N=126)	Sample 5 (Special Course 1976 N=73)	Sample 6 (Graduate Entry 1972 N=86)	Sample 7 (Graduate Entry 1976 N=109)
Group Discussion	0.20**	0.09**	0.28**	0.30**
Written Appreciation	0.08**	0.09**	0.22**	0.05**
First Interview	0.14**	0.19**	0.24**	0.22**
Matrices	0.01	0.00	0.01*	0.00
Drafting Test	0.03**	0.11**	0.03**	0.00
Committee Chairman)	0.04**	0.07**	0.04**	0.02*
Committee Member)	0.01	0.01	0.00	0.00
Abstractions	0.00	0.01	0.01*	0.01
Second Interview	0.04**	0.17**	0.01	0.08**
Senior Officer)	0.01	0.00	0.00	0.00
Holiday Companion)	0.00	0.03**	0.00	0.00
GIT)	0.00	0.00	0.00	0.00
	R=0.75**	R=0.89**	R=0.92**	R=0.83**

Note Brackets indicate measures which were obtained by the assessors approximately simultaneously.
* - $p<0.05$; ** - $p<0.01$

figure 8. It is apparent that for all four samples most of the predictable variance in Final Mark is accounted for by the group and written exercise and interview measures. Interviews seem to make a particularly important contribution as evidenced by the amount of variance accounted for by First Interview and the fact that, for three of the four samples, Second Interview, though late in the EI procedure, makes a significant addition to the total variance accounted for. The pencil and paper test and peer nomination measures, however, tend to contribute little or no unique variance. In interpreting these findings it should be borne in mind that they are partly dependent on the order of the EI procedures; it may be that if parts of the procedure were changed round, for example conducting the Group Discussion after the Committee Exercise, the picture would alter somewhat. Ideally one would carry out this type of regression analysis alongside experimental changes of order.

SUMMARY OF STUDY 2 RESULTS

In summary, though it is impossible to be certain of causality, there seems good reason to think that **assessors pay attention to group exercise, written exercise and interview performance in arriving at Final Mark, but that little account is taken of pencil and paper test and peer nomination measures.**

Study 3: Judgemental vs. Mechanical Combination of EI Information

Consideration was given to comparing the validities of EI Final Mark for different criteria against validities obtained using regression-weighted composites of EI measures. This would be reasonable if there were only one criterion of interest. But there are several important criteria which are predicted by different subsets of EI measures, and it is clear that different regression weights would be necessary for optimum prediction of each. A comparison of Final Mark which is a 'one-off' combination of information with several different weightings/combinations would not represent a fair judgemental versus mechanical test. It was decided instead to compare Final Mark with a unit-weighted composite of EI measures.

Figure 11 - Pearson Correlation Coefficients - Final Mark and a Unit-Weighted Composite with Five Criteria - Sample 2 (Special Course)

	Overall Perf. (Training)	Rank	Factor 1	Potential	Overall Perf. (Job)
Unit-Weighted Composite	0.26** N=155	0.13 N=138	0.19* N=127	0.18* N=126	0.23** N=126
Final Mark	0.11 N=156	0.00 N=139	0.15 N=128	0.11 N=127	0.18* N=127

Note The unit-weighted composite comprised four measures: Committee Chairman, Committee Member, Drafting Test, and Senior Officer.
* - $p<0.05$; ** - $p<0.01$ (two-tailed)

JUDGEMENTAL vs. MECHANICAL - SAMPLE 2

For Sample 2 four measures - Committee Chairman, Committee Member, Drafting Test and Senior Officer - had been shown to account for all the predictable variance in four criterion measures. It was therefore decided to form a unit-weighted composite from these measures. The composite consisted simply of the sum of the standard scores on the four measures. In figure 11, correlations of the composite and Final Mark with the five main criteria - Overall Performance (training), Rank, Factor 1, Potential and Overall Performance (job) - are compared. It may be seen that for each criterion the coefficient for the composite is higher than the corresponding coefficient for Final Mark. Also four of the five coefficients for the unit-weighted composite are significant compared with one significant coefficient for Final Mark. It is also worthy of note that the composite significantly predicts Potential whereas, as can be seen from figure 5, none of the individual EI measures correlate significantly and positively with this criterion.

CAUTIONARY REMARKS

On the face of things, these findings would appear to support the hypothesis that mechanical combination of EI information would prove as or more valid than judgemental combination. However the findings should be interpreted cautiously for two reasons. Firstly correlations with Final Mark are likely to be affected more by range restriction than correlations with the unit-weighted composite (i.e. direct as opposed to indirect

restriction). As a test of this differential effect, the corrected coefficients for Overall Performance (job) were compared. As has already been reported, the 0.18 coefficient for Final Mark rises to 0.33 after correction for direct range restriction. The comparable correlation of 0.23 for the unit-weighted composite rises to 0.34 after correction for indirect range restriction. It was not seen as legitimate to conduct such comparisons for the other criteria since this would have involved correcting coefficients which were not significantly different from zero. Nevertheless this single comparison is quite informative in suggesting that while range restriction differentially affects Final Mark and the unit-weighted composite, the size of this differential effect may not be very large. It would seem unlikely that the general superiority of the unit-weighted composite observed in figure 11 could be fully explained in this way. A second reason for caution is the need for cross-validation of the findings. The constituents of the unit-weighted composite were chosen on the basis of significant results found in this study; it remains to be seen whether these measures would prove as predictive for another follow-up sample.

JUDGEMENTAL vs. MECHANICAL - SAMPLE 3

Another unit-weighted composite was formed for Sample 3 (Graduate Entrants) though, because of the lack of EI measures found to be significant predictors in their own right, different criteria for inclusion were adopted. Correlations of EI measures with the

five main criteria - Overall Performance (training), Rank, Factor 1, Potential and Overall Performance (job) - were examined and those EI measures which correlated positively with all these criteria, regardless of the size of the coefficients, were included. As a result the composite consisted of four measures - Drafting Test, General Information Test, Senior Officer and Holiday Companion. The unit-weighted composite was formed by summing the standard scores on the four measures. In figure 12, correlations of the composite and Final Mark with the different criteria are compared. It is again seen that for each criterion the coefficient for the composite is higher than the corresponding coefficient for Final Mark. Two of the five coefficients for the unit-weighted composite are significant compared with none for Final Mark. The superiority of the composite as a predictor of training performance is particularly marked. Even allowing for the 0.28 correlation of Senior Officer

Figure 12 - Pearson Correlation Coefficients - Final Mark and a Unit-Weighted Composite with Five Criteria - Sample 3 (Graduate Entry)

	Overall Perf. (Training)	Rank	Factor 1	Potential	Overall Perf. (Job)
Unit-Weighted Composite	0.37** N=52	0.16 N=56	0.17 N=50	0.29* N=50	0.27 N=50
Final Mark	0.07 N=54	0.05 N=58	0.07 N=52	0.21 N=52	0.14 N=52

Note The unit-weighted composite comprised four measures: Drafting Test, General Information Test, Senior Officer and Holiday Companion.
* - $p<0.05$; ** - $p<0.01$ (two-tailed)

with this criterion, it is clear that the other components of the composite account for substantial additional variance. It is also worthy of note that Potential is significantly predicted by the composite though not by any of the EI measures individually.

SUMMARY OF STUDY 3 RESULTS

Taking results of the comparisons for the two samples together, there would seem to be **support for the hypothesis that mechanical combination of EI information would prove as or more valid than judgemental combination.** However the results should be interpreted conservatively with respect to the possibility of differential effects of range restriction and the need for cross-validation.

Chapter 8 - DISCUSSION AND CONCLUSIONS

Supervisory Ratings

Results of the factor analysis of supervisory ratings in this study conform to the findings of much previous research in indicating a pervasive global evaluation effect across ratings on individual performance dimensions. No other factor of anything like comparable size was derived. A second factor seems to represent a contrast between dimensions with a social skills component and those with a task/motivational component, indicative of differences in style rather than in level of job performance. The finding that this factor correlated positively with the Holiday Companion peer nomination (figure 5) would seem to tie in with the definition of the factor as characterized by social skills at the positive end. No such clear-cut interpretation suggests itself for this factor's significant negative correlation with Chairman's interview mark (figure 5). However due to the small size of the factor in terms of the proportion of variance for which it accounts it would seem unwise, without corroborative evidence, to accept it as representing a reliable performance distinction.

Interrelationships of Criteria

Moderate to high correlations of training criteria with each other and of supervisory rating criteria with each other were to be expected given the interdependence of measures within each of these sets. Of more interest are the correlations of the overall training performance measure with measures of rank, job performance, and potential, obtained 3 to 11 years later; significant but low/moderate coefficients ranging from 0.19 to 0.29 indicate generally weak relationships. Similarly correlations of rank with overall job performance and potential are significant but low (range r=0.18 to r=0.24), indicating that the rank and supervisory rating criteria were substantially independent.

Predictive Validity

As regards the correlations of EI measures with the above criteria, it is somewhat surprising, in the light of Mays' (1972) finding of a moderate and significant correlation of EI Final Mark and post-training assessments, to find that the comparable coefficients obtained in this study were small and non-significant for both Special Course and Graduate Entry samples. An important difference between the studies, however, was that Mays used a specifically designed follow-up form for gathering training performance information whereas in this study it was necessary, due to the retrospective nature of the data collection, to place reliance on routine end-of-course

assessments. It could be that the measure used by Mays was a somewhat more sensitive indicator of training performance differences. The presence of range restriction should also be borne in mind in interpreting the findings, though this applied equally to Mays' research in which the significant results were obtained with much smaller sample sizes.

Nevertheless there are some indications from the present research that training criteria were predicted by EI measures. Three EI measures were found to correlate significantly with overall training performance for Special Course successful candidates, and though only one measure predicted this criterion for Graduate Entrants, unit-weighted composites of EI measures were found to be moderately predictive of training performance for both groups. So in general it seems that some EI components and combinations of EI components have validity for a training criterion, though this validity has not been shown to be reflected in EI Final Mark.

For Special Course successful candidates, EI Final Mark emerges as a stronger predictor of supervisory ratings than of training grades. Its correlation, after correction for range restriction, of 0.33 with Overall Performance (job) indicates a moderate prediction of general performance evaluations 7 to 12 years later. Though the corresponding correlation with the closely related measure Factor 1 is non-significant, it is of a comparable order of magnitude. The correlation with rated potential, however, is clearly lower. Corresponding coefficients

for Graduate Entrants are all positive but non-significant. Final Mark has therefore not been shown to predict rated performance or potential for the Graduate group, though in interpreting this the presence of range restriction combined with small sample size should be borne in mind.

The significant relationship of Final Mark with performance ratings is again at variance with Mays' (1972) findings; he found non-significant negative correlations of EI Final Mark and ratings of performance and potential five years after the course (i.e. approximately six years after EI assessment). A possible explanation for this discrepancy is the difference in sample sizes. At the five year follow-up in Mays' study data were available on only 35 successful Special Course candidates, conditions which may not have been conducive to obtaining reliable validity estimates. Corresponding numbers at the 7 to 12 year follow-up in the present study were over 140.

As in the case of training criteria several EI component measures emerged, for Special Course successful candidates, as significant predictors of job performance ratings in their own right. However the finding of no overlap between the subsets of measures predicting training grades and job performance ratings suggests that these criteria are of distinctly different kinds; this ties in with the weak correlations between training and job performance measures. Another finding of interest is the correlation of a unit-weighted composite of EI measures with rated potential; it appears that EIs as a whole may have some

validity in relation to this criterion despite the fact that no single EI measure significantly predicted it. Unit-weighted composite results are also of importance in relation to Graduate Entrants. Although no single EI measure significantly predicted ratings of overall job performance or potential for this group, the composite correlated significantly (r=0.29) with rated potential and at a similar though non-significant level (r=0.27) with overall job performance. Taken together the findings in relation to supervisory rating criteria parallel those for training criteria to the extent that EIs as a whole appear to have more validity than is reflected in Final Mark.

Correlations of Final Mark with rank attained for Special Course successful candidates were of a much lower order than those with training grades and performance ratings. Non-significant coefficients of 0.00 in the 7 to 12 year follow-up and 0.10 in the independent 14 to 19 year follow-up indicate negligible validity for this criterion, and the corresponding coefficient of 0.05 in the 8 to 12 year follow-up of Graduate Entrants points in the same direction. For Special Course candidates the contrast between the validities obtained using rank and performance measures provides further evidence that these are substantially independent criteria.

Rank/advancement may be a somewhat crude and unreliable type of criterion (though no means of testing its reliability is apparent) subject as it is to very many chance/opportunity factors. In the present study promotions resulted from decisions made by 44 police forces in England, Wales and Northern Ireland, so that it is quite likely rank was partly a function of factors such as differences in force manpower requirements and force promotion practices. Nevertheless many other studies have successfully validated ACs using status/advancement criteria (see Thornton and Byham, 1982), and in this study the finding that rank was moderately correlated with both the EI measure Senior Officer (for Special Course candidates) and with overall training performance would seem to demonstrate its predictability.

Validity of Peer Nominations

It is of some interest that the Senior Officer peer nomination significantly predicted rank whereas EI Final Mark did not and that, conversely, Final Mark predicted job performance ratings though Senior Officer did not. It may be recalled from chapter 5 that three other studies have compared the validities of post-AC peer and assessor evaluations. Two which used job performance/potential ratings as criteria (Vernon, 1950; Turnage and Muchinsky, 1984) found assessor evaluations to be superior while the third, which used a salary growth criterion (Mitchel, 1975), found assessor and peer evaluations to be about equally predictive. So the findings of the present study in relation to supervisory rating criteria appear in line with earlier findings

for similar criteria, while the finding that peer evaluations predicted rank more strongly than assessor evaluations partially ties in with Mitchel's finding for a closely related criterion. More research comparisons are needed to establish whether peer and assessor evaluations are consistently differentially predictive across criteria. If such differences were established they might throw some light on the nature of different criteria. In ACs assessors' observations of candidates typically take place in task-centred situations. Peers, on the other hand, may observe less of the detail of task performance but more of candidates in off-task situations. The extent to which criteria are differentially predicted by these two sources of rating evaluations might thus suggest components of those criteria. If, hypothetically, peer ratings were consistently found to be as or more predictive of advancement-related criteria than assessor ratings, while the reverse was found to be true for job performance criteria, then this might point to the effect of 'non-performance' factors on the former type of criterion, for example, Wallace's (1974) notional 'ability to make people say good things about oneself' (p.404).

Invalidity of Pencil and Paper Tests

The lack of correlation between pencil and paper test results and the criteria in this study was unexpected given the demonstrated validity of cognitive tests in a variety of contexts. Two of the tests - the Abstractions and the Advanced Progressive Matrices (APM) - could be described as 'general ability' measures, and in

metaanalyses Schmitt, Gooding, Noe and Kirsch (1984) found an average uncorrected validity for such measures across criteria of 0.248 (53 coefficients), while Ghiselli (1973) found an average validity for 'intelligence' with job proficiency criteria for executives and administrators of 0.30. It is particularly surprising that the APM was not found valid given that this is a generally accepted marker for 'fluid intelligence' (Psychometric Research Unit Hatfield, 1985). An explanation for these results could lie in the ability levels of the samples in this study; it may be, as Huck (1977) suggests, that beyond a certain level of 'intelligence' other factors, such as interpersonal and administrative skills, become more important as determinants of managerial effectiveness. A recent trial of the APM with undergraduates produced a mean of 26.8 with standard deviation 4.4 (Psychometric Test Unit Hatfield, 1985); in the present study the means for Sample 2 (Special Course successful candidates) and Sample 3 (Graduate Entrants) were 27.4 and 26.9 respectively with standard deviations of 3.8 and 4.3 respectively. So it would appear that both the Special Course and Graduate Entry groups were of about normal graduate level ability. The APM manual locates the 95th percentile, corresponding roughly to an IQ of 125, at APM scores of 24 at age 20 and 23 at age 30 (Raven, Court and Raven, 1983); by this yardstick the average scores for successful EI candidates appear very high indeed. There would seem to be considerable support here for the notion that ability levels were too high for test scores to discriminate usefully.

Performance versus Potential

Another point of theoretical interest centres on the finding that EI Final Mark significantly predicted performance ratings but not potential ratings or rank. This runs directly counter to Klimoski and Strickland's (1977) suggestion that ACs are 'prescient' rather than valid, that is, that AC assessors are better at predicting promotion decisions made by operating managers than at predicting job performance itself. The present findings thus reinforce the conclusion of Thornton and Byham's (1982) review that ACs are just as valid for performance ratings as for advancement-related criteria (see chapter 5).

Conclusions on Predictive Validity

Overall the only validity positively identified for EI Final Mark in this study was in relation to job performance ratings for successful Special Course candidates. While job performance is clearly a central criterion, it is evident from the background to the Special Course that the orientation is towards leadership of the police service. To this extent the rank criterion is important and the clear finding of negligible validity in relation to it may be seen as a cause for concern.

On the basis of a correlation between Final Mark and an overall job performance rating of 0.18 rising to 0.33 after correction (accounting for approximately 11% of the criterion variance) it seems probable that EIs are performing a worthwhile function, at least for Special Course candidates. Ideally this statement would be supported by a utility estimate, but, in the absence of

data on which to base such an estimate, one can refer to Cascio and Silbey's (1979) work (see chapter 5) suggesting that reasonably valid ACs will prove cost effective in most situations. However in relation to the general run of reported AC validities the validity found here is low. In metaanalyses Schmitt, Gooding, Noe and Kirsch (1984) have observed an average **uncorrected** validity for ACs across different criteria of 0.407 (21 coefficients), with average validities for 'performance' and 'status change' criteria both above 0.4. And to cite two relevant British comparisons, Anstey (1977) found a validity coefficient for the CSSB-based 'Final Selection Board' mark of 0.354 (0.660 after correction for range restriction; N=301) with rank attained by Civil Service Administrators after 30 years, while Jones (1984) has reported a recent validity coefficient for the Admiralty Interview Board with a training criterion of 0.36 (N=565) rising to 0.52 after correction for range restriction.

As a whole the present findings of negligible to relatively low EI validity depending on the criteria used are perhaps not surprising given that the procedure, as discussed in chapter 6, was not designed for this police application and does not appear to be very job-related. All in all there would now seem to be a case for taking a fresh look at Special Course and Graduate Entry EIs in the light of job analysis information and with a view to revision.

Efficiency

A link between external and internal validity findings is forged by the concept of efficiency/redundancy. The distinction, made in chapter 5, between redundancy as a function of the collection of information and redundancy as a function of the combination/processing of information is relevant here. For Special Course candidates, stepwise multiple regressions showed that four of the thirteen EI components accounted for all of the explainable variance in four criteria. Thus it appears that more information was collected than was needed. This may reflect the low validity found for EIs. But it seems in general to be the case that where lengthy multi-component test procedures are developed on the basis of job analysis (as was the case when EIs were originally developed in the Civil Service administrative context), redundancy of measurement is a likely result (Tenopyr, 1977). The few AC studies which have provided relevant data have indeed indicated such redundancy (see chapter 5), and the present study provides further evidence that AC practitioners should pay more attention to measurement efficiency.

Internal Validity

Redundancy as a function of assessors' processing of information has been explored within an internal validity framework. The modelling of decision making using hierarchical multiple regression proved a useful way of clarifying the picture which had emerged from bivariate correlations. The hierarchical method

seems particularly appropriate where marks awarded for different AC component techniques are to some extent interdependent; the comparison of predictions of Final Mark by First Interview and Second Interview in this study supported the notion that such interdependence was a feature of EIs.

One can use hierarchical regression to test whether assessors may be taking a particular technique into account in arriving at decisions, since one can see whether measures derived from that technique contribute significant unique variance to the prediction of OAR when entered into the regression equation after other measures. In the present study interviews generally 'passed' this test while pencil and paper tests generally 'failed' it. Ideally such analyses would be carried out in conjunction with experimental changes of AC order; by systematic trials which place each technique towards the end of the AC, it should be possible to identify techniques which consistently figure in decision making.

A general conclusion emerging from the use of these analyses in the present study is that assessors, in deciding on EI Final Marks, seem to focus on those parts of the procedure which they have subjectively marked - interviews and group and written exercises; interviews appear particularly influential. However, parts of the procedure which assessors have not subjectively marked - pencil and paper tests and peer nominations - seem hardly to be taken into account.

Internal-External Validity Comparisons

From the point of view of the overall efficiency of the EI, it is important to ascertain the extent to which this pattern of information use relates to the validity of the component techniques as assessed against external criteria. For Graduate Entrants, the uniformly negligible external validity found for EIs renders internal-external validity comparisons meaningless. The points which follow, therefore, refer to the findings for Special Course candidates.

The use which assessors make of the information obtained from group and written exercise performance would appear to be justified in terms of external validity, though there is a question mark over the Written Appreciation which does not correlate significantly with any of the major criteria. The situation as regards interviews is somewhat less clear. Though one of the interview marks (NSM) is found to correlate significantly with Overall Performance (job), in multiple regression it contributes no unique variance to the prediction of this criterion. Taken as a whole interview marks do not appear to be very predictive of external criteria, and on this basis the weight accorded to interviews by assessors seems unjustified. As regards the pencil and paper tests, the internal and external validity evidence is very much in line; these tests appear to contribute neither to decision making nor to the prediction of external criteria. A different picture emerges for the peer nominations, however. Although they appear to contribute little

to decision making, the Senior Officer nomination is one of the strongest predictors overall. It is also the only measure to predict an important criterion for Sample 3.

Overall it appears that the pattern of assessors' information use is inefficient to the extent that interviews are over-emphasized while peer nominations are hardly taken into account. Though most of the group and written exercise measures emerge as predictive of at least one important criterion, the Written Appreciation does not, and so its apparent contribution to decision making could be seen as another source of inefficiency. If EIs were to undergo revision there would, on this evidence, be a case for considering a reduction in the number of interviews in EI and removal/replacement of the Written Appreciation, the job-relatedness of which seems in any case questionable (see comment in chapter 6, page 163).

Mechanical versus Clinical/Judgemental Combination

The finding that mechanical combination of EI information was superior to clinical/judgemental combination, albeit with the caveat of the need for cross-validation, is in line with previous research in this area (see chapter 2). The results also provide support for Cascio, Valenzi and Silbey's (1978,1980) findings that unit-weighted composites increase the statistical power of validation efforts. The present research represents something of an advance over previous AC studies in its use of mechanical prediction for multiple criteria. Previous AC studies (Wollowick

and McNamarra, 1969; Mitchel, 1975; Borman, 1982) and general theoretical discussions (Dawes and Corrigan, 1974; Einhorn and Hogarth, 1975) in this area have considered only the one criterion case. It seems unlikely, however, that in many real selection situations one criterion would suffice to take account of the full range of performance in which one is potentially interested. To the extent that AC assessors, in arriving at OARs, are meant to take account of this full range of performance, the multiple criteria comparisons of the present study are perhaps more realistic. Where more than one criterion is used it becomes clear that regression-weighted composites, with weights chosen according to the prediction of each individual criterion, are inappropriate. The unit-weight approach, however, remains viable. In this study its superiority over clinical/judgemental combination was consistent both for different samples and across different criteria.

These comparisons would seem to indicate that in any future review of EI procedures the possibility of introducing a mechanical element into decision making should be considered. Clearly, however, a balance has to be struck between predictive efficiency and acceptability to AC users.

Conclusions

In summary, the central conclusion of the present study is that EI OAR (Final Mark) has some small validity in relation to a job performance criterion for Special Course successful candidates. However no such validity has been determined for other types of criteria (training and rank) or for Graduate Entrants. Relatively low validity overall is interpreted principally in terms of the apparent lack of job relatedness of parts of the EI procedure, but also in terms of possible insensitivity of the training performance measures used, inefficiencies in EI decision making, range restriction and, in the case of Graduate Entrants, small sample size. Inefficiencies in decision making appear to stem in particular from over-emphasis on interviews and the Written Appreciation and neglect of peer nominations, and in general from the use of clinical/judgemental as opposed to mechanical means of combining EI information. Overall it seems that EIs should be reviewed with regard to the possibilities of reducing the number of interviews in the procedure, revising/replacing component techniques in relation to external validity and job analysis information, and introducing a mechanical element into decision making providing this could be made acceptable to AC users.

These findings are of relevance to the wider selection research field since there have been very few criterion-related validations of police ACs and relatively few validations of 'British style' ACs. The finding of relatively low validity

overall also lends support to the conventional wisdom that content validity or job relatedness considerations are an important element in the AC design process. Additionally the findings relating to AC efficiency are of potential general interest given that this area has been largely neglected by AC researchers. The demonstration that there is considerable redundancy in the prediction of external criteria by AC measures is in line with the expectations generated by the few previous published research findings in this area, and the practical value of conducting such research as an integral part of the criterion-related validation process is apparent. Similarly the superiority of mechanical over judgemental processing of information is in line with the findings of the few previous relevant AC studies, and the practical importance of such research in terms of improving selection validity is clear. Where this study seems to break quite new ground is in its identification of specific sources of decision making inefficiency, which in this case appear to centre on the over-emphasis of information subjectively derived by the assessors, particularly from interviews. The potential value of hierarchical multiple regression as a tool for investigating decision making processes has also been demonstrated.

On a more general note, this study has indicated that the use of multiple predictors and multiple criteria facilitates exploration of differences among criterion measures. Here this approach has led to the identification of a line of investigation into the

relationship of supervisory rating and rank/advancement types of criteria, and to evidence which directly bears upon the suggestions of Klimoski and Strickland (1977) as to the sorts of criteria most effectively predicted by ACs. The use of multiple criteria also gives a more realistic representation of the scope of typical selection decision making than is achieved by more common single criterion investigations.

APPENDIX

NAME _____

The above named officer attended the Police Special Course several years ago and a follow up of such officers is being conducted. This questionnaire is designed to gather information about the officer's job performance and potential, and it is similar to a standard annual report form. The questionnaire is for research purposes only, and it should not be added to the officer's personal file.

In completing the questionnaire you are asked not to take attendance at the Special Course into account but to compare the officer's performance with that of other officers in the same rank. It is acknowledged that the reporting officer will not know the qualities of officers throughout the Force, but it is assumed that the qualities of officers under his command will generally be similar to the distribution in the Force as a whole.

For each item please circle the appropriate statement. For example, if the officer held the rank of inspector and you felt his "Organization of work" to be above average but not as good as the best 5% of inspectors, you would complete the first item as follows:

ORGANIZATION OF WORK

Concerns the ability to cope with a number of tasks at once and decide upon a sensible order of priorities.

Always takes new tasks in his stride while continuing to cope well with his other work	Best 5%	(Next 25%)	Middle 40%	Next 25%	Bottom 5%	Sometimes flustered by additional work demands

Please complete all items in this way.

ORGANIZATION OF WORK

Concerns the ability to cope with a number of tasks at once and decide upon a sensible order of priorities.

Always takes new tasks in his stride while continuing to cope well with his other work	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Sometimes flustered by additional work demands

JUDGEMENT

Concerns the use of common sense, powers of reasoning and discernment in assessing a situation or problem in order to discriminate between the various alternatives and come to a sound conclusion.

Proposals or decisions are consistently sound	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Poor perception of relative merits or feasibility in some situations

FORESIGHT

Concerns the ability to look ahead, anticipate problems and difficulties and make plans to deal with them.

Consistently anticipates problems and develops solutions in advance	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Tends to handle problems only after they arise

ORAL EXPRESSION

Concerns the ability to express oneself clearly and concisely, and effectively communicate ideas, instructions etc. when the need arises.

An extremely fluent and effective communicator	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Verbal expression is a weak point

WRITTEN WORK

Concerns the quality of reports, statements, and all other written work and includes accuracy, attention to detail, clarity and conciseness.

Outstanding written work	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Written work is a weak point

PROFESSIONAL KNOWLEDGE

Concerns general knowledge of law and practice, relevant police procedures, general orders and instructions etc., especially in relation to his particular job but also in connection with developments in the police service as a whole.

Good general knowledge with appropriate breadth and depth	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Displays some gaps, weaknesses or limitations in knowledge

APPLICATION OF PROFESSIONAL KNOWLEDGE

Concerned not so much with what an officer knows as what he does i.e. the effectiveness with which theoretical knowledge is applied in a practical way to achieve the objectives of the job in hand.

Shows a very high standard of practical application in all aspects of his work	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Practical application somewhat poor

MANNER WITH THE PUBLIC

Concerns an officer's general attitude to and manner of dealing with people, whether offenders or ordinary members of the public, with whom he comes into contact in his official capacity; covers both face to face and telephone interactions.

Exceptional ability in handling people in a variety of situations	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Dealing with the public is not a strong point; sometimes rubs people up the wrong way

ASSESSMENT OF PEOPLE

Concerns ability to judge people from many different walks of life and to understand their behaviour in a variety of contexts.

A shrewd judge of people and situations	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Sometimes mistaken in his judgements of people

RELATIONSHIPS WITH COLLEAGUES

Concerns the ability to mix well and cooperate with colleagues, both in his own section and in other departments, thus assisting effective teamwork.

Establishes excellent working relationships at all levels	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Can be a difficult colleague

RELIABILITY UNDER PRESSURE

Concerns the ability to deal with pressures of work and demanding situations while retaining a rational and balanced outlook.

Well able to cope with severe pressures	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Some deterioration in performance when under pressure

INITIATIVE

Concerns the officer's ability to think for himself and initiate positive action without having to wait for a lead from others.

Exceptional enterprise and resourcefulness in all types of situations	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Somewhat unimaginative; tends to need prompting

EFFORT AND VITALITY

Concerns drive and energy, i.e. the degree to which an officer exerts himself in particular tasks and in his work in general.

Spares no effort in all aspects of his job	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Makes minimum effort to get by

ACCEPTANCE OF RESPONSIBILITY

Concerns the degree of willingness to accept responsibility and not the judgement exercised in using that responsibility.

Seeks and accepts responsibility at all times	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Somewhat reluctant to take on responsibility

TURN-OUT

Concerns appearance when on duty; includes standards of dress, general bearing and personal neatness from which an overall impression of turn-out is formed.

Appearance and bearing always impeccable	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Pays insufficient attention to turn-out

LEADERSHIP

Concerns ability to inspire other people to exert their efforts towards a common objective.

Always inspires and leads by example	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Somewhat lacking in leadership qualities

SUPERVISORY ABILITY

Concerns the ability to maintain good discipline and control of subordinates without the need for oppressive behaviour.

Maintains good control at all times while encouraging a constructive work atmosphere	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Poor in his control of subordinates

OVERALL PERFORMANCE

Concerns the officer's overall effectiveness in his present rank.

Outstanding	Best 5%	Next 25%	Middle 40%	Next 25%	Bottom 5%	Poor

POTENTIAL

Please estimate the officer's prospects within the service by circling the appropriate statement below:

Likely to reach the highest ranks in the service	Should rise two or more ranks	Should rise one rank but probably no further	Limited potential beyond present rank

REFERENCES

ANSBACHER, H.L. (1941). German military psychology. **Psychological Bulletin, 38**, 370-392.

ANSTEY, E. (1966). The civil service administrative class and the diplomatic service: A follow-up. **Occupational Psychology, 40**, 139-151.

ANSTEY, E. (1977). A 30-year follow-up of the CSSB procedure, with lessons for the future. **Journal of Occupational Psychology, 50**, 149-159.

AMERICAN PSYCHOLOGICAL ASSOCIATION (1954). **Technical Recommendations for Psychological Tests and Diagnostic Techniques**, Washington DC: American Psychological Association.

ARBOUS, A.G. & MAREE, J. (1951). Contribution of two group discussion techniques to a validated test battery. **Occupational Psychology, 25**, 73-89.

ARCHAMBEAU, D.J. (1979). Relationships among skill ratings assigned in an assessment center. **Journal of Assessment Center Technology, 2**, 7-20.

ARVEY, R.D. & CAMPION, J.E. (1982). The employment interview: A summary and review of recent research. **Personnel Psychology, 35**, 281-322.

BASS, B.M. (1954). The leaderless group discussion. **Psychological Bulletin, 51**, 465-492.

BERRIEN, F.K. (1976). A general systems approach to organizations. In M.D. Dunnette (ed.), **Handbook of Industrial and Organizational Psychology**, Chicago: Rand McNally.

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

BOCHE, A. (1977). Management concerns about assessment centers. In J.L. Moses and W.C. Byham (eds), **Applying the Assessment Center Method**, New York: Pergamon Press.

BORMAN, W.C. (1982). Validity of behavioral assessment for predicting military recruiter performance. **Journal of Applied Psychology, 67**, 3-9.

BRAY, D.W. & CAMPBELL, R.J. (1968). Selection of salesmen by means of an assessment center. **Journal of Applied Psychology**, 52, 36-41.

BRAY, D.W., CAMPBELL, R.J. & GRANT, D.L. (1974). **Formative Years in Business: A Long-Term AT&T Study of Managerial Lives**, New York: Wiley.

BRAY, D.W. & GRANT, D.L. (1966). The assessment center in the measurement of potential for business management. **Psychological Monographs**, 80, 1-27.

BYHAM, W.C. (1970). Assessment centers for spotting future managers. **Harvard Business Review**, July-August, 150-160.

CAMPBELL, D.T. & FISKE, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. **Psycholocical Bulletin**, 56, 81-105.

CAMPBELL, J.P. (1976). Psychometric theory. In M.D. Dunnette (ed.), **Handbook of Industrial and Organizational Psychology**, Chicago: Rand McNally.

CASCIO, W.F. (1982). **Costing Human Resources: The Financial Impact of Behavior in Organizations**, New York: Van Nostrand Reinhold.

CASCIO, W.F. & SILBEY, V. (1979). Utility of the assessment center as a selection device. **Journal of Applied Psychology**, 64, 107-118.

CASCIO, W.F., VALENZI, E.R. & SILBEY, V. (1978). Validation and statistical power: Implications for applied research. **Journal of Applied Psychology**, 63, 589-595.

CASCIO, W.F., VALENZI, E.R. & SILBEY, V. (1980). More on validation and statistical power. **Journal of Applied Psychology**, 65, 135-138.

CASTLE, P.F.C. & GARFORTH, F.I. de la P. (1951). Selection, training and status of supervisors: I Selection. **Occupational Psychology**, 25, 109-123.

CIVIL SERVICE COMMISSION (1977). A Guide to the Civil Service Selection Board, London: HMSO.

COMMISSION FOR RACIAL EQUALITY (1983). **Code of Practice**, London: Commission for Racial Equality.

COOPER, W.H. (1981). Ubiquitous halo. **Psychological Bulletin, 90**, 218-244.

CRONBACH, L.J. (1970). **Essentials of Psychological Testing** (3rd ed.), New York: Harper and Row.

CRONBACH, L.J. (1980). Selection theory for a political world. **Public Pesronnel Management,** January-February, 37-50.

CROOKS, L.A. (1977). The selection and development of assessment center techniques. In J.L. Moses and W.C. Byham (eds), **Applying the Assessment Center Method,** New York: Pergamon Press.

DAVIES, J.G.W. (Chairman;1969). **The Method II System of Selection,** Report of the Committee of Enquiry, London: HMSO.

DAWES, R.M. & CORRIGAN, B. (1974). Linear models in decision making. **Psychological Bulletin, 81**, 95-106.

DODD, W.E. (1977). Attitudes toward assessment center programs. In J.L. Moses and W.C. Byham (eds), **Applying the Assessment Center Method,** New York: Pergamon Press.

DRAKELEY, R. (1984). **The use of Biographical Data in the Selection of Royal Naval Officers.** Presented at 26th Annual Conference of the Military Testing Association, Munich, Federal Republic of Germany.

DULEWICZ, V. & FLETCHER, C. (1982). The relationship between previous experience, intelligence and background characteristics of participants and their performance in an assessment centre. **Journal of Occupational Psychology**, 55, 197-207.

DULEWICZ, V., FLETCHER, C. & WOOD, P. (1983). A study ofhe internal validity of an assessment centre and of participants' background characteristics and attitudes: A comparison between British and American findings. **Journal of Assessment Center Technology**, 6, 15-24.

DUNNETTE, M.D. (1963). A modified model for test validation and selection research. **Journal of Applied Psychology, 47**, 317-323.

DUNNETTE, M.D. (1976). Aptitudes, abilities and skills. In M.D. Dunnette (ed.), **Handbook of Industrial and Organizational Psychology**, Chicago: Rand McNally.

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

DUNNETTE, M.D. & MOTOWIDLO, S.J. (1976). Police selection and career assessment. Washington D.C.: National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice, November.

EATON, N.K., WING, H. & MITCHELL, K.J. (1985). Alternate methods of estimating the dollar value of performance. **Personnel Psychology**, 38, 27-40.

EINHORN, H.J. & HOGARTH, R.M. (1975). Unit weighting schemes for decision making. **Organizational Behavior and Human Performance**, 13, 171-192.

FARAGO, L. (1972). **German Psychological Warfare**, New York: Arno Press.

FILER, R.J. (1979). The assessment center method in the selection of law enforcement officers. In C.D. Spielberger (ed.), **Police Selection and Evaluation: Issues and techniques**, Washington: Praeger Publishers.

FINKLE, R.D. (1976). Managerial assessment centers. In M.D. Dunnette (ed.), **Handbook of Industrial and Organizational Psychology**, Chicago: Rand McNally.

FITTS, P.M. (1946). German applied psychology during World War II. **American Psychologist**, 1, 151-161.

FITZGERALD, L.F. & QUAINTANCE, M.K. (1982). Survey of assessment center use in state and local government. **Journal of Assessment Center Technology**, 5, 9-21.

FLANAGAN, J.C. (1949). Critical requirements: A new approach to employee evaluation. **Personnel Psychology**, 2, 419-425.

FLANAGAN, J.C. (1954). The critical incident technique. **Psychological Bulletin**, 51, 327-358.

FLETCHER, C.A. (1982). Assessment centres. In D. MacKenzie-Davey and M. Harris (eds), **Judging People: A Guide to Orthodox and Unorthodox Methods of Assessment**, London: McGraw-Hill.

FREDERIKSEN, N. (1961). **Consistency of Performance in Simulated Situations**, Princeton, New Jersey: Educational Testing Service.

FULTON, Lord (Chairman;1968). **The Civil Service: Vol. 3(2) Surveys and Investigations**, London: HMSO.

GALBRAITH, J.K. (1974). **The New Industrial State**, Harmondsworth, Middx.: Pelican Books.

GARDNER, K.E. & WILLIAMS, A.P.O. (1973). A twenty-five year follow-up of an extended interview selection procedure in the Royal Navy: Part 1: Introduction and preliminary analysis. **Occupational Psychology**, 47, 1-13.

GARFORTH, F.I.. de la P. (1945). War Office selection boards. **Occupational Psychology**, 19, 97-108.

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.

GINSBURG, L.R. & SILVERMAN, A. (1972). The leaders of tomorrow: Their identification and development. **Personnel Journal**, 51, 662-666.

GLASER, R., SCHWARZ, P.A. & FLANAGAN, J.C. (1958). The contribution of interview and situational performance procedures to the selection of supervisory personnel. **Journal of Applied Psychology**, 42, 69-73.

GRATTON, L. (1985). **Assessment Centres: What's gone wrong?.** Presented at the British Psychological Society Occupational Psychology Conference, Sheffield.

GUDJONSSON, G.H. & ADLAM, K.R.C. (1985). Occupational Stressors among British Police Officers. ?

GUEST, D. (1984). What's new in selection. **Personnel Management**, January, 14-16.

GUILFORD, J.P. (1956). **Fundamental Statistics in Psychology and Education** (3rd ed.), New York: McGraw-Hill.

GUION, R.M. (1976). Recruiting, selection and job placement. In M.D. Dunnette (ed.), **Handbook of Industrial and Organizational Psychology**, Chicago: Rand McNally.

GUION, R.M. (1980). On trinitarian doctrines of validity. **Professional Psychology**, 11, 385-398.

GUION, R.M. & CRANNY, C.J. (1982). A note on concurrent and predictive validity designs: A critical reanalysis. **Journal of Applied Psychology**, 67, 239-244.

HAIRE, M. (1959). Biological models and empirical histories of the growth of organizations. In M. Haire (ed.), **Modern Organization Theory**. New York: Wiley.

HAMBLETON, R.K. & COOK, L.L. (1977). Latent trait models and their use in the analysis of educational test data. **Journal of Educational Measurement**, 14, 75-96.

HANDYSIDE, J.D. & DUNCAN, D.C. (1954). Four years later: A follow-up of an experiment in selecting supervisors. **Occupational Psychology**, 28, 9-23.

HARRIES-JENKINS, G. (1980). Bureaucracy in Great Britain in the 1980s. **The Journal of Applied Behavioural Science**, 16, 317-335.

HARRIS, H. (1949). **The Group Approach to Leadership Testing**, London: Routledge and Kegan Paul.

HAYES, B. (1984). Setting a PC to pick a PC. **Police Review**, 4th May, 870-871.

HAYMAKER, J.C. & GRANT, D.L. (1982). Development of a model for content validation of assessment centers. **Journal of Assessment Center Technology**, 5, 1-7.

HEMPHILL, J.K. (1960). **Dimensions of executive positions: A study of the basic characteristics of the positions of ninety-three business executives.** Bureau of Business Research Monograph no. 98. Columbus: Ohio State University, College of Commerce and Administration.

HERRIOT, P. (1984). **Down from the Ivory Tower: Graduates and their Jobs**, Chichester: Wiley.

HERRIOT, P. & WINGROVE, J. (1984). **Is Washing-Up Really Necessary?**. Presented at the 1984 BPS Occupational Psychology Conference, York.

HINRICHS, J.R. (1969). Comparison of "real life" assessments of management potential with situational exercises, paper- and pencil- ability tests, and personality inventories. **Journal of Applied Psychology**, 53, 425-432.

HINRICHS, J.R. (1978). An eight-year follow-up of a management assessment center. **Journal of Applied Psychology**, 63, 596-601.

HINRICHS, J.R. & HAANPERA, S. (1976). Reliability of measurement in situational exercises: An assessment of the assessment center method. **Personnel Psychology**, 29, 31-40.

HOME OFFICE (1961). **Police Training in England and Wales.** Presented to Parliament by the Secretary of State for the Home Department in August, 1961. London: HMSO.

HUCK, J.R. (1977). The research base. In J.L. Moses and W.C. Byham (eds), **Applying the Assessment Center Method,** New York: Pergamon Press.

HUCK, J.R. & BRAY, D.W. Management assessment center evaluations and subsequent job performance of white and black females. **Personnel Psychology, 29,** 13-30.

HUNTER, J.E. & SCHMIDT, F.L. (1983). Quantifying the effects of psychological interventions on employee job performance and work-force productivity. **American Psychologist, 78,** 473-478.

HURLEY, K., WONG, R., & JOINER, D.A. (1982). Description of the San Francisco police captain assessment center. **Journal of Assessment Center Technology, 5,** 23-28.

JACOBS, R., KAFRY, D., & ZEDECK, S. (1980). Expectations of behaviorally anchored rating scales. **Personnel Psychology, 33,** 595-640.

JESWALD, T.A. (1977). Issues in establishing an assessment center. In J.L. Moses and W.C. Byham (eds), **Applying the Assessment Center Method,** New York: Pergamon Press.

JONES, A. (1981). Inter-rater reliability in the assessment of group exercises at a UK assessment centre. **Journal of Occupational Psychology, 54,** 79-86.

JONES, A. (1984). **Royal Navy Officer Selection: Developments, Current Procedures and Research.** Presented at 26th Annual Conference of the Military Testing Association, Munich, Federal Republic of Germany.

KIM, J-O. & KOHOUT, F.J. (1975). Multiple regression analysis: Subprogram regression. In N.H. Nie, C.H. Hull, J.G. Jenkins, K. Steinbrenner and D.H. Bent (eds), **Statistical Package for the Social Sciences** (2nd ed.), New York: McGraw-Hill.

KLIMOSKI, R.J. & STRICKLAND, W.J. (1977). Assessment centers: Valid or merely prescient. **Personnel Psychology, 30,** 353-361.

LANDY, F.J. & FARR, J.L. (1980). Performance rating. **Psychological Bulletin, 87,** 72-107.

LAWLER, III, E.E. (1967). The multitrait-multirater approach to measuring managerial job performance. **Journal of Applied Psychology**, 51, 369-381.

LEFKOWITZ, J. (1975). Psychological attributes of policemen: A review of research and opinion. **Journal of Social Issues, 31**, 3-26.

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

LINNANE, J. (1985). **National Survey of Police Recruiting Practice.** Police Extended Interview Office. Internal report.

MacKINNON, D.W. (1977). From selecting spies to selecting managers: The OSS assessment program. In J.L. Moses and W.C. Byham (eds), **Applying the Assessment Center Method**, New York: Pergamon Press.

MAGALDI, R.J., MENDOZA, R.H., STAFFORD, G.T. & FRANK, F.D. (1984). Police promotional level assessment centers: The Metro-Dade Police Department experience - focus on race, sex and assessment center cycle. **Journal of Assessment Center Technology**, 7(2), 9-16.

MALLOY, T.E. & MAYS, G.L. (1984). The police stress hypothesis: A critical evaluation. ***Criminal Justice and Behavior***, 11, 197-224.

MANT, A. (1974). A European look at assessment centers. **European Business**, Summer, 35-39.

MAYS, R. (1972). **A follow-up of the police special course.** Civil Service Department Behavioural Sciences Research Division. Internal report.

MAYFIELD, E.C. (1964). The selection interview: A reevaluation of published research. **Personnel Psychology**, 17, 239-260.

McLEOD, D. (1982). **Early History of Assessment Centres.** Paper based on material used for a presentation to the Tenth International Congress on the Assessment Center Method, Pittsburg, USA.

McCONNELL, J.J. & PARKER, T.C. (1972). An assessment center program for multi-organizational use. **Training and Development Journal**, March, 6-14.

McCORMICK, E.J. & ILGEN, D.R. (1980). **Industrial Psychology** (7th ed.), London: Goerge, Allen and Unwin.

MEEHL, P.E. (1954). **Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.** Minneapolis: University of Minnesota Press.

MILLS, R.B. (1976). Simulated stress in police recruit selection. **Journal of Police Science and Administration, 4,** 179-186.

MITCHEL, J.O. (1975). Assessment center validity: A longitudinal study. **Journal of Applied Psychology, 60,** 573-579.

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

MORRIS, B.S. (1949). Officer selection in the British Army 1942-1945. **Occupational Psychology, 23,** 219-234.

MOSES, J.L. (1973). The development of an assessment center for the early identification of supervisory potential. **Personnel Psychology, 26,** 569-580.

MOUNT, M.K. (1984). Psychometric properties of subordinate ratings of managerial performance. **Personnel Psychology, 37,** 687-702.

NEIDIG, R.D., MARTIN, J.C. & YATES, R.E. (1979). The contribution of exercise skill ratings to final assessment center evaluations. **Journal of Assessment Center Technology, 2,** 21-23.

NEIDIG, R.D. & NEIDIG, P.J. (1984). Multiple assessment center exercises and job relatedness. **Journal of Applied Psychology, 69,** 182-186.

NIE, N.H., HULL, C.H., JENKINS, J.G., STEINBRENNER, K. & BENT, D.H. (1975). **Statistical Package for the Social Sciences** (2nd ed.), New York: McGraw-Hill.

NUNNALLY, J.C. (1970). **Introduction to Psychological Measurement.** New York: McGraw-Hill.

OFFICE OF STRATEGIC SERVICES ASSESSMENT STAFF (1948). **Assessment of Men: Selection of Personnel for the Office of Strategic Services.** New York: Rinehart.

P.A. MANAGEMENT CONSULTANTS LTD. (1968). **Job Evaluation in the Police Service.** Confidential report to the Superintendents' Association and Police Federation.

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

POLICE EXTENDED INTERVIEW OFFICE (1980). **Police Extended Interviews: Special Course 1962-1979.** Internal report.

POLICE EXTENDED INTERVIEW OFFICE (1985). **The Special Course: Information for Potential Candidates.** Unpublished draft.

PRIEN, E.P. (1977). The function of job analysis in content validation. **Personnel Psychology, 30,** 167-174.

PRIEN, E.P. & RONAN, W.W. (1971). Job analysis: A review of research findings. **Personnel Psychology, 24,** 371-396.

PSYCHOMETRIC RESEARCH UNIT, HATFIELD POLYTECHNIC (1985). **Graduate and Managerial Assessment: Manual and User's Guide,** Windsor: NFER-Nelson.

RAJU, N.S., EDWARDS, J.E. & LoVERDE, M.A. (1985). Corrected formulas for computing sample sizes under indirect range restriction. **Journal of Applied Psychology, 70,** 565-566.

RAVEN, J.C., COURT, J.H. & RAVEN, J. (1983). **Manual for Raven's Progressive Matrices and Vocabulary Scales Section 4: Advanced Progressive Matrices,** London: H.K. Lewis.

REEVE, E.G. (1971). **Validation of Selection Boards,** London: Academic Press.

REILLY, R.R. & CHAO, G.T. (1982). Validity and fairness of some alternative employee selection procedures. **Personnel Psychology, 35,** 1-62.

ROBERTSON, I.T. & KANDOLA, R.S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. **Journal of Occupational Psychology, 55,** 171-183.

ROCHE, W.R. (1965). A dollar criterion in fixed-treatment employee selection. In L.J. Cronbach and G.C. Gleser, **Psychological Tests and Personnel Decisions** (2nd ed.), Urbana: University of Illinois Press.

ROSS, J.D. (1980). Determination of the predictive validity of the assessment center approach to selecting police managers. **Journal of Criminal Justice**, 8, 89-96.

ROTHWELL, S. (1984). Integrating the elements of a company employment policy. **Personnel Management**, November, 31-33.

RUNDQUIST, E.A. (1969). The prediction ceiling. **Personnel Psychology**, 22, 109-116.

RUNNYMEDE TRUST AND BRITISH PSYCHOLOGICAL SOCIETY (1980). **Discriminating Fairly: A Guide to Fair Selection**, London: Runnymede Trust.

SACKETT, P.R. (1982). A critical look at some common beliefs about assessment centers. **Public Personnel Management Journal**, 11, 140-147.

SACKETT, P.R. & DREHER, G.F. (1982). Constructs and assessment center dimensions: Some troubling findings. **Journal of Applied Psychology**, 67, 401-410.

SACKETT, P.R. & DREHER, G.F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. **Journal of Applied Psychology**, 69, 187-190.

SACKETT, P.R. & WADE, B.E. (1983). On the feasibility of criterion-related validity: The effects of range restriction assumptions on needed sample size. **Journal of Applied Psychology**, 68, 374-381.

SACKETT P.R. & WILSON M.A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. **Journal of Applied Psychology**, 67, 10-17.

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

SCHMIDT, F.L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. **Educational and Psychological Measurement**, 31, 699-714.

SCHMIDT, F.L., HUNTER, J.E., McKENZIE, R.C. & MULDROW, T.W. (1979). Impact of valid selection procedures on work-force productivity. **Journal of Applied Psychology**, 64, 609-626.

SCHMIDT, F.L., HUNTER, J.E. & URRY, V.W. (1976). Statistical power in criterion-related validity studies. **Journal of Applied Psychology**, 61, 473-485.

SCHMITT, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. **Journal of Applied Psychology**, 62, 171-176.

SCHMITT, N., GOODING, R.Z., NOE, R.A. & KIRSCH, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. **Personnel Psychology**, 37, 407-422.

SCHMITT, N., NOE, R.A., MERITT, R. & FITZGERALD, M.P. (1984). Validity of assessment center ratings for the prediction of performance ratings and school climate of school administrators. **Journal of Applied Psychology**, 69, 207-213.

SEASHORE, S.E. & YUCHTMAN, E. (1967). Factorial analysis of organizational performance. **Administrative Science Quarterly**, 12, 377-395.

SHERRID, S.D. (1979). Changing police values. In C.D. Spielberger (ed.), **Police Selection and Evaluation: Issues and techniques**, Washington: Praeger Publishing.

SISSON, E.D. (1948). Forced choice: The new army rating. **Personnel Psychology**, 1, 365-381

SKITT, B.H. (1982). **The Special Course.** Unpublished paper produced for the 19th Senior Command Course, Police Staff College, Bramshill.

SMITH, Adam. The Wealth of Nations. London: Dent 1975.

SMITH, P.C. (1976). Behaviours, results, and organizational effectiveness: The problem of criteria. In M.D. Dunnette (ed.), **Handbook of Industrial and Organizational Psychology**, Chicago: Rand McNally.

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.

STANDING, T.E. (1977). Assessment and management selection. In J.L. Moses and W.C. Byham (eds), **Applying the Assessment Center Method**, New York: Pergamon Press.

STEAD, P.J. (1973). The idea of a police college. In J.C. Alderson and P.J. Stead (eds), **The Police We Deserve**, London: Wolfe.

STEWART, A. & STEWART, V. (1981). **Tomorrow's Managers Today.** London: Institute of Personnel Management.

STEWART, R. (1979). **The Reality of Management** (revised ed.). London: Pan Books.

STEWART, R. (1983). Managerial behaviour: How research has changed the traditional picture. In M.J. Earl (ed.) **Perspectives on Management: A Multidisciplinary Analysis.** Oxford: Oxford University Press.

STEWART, V. & STEWART, A. (1978). **Managing the Manager's Growth.** Westmead, Hants: Gower.

TAFT, R. (1959). Multiple methods of personality assessment. **Psychological Bulletin, 56,** 333-352.

TASK FORCE ON ASSESSMENT CENTER STANDARDS (1980). Standards and ethical considerations for assessment center operations. **Personnel Administrator,** February, 35-38.

TEEL, K.S. & DuBOIS, H. (1983). Participants' reactions to assessment centers. **Personnel Administrator, 28,** 85-91.

TENOPYR, M.L. (1977) Content-construct confusion. **Personnel Psychology, 30,** 47-54.

TENOPYR, M.L. & OELTJEN, P.D. (1982). Personnel selection and classification. **Annual Review of Psychology, 33,** 581-618.

THOMSON, H. (1970). Comparison of predictor and criterion judgments of managerial performance using the multitrait-multimethod approach. **Journal of Applied Psychology, 54,** 496-502.

THOMSON, R. (1968). **The Pelican History of Psychology,** Harmondsworth, Middx.: Penguin.

THORNTON, G.C., & BYHAM, W.C. (1982). **Assessment Centers and Managerial Performance,** London: Academic Press.

THORNDIKE, R.L. (1949). **Personnel Selection: Tests and Measurement Techniques.** New York: Wiley.

TURNAGE, J.J. & MUCHINSKY, P.M. (1982). Transsituational variability in human performance within assessment centers. **Organizational Behavior and Human Performance, 30,** 174-200.

TURNAGE, J.J. & MUCHINSKY, P.M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. **Journal of Applied Psychology**, 69, 595-602.

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

ULRICH, L. & TRUMBO, D. (1965). The selection interview since 1949. **Psychological Bulletin**, 63, 100-116.

UNGERSON, B. (1974). Assessment centres: a review of research findings. **Personnel Review**, 3, 4-13.

VERNON, P.E. (1950). The validation of Civil Service Selection Board procedures. **Occupational Psychology**, 24, 75-95.

VERNON, P.E. & PARRY, J.B. (1949). **Personnel Selection in the British Forces**, London: University of London Press.

WAGNER, R. (1949). The employment interview: A critical summary. **Personnel Psychology**, 2, 17-46.

WALLACE, S.R. (1974). How high the validity? **Personnel Psychology**, 27, 397-407.

WEEKLEY, J.A., FRANK, B., O'CONNOR, E.J. & PETERS, L.H. (1985). A comparison of three methods of estimating the standard deviation of performance in dollars. **Journal of Applied Psychology**, 70, 122-126.

WERNIMONT, P.F. & CAMPBELL, J.P. (1968). Signs, samples and criteria. **Journal of Applied Psychology**, 52, 372-376.

WHYTE, W.H. (1956). **The Organization Man**, New York: Simon and Schuster.

WIGGINS, J.S. (1973). **Personality and Prediction: Principles of Personality Assessment**, Reading, Mass.: Addison-Wesley.

WILLIAMS, A.P.O. (1984). **The Neglected Process of Self-Selection.** Presented at 26th Annual Conference of the Military Testing Association, Munich, Federal Republic of Germany.

WILSON, N.A.B. (1948). The work of the Civil Service Selection Board. **Occupational Psychology**, 22, 204-212.

WINGROVE, J., JONES, A. & HERRIOT, P. (1985). The predictive validity of pre- and post-discussion assessment centre ratings. **Journal of Occupational Psychology, 58**, 189-192.

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

WORKING PARTY ON THE SPECIAL COURSE (1974). Internal Home Office Report.

ZEDECK, S. & CASCIO, W.F. (1984). Psychological issues in personnel decisions. **Annual Review of Psychology, 35**, 461-518.