Guidelines for the Validation and Use of Assessment Procedures for the Workplace

The Psychological Society of South Africa
SOCIETY FOR INDUSTRIAL & ORGANISATIONAL PSYCHOLOGY of SA

Published 2005

in association with People Assessment in Industry (PAI)

The Psychological Society of South Africa
SOCIETY FOR INDUSTRIAL & ORGANISATIONAL PSYCHOLOGY of SA

Copies of this publication are available from:

Secretary: Society for Industrial and Organisational Psychology of SA (SIOPSA)

Kindly contact Judith Williamson on:
Fax: (012) 998-1055
Cell: 083 304 6068
E-Mail: siopsa@worldonline.co.za
Website: www.siopsa.org.za

The price of the publication is available on request. Copies of the publication will only be dispatched once payment has been received.

Cheques: Please make cheques payable to *SIOPSA* and post to:
SIOPSA
c/o TimeAfrica
PO Box 1067
GARSFONTEIN
0042

Electronic Transfer:

Bank Details	ABSA
Branch Code	508005
Account Number	712519797

FOREWORD

It is with pride that the Society of Industrial and Organisational Psychology of South Africa presents the 3rd edition of the Guidelines for the Validation and Use of Assessment Procedures for the Workplace. There were two main objectives for the revision: to update the guidelines to be consistent with the latest research; and to ensure the guidelines reflect the current thinking around validity and validation.

The last guidelines were published in 1998 and South Africa has undergone dramatic transformation over the last decade. The changes that have been brought about have impacted on all facets of life – including the field of psychological and other assessments of people in the workplace. The professional societies for persons involved in assessment in industry regularly update their documents and guidelines to keep track with the contextual aspects of testing and assessment. The current document provides guidelines for practitioners to ensure that their instruments, tools, processes and practices comply with scientific requirements and with international best practices.

The Employment Equity Act highlights the importance of the validation of any instruments to be used for assessment and selection purposes. There are two main objectives in the Act: on the one hand to ensure that only valid and reliable assessments are used, and on the other hand, to ensure that psychometric tests and other assessments are used in a manner that is considered fair and free from bias.

The Employment Equity Act (1998) states the following:

> "Psychological testing and other similar assessments of an employee are prohibited unless the test or assessment being used –
> a) has been scientifically shown to be valid and reliable;
> b) can be applied fairly to all employees; and
> c) is not biased against any employee or group."

Over and above the legal, professional and ethical guidelines and requirements, other issues such as the language proficiency, test sophistication, educational and socio-economic background of persons being assessed should always be considered during any assessment process. The aforementioned issues, together with an ongoing focus on and gathering of cross-cultural comparative information will likely be needed for the foreseeable future to ensure fairness of assessment in the South African context.

There is a need for continued and specialised training to understand the contribution of assessments in particular circumstances and to interpret the results responsibly and ethically. We hope that this document will be helpful to practitioners, scientists and students as a reference on current best practices, globally and in South Africa.

Prof Hennie Kriek
Task Group Chairperson

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The Society for Industrial and Organisational Psychology (SIOPSA) expresses its sincere appreciation to the Executive Committee of the Society for Industrial and Organizational Psychology (SIOP), Division 14 of the American Psychological Association, for its kind consent to use, with minor changes, much of the material (pp.1-61) contained in the 2003 publication *Principles for the validation and use of personnel selection procedures* (4th edition) in Section 2 of this document.

In addition, SIOPSA hereby acknowledges the use of material (pp. 480-487) with minor changes from the 2005 publication *Applied psychology in human resource management* by Wayne F Cascio and Herman Aguinis (6th edition, USA: Pearson Prentice Hall) in Section 3 of this document.

Although SIOP, Dr Cascio and Dr Aguinis have given permission for the use of specific material in these *Guidelines*, the final responsibility for the development, interpretation and application of the *Guidelines for the validation and use of assessment procedures in the workplace* rests with SIOPSA, a division of the Psychological Society of South Africa.

SECTION 1
INTRODUCTION

1.1 AIM OF THIS DOCUMENT

The aim of the *Guidelines for the validation and use of assessment procedures in the workplace* (hereafter, the *Guidelines*) is to specify established scientific findings and generally accepted professional practice in the field of personnel assessment in the choice, development, evaluation, and use of personnel assessment procedures.

The underlying assumption of any personnel selection procedure is that the procedures used can predict one or other important and relevant behavioural requirement or job performance aspect of the position. The evaluation of any assessment procedure is thus based on the fact that sufficient proof can be found that the procedures used are indeed relevant to the position or work concerned.

The implementation of effective and defensible procedures to investigate the validity of instruments being considered by an organisation for use in assessment (or which are already being used for this purpose) is a primary responsibility of assessment professionals in the organisation. The responsibility for making fair and equitable decisions or recommendations on the career and future of an individual also has ethical implications as regards those who do not receive the necessary specialised training. An example here would be an individual who is a therapeutic psychologist by profession but who controls the personnel selection procedures in an organisation.

Further onus is placed on psychologists by the numerous selection instruments available on the market: the psychologist must consider the effectiveness and usefulness of the available selection instruments in the development of personnel selection procedures. There are, however, other issues that also warrant careful attention.

The most important of these are fairness in the interpretation of scores (particularly with regard to the comparability of scores across groups) and fairness in the combination of the information on an individual (including test scores) to make a final decision about that individual. It should be noted that such decisions are made not only at the point of entry into an organisation, but are also made to determine who is promoted and who is given valuable developmental opportunities (such as training, special education or other enriching experiences). The interests of both employer and employee are at stake in the evaluation and decision process: the former is concerned with productivity (and possibly also a respectable image in the community) and the latter with not being prejudiced by biased or ineffective decision-making tools and procedures.

Fairness is a complex subject, especially in a multicultural country such as South Africa. Over the past few years, fairness has occupied and still does occupy a central place in assessment practices as South Africa moves towards an equitable socio-political community. Gender, race and age are the main factors against which fairness of opportunity in the workplace is evaluated.

Although the validity of an assessment instrument is crucial to fairness (there should be acceptable and comparable validities across the classes of individuals as defined e.g. by race and gender), achieving an equitable state of affairs does not depend only on validity. The notion of fairness extends beyond the realm of psychometrics into the socio-political domain. There are major philosophies of fairness, each underpinned by a particular system of social values, into which fairness approaches may be classified.

This document does not focus on psychometric issues but rather on the problems of making decisions in employee assessment, placement, promotion, etc. The primary concern is that performance on a test (or any other basis for decision-making) is related to performance on the job or other measures of job success. If more information is required on psychometric issues, the publication entitled *Standards for*

educational and psychological testing (hereafter, the *Standards*), which is published by the American Educational Research Association, American Psychological Association and the National Council on Measurement in Education (AERA et al., 1999) is recommended.

The *Guidelines* is intended to be consistent with the *Standards*. This edition brings the *Guidelines* up-to-date with regard to current scientific knowledge, and further guides sound practice in the use of personnel assessment procedures. The *Guidelines* should be taken in its entirety rather than considered as a list of separately enumerated principles. National and local statutes, regulations, and case law regarding employment decisions exist. The *Guidelines* is not intended to interpret these statutes, regulations, and case law, but can inform decision-making related to them.

This document provides the reader with:

- principles regarding conducting selection and validation research;
- principles regarding the application and use of selection procedures;
- information for those responsible for authorising or implement validation efforts; and
- information for those who evaluate the adequacy and appropriateness of selection procedures.

The *Guidelines* is intended to address the needs of persons involved in personnel assessment, covering many aspects of validation and personnel selection. However, other professional documents (e.g., *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists* and *International guidelines for test use* of the International Test Commission*)* may also provide guidance in particular situations.

The *Guidelines* is largely a technical document, but it is also an informational document. It is important to recognise that this document constitutes pronouncements that guide, support, or recommend, but do not mandate, specific approaches or actions. This document is intended to be aspirational and to facilitate and assist the validation and use of selection procedures. It is not intended to be mandatory, exhaustive, or definitive, and may not be applicable to every situation.

Sound practice requires professional judgment to determine the relevance and importance of the *Guidelines* in any particular situation. The *Guidelines* is not intended to mandate specific procedures independent of the professional judgment of those with expertise in the relevant area. In addition, this document is not intended to provide advice on complying with local or national laws that might be applicable to a specific situation.

The *Guidelines* expresses expectations toward which the members of SIOPSA and other researchers and practitioners should strive. Evidence for the validity of the inferences from a given selection procedure may be weakened to the extent that the expectations associated with professionally accepted practice, and consequently the *Guidelines*, are not met. However, circumstances in any individual validation effort or application affect the relevance of a specific principle or the feasibility of its implementation. Complete satisfaction of the *Guidelines* in a given situation may not be necessary or attainable.

The *Guidelines* is intended to represent the consensus of professional knowledge and practice as it exists today. However, personnel selection research and development is an evolving field in which techniques and decision-making models are subject to change. Acceptable procedures other than those discussed in this edition of the *Guidelines* may be developed in the future. In certain instances, references are cited that provide support for the principles, but these citations are selective rather than exhaustive. Both researchers and practitioners are expected to maintain an appropriate level of awareness of research developments relevant to the field of personnel selection.

The *Guidelines* is not intended:

- to be a substitute for adequate training in validation procedures;
- to be exhaustive (although it covers the major aspects of selection procedure validation and use);

- to be a technical translation of existing or future regulations;
- to freeze the field to prescribed practices and so limit creative endeavours; or
- to provide an enumerated list of separate principles.

Selection Procedures Defined. Selection procedures refer to any procedure used singly or in combination to make a personnel decision including, but not limited to, paper-and-pencil tests, computer-administered and Internet-delivered tests, performance tests, work samples, inventories (e.g., personality, interest), projective techniques, individual assessments, assessment centre evaluations, biographical data forms or scored application blanks, interviews, educational requirements, experience requirements, reference checks, background investigations, physical requirements (e.g., height or weight), physical ability tests, appraisals of job performance, computer-based test interpretations, and estimates of advancement potential. These selection procedures include methods of measurement that can be used to assess a variety of individual characteristics that underlie personnel decision-making.

The terms "selection procedure," "test," "predictor," and "assessment" are used interchangeably throughout this document. Personnel decisions are employment-related decisions to hire, train, place, certify, compensate**,** promote, terminate, transfer, and/or take other actions that affect employment status.

1.2 PROCEDURES IN THE DEVELOPMENT OF THESE GUIDELINES

Historically, the international community of industrial psychologists recognised the need for guidelines for personnel assessment procedures. The need for guidelines for the validation and use of personnel assessment procedures in South Africa was raised during the 1991 Psychometrics Congress of the Society for Industrial Psychology, and a task group was formulated to address this. The assignment given to the task group was to compile a consensus document that would reflect the current state of research and practice in South Africa and in the international community, which could be used as a basis for establishing sound practices in personnel decisions in the South African context.

The task group formed first reviewed the American guidelines, and a discussion document for the South African context was drafted. Following a number of meetings, discussions and review, the final document was presented to the Executive Committee of the Society during May 1992.

The 1997 Industrial Psychology Conference saw the establishment of the Psychological Assessment Initiative, with one of its main objectives being to develop and publish theoretically sound and practically useful criteria for evaluating psychometric instruments to be used for organisational applications. This objective translated into the revision of the *Guidelines* for the validation and use of assessment procedures for the workplace. The revised guidelines enable practitioners across the country to engage in validity studies that will yield acceptable results for defending the use of the instrument or technique. This is particularly important in view of South Africa's labour legislation as, in the case of a labour dispute, test users have to be able to defend the use of their instruments in a court of law. The maintenance of nationally acceptable validity standards can assist in eliminating a great deal of argument and litigation.

The process followed in revising these guidelines was comprehensive and democratic. The guidelines first published in 1992 were distributed to a number of experts in the field of psychometrics and industrial psychology and to HR practitioners. Following this a workshop open to all interested parties was held to discuss shortfalls in the initial guidelines, as well as additional information that had to be included. After the meeting several changes were made and a blueprint of the final guidelines document distributed to the delegates of the workshop for comments. Based on feedback, minor changes were made, following which the final document was presented in June 1998.

During the SIOPSA conference in 2004 the decision was made to update these guidelines once more, and to publish them and present them at the SIOPSA conference in 2005. With this goal in mind, a task force was formed at the 2004 conference to work on the new guidelines. The American principles document was circulated to the task force, and feedback was requested on updating the guidelines within the South African context. Feedback was integrated into a draft document, which was the discussion topic at a workshop to discuss inconsistencies in the feedback. Following the workshop, the document was finalised.

The task force consisted of:

Hennie Kriek (Chairman)	UNISA/SHL
Abed Moola	University of KZN-Westville
Aletta Odendal	UNISA/Aletta Odendal Consulting
Andrew Thatcher	WITS
Anne Buckett	Precision HR
Callie Theron	University of Stellenbosch
Cas Prinsloo	HSRC
Cheryl Foxcroft	University of Port Elizabeth
Deon Meiring	SAP Psychological Services
Griffiths Lubisi	SHL
Hennie Trytsman	Maccauvlei Conference Centre
HS van der Walt	Independent Practice
Kasthuri Nainaar	University of Johannesburg
Kim Dowdeswell (task force co-ordinator)	SHL
Marie de Beer	UNISA
Nkhabele Marumo	CCDF
Tina van Schalkwyk	Anglo Platinum
Vasie Naidoo	UNISA

1.3 HOW TO USE THIS DOCUMENT

The basic guidelines for the validation and use of assessment procedures (Section 2) involve a technical and often complex summary of the current state of research with regard to the validation of assessment procedures. This section deals in more detail with the most important aspects to be kept in mind during the evaluation of assessment procedures.

The control list (Section 3) is of particular value for the practitioner who wants to ensure sound practice. It will thus be possible for the practitioner to evaluate the extent to which sound practices are being adhered to. It should be emphasised that, while the standards set in this document are very high, they nevertheless represent ideals towards which every professional, practitioner, researcher and member of SIOPSA should strive.

SECTION 2
VALIDATION AND USE OF ASSESSMENT PROCEDURES

The essential principle in the evaluation of any selection procedure is that evidence must be accumulated to support an inference or assumption of job relatedness, as is required by the Employment Equity Act (No. 55 of 1998, Section 8). Selection procedures are demonstrated to be job related when evidence supports the accuracy of inferences made from scores on, or evaluations derived from, those procedures with regard to some important aspect of work behaviour (e.g., quality or quantity of job performance, performance in training, advancement, tenure, termination, or other organisationally pertinent behaviour). Although this document focuses on individual performance, group and organisational performance may also be relevant criteria.

2.1 OVERVIEW OF THE VALIDATION PROCESS

Any claim of validity made for a selection procedure should be documented with appropriate research evidence built on a foundation of systematic procedures and the principles discussed in this document. Promotional literature or testimonial statements should not be used as evidence of validity.

The *Standards for Educational and Psychological Testing,* as published by the American Education Research Association, American Psychological Association and the National Council on Measurement in Education (AERA et al., 1999, p. 184) define validity as "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test".

Validity is the most important consideration in developing and evaluating selection procedures. Since validation involves the accumulation of evidence to provide a sound scientific basis for the proposed score interpretations, it is the interpretations of these scores required by the proposed uses that are evaluated, not the selection procedure itself. The *Standards* notes that validation begins with "an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation refers to the construct or concepts the test is intended to measure" (AERA et al., 1999, p. 9). Examples of constructs or concepts include arithmetic proficiency, managerial performance, ability to design a Web page, oral presentation skills, conscientiousness, and ability to trouble-shoot technical problems with equipment on an assembly line. A clear description of the construct or conceptual framework that delineates the knowledge, skills, abilities, processes, and characteristics assessed should be developed.

Huysamen (2002) states that probably the most important implication of the *Standards* for South African test users is that far more research is needed on the validity, bias and fairness of assessment procedures used locally. For example, in the case where there is little or nothing known about the susceptibility of locally used assessment procedures to construct-irrelevant variance, no deliberate effort can be made to comply with standards dealing with this issue. Before efforts can be made to identify and remove sources of predictive bias, the presence of such bias should be documented. Even the *Standards* realises that it may be completely unfeasible to study some of these topics before a test is released, but it encourages test users and researchers to accumulate results obtained with such tests with a view to providing a database for later empirical analyses. On this score as well, much remains to be done by South African test users.

Geisinger and Carlson (1998) stress that training in psychological testing should not only include psychometric concepts such as reliability, validity, test standardisation and test bias, but also the complexities of testing individuals from culturally diverse groups. This requirement is even more relevant to the South African situation. Earlier research by Bedell, van Eeden and van Staden (1999) reported that if cross-cultural validity has not been unequivocally determined for a test it implies that cross-group comparison of scores could yield information that is discriminatory if used unconditionally.

In the early 1950s, three different aspects of test validity were discussed: content, criterion-related, and construct validity. Since that time the conceptualisation of validity evidence has undergone some modification, moving from three separate aspects of validity evidence to the current *Standards*' view of validity as a unitary concept with different sources of evidence contributing to an understanding of the inferences that can be drawn from a selection procedure. Nearly all information about a selection procedure, and inferences about the resulting scores, contributes to an understanding of its validity. Evidence concerning content relevance, criterion relatedness, and construct meaning is subsumed within this definition of validity.

The validity of any inference can be determined through a variety of different strategies for gathering evidence. The *Standards* notes that while different strategies for gathering evidence may be used, the primary inference in employment contexts is that a score on a selection procedure predicts subsequent work behaviour. Even when the validation strategy used does not involve empirical predictor-criterion linkages, such as when a user relies on test content to provide validation evidence, there is still an implied link between the test score and a criterion. Therefore, even when different strategies are employed for gathering validation evidence, the inference to be supported is that scores on a selection procedure can be used to predict subsequent work behaviour or outcomes. Professional judgment should guide the decisions regarding the sources of evidence that can best support the intended interpretation and use.

The quality of validation evidence is of primary importance. In addition, where contradictory evidence exists, comparisons of the weight of evidence supporting specific inferences to the weight of evidence opposing such inferences are desirable.

2.1.1 SOURCES OF EVIDENCE

The *Standards* discusses five sources of evidence that can be used in evaluating a proposed interpretation of selection procedure test scores for a particular use: (a) relationships between predictor scores and other variables, such as test-criterion relationships; (b) content; (c) internal structure of the test; (d) response processes; and (e) consequences of testing. Given that validity is a unitary concept, such categorisations refer to various sources of evidence rather than distinct types of validity. It is not the case that each of these five sources is an alternative approach to establishing job relatedness. Rather, each provides information that may be highly relevant to some proposed interpretations of scores, and less relevant, or even irrelevant to others.

2.1.1.1 Evidence Based on the Relationship Between Scores on Predictors and Other Variables

This form of evidence is based on the empirical relationship of predictor scores to external variables. Two general strategies for assembling empirical evidence apply. The first strategy involves examining the relationship between scores on two or more selection procedures measuring the same construct hypothesised to underlie the predictor measure. Evidence that two measures are highly related and consistent with the underlying construct can provide convergent evidence in support of the proposed interpretation of test scores as representing a candidate's standing on the construct of interest. Similarly, evidence that test scores relate differently to other distinct constructs can contribute to discriminant evidence of validity. Note that convergent and discriminant evidence does not in and of itself establish job relatedness. Establishing job-relatedness requires evidence that links selection procedure scores to work-relevant behaviour.

A second strategy typically involves relating a test or other selection procedure to a criterion. This strategy has historically encompassed two study designs: predictive and concurrent. A predictive study examines how accurately test scores predict future performance. In a concurrent study predictor and criterion data are collected during a relatively simultaneous time frame, although the objective remains to predict performance.

2.1.1.2 Content-Related Evidence

Test content includes the questions, tasks, format, and wording of questions, response formats, and guidelines regarding administration and scoring of the test. Evidence based on test content may include logical or empirical analyses that compare the adequacy of the match between test content and work content, worker requirements, or outcomes of the job.

2.1.1.3 Evidence Based on the Internal Structure of the Test

Studies that examine the internal structure of a test and the relationship among its items or tasks (e.g., work samples) can provide additional evidence of how test scores relate to specific aspects of the construct to be measured. Such evidence typically includes information concerning the relationships among items and the degree to which they represent the appropriate construct or content domain. For example, evidence of a high degree of item homogeneity is appropriate when a single dimension or construct is to be measured. If the conceptual framework requires a more complex structure, overall consistency among items may not provide appropriate evidence of the internal structure of the test. When a multidimensional factor structure is proposed, evidence that supports inferences concerning the validity of score interpretations for the subcomponents in the predictor may be appropriate.

2.1.1.4 Evidence Based on Response Processes

In employment contexts, evidence based on response processes is necessary when claims are made that scores can be interpreted as reflecting a particular response process on the part of the examinee. For example, if a claim is made that a work sample measures the use of proper techniques for resolving customer service problems, simply assessing whether the problem is resolved is not sufficient evidence of validity. Evidence based on both cognitive and physical response processes may provide additional evidence of validity. Examining the processes used by individuals in responding to performance tasks or test questions can provide such evidence. Often evidence regarding individual responses can be gathered by (a) questioning test takers about their response strategies, (b) analysing examinee response times on computerised assessments, or (c) conducting experimental studies where the response set is manipulated. Observations of how individuals engage in performance tasks can also illustrate the extent to which the task is eliciting behaviour related to the intended construct as opposed to behaviour more related to irrelevant constructs. However, in many employment contexts such evidence is irrelevant to the proposed use, as is the case where the only claim made is that the scores on the selection procedure are predictive of a particular work outcome.

2.1.1.5 Evidence Based on Consequences of Personnel Decisions

In recent years, one school of thought has advocated incorporating examination of consequences of the use of predictors in the determination of validity. This perspective views unintended negative consequences as weakening the validity argument. Although evidence of negative consequences may influence policy or practice decisions concerning the use of predictors, these *Guidelines* and the *Standards* take the view that such evidence is relevant to inferences about validity only if the negative consequences can be attributed to the measurement properties of the selection procedure itself. Subgroup differences resulting from the use of selection procedures are often viewed as a negative consequence of employment selection. Group differences in predictor scores and selection rates are relevant to an organisation and its employment decisions, yet such differences alone do not detract from the validity of the intended test interpretations. If the group difference can be traced to a source of bias or contamination in the test, then the negative consequences do threaten the validity of the interpretations. Alternatively, if the group difference on the selection procedure is consistent with differences between the groups in the work behaviour or performance predicted by the procedure, the finding of group differences could actually support the validity argument. In this case, negative consequences from test use constitute a policy issue for the user, rather than indicate negative evidence concerning the validity of the selection procedure.

A different example of negative consequences is also helpful. An organisation that introduces an integrity test to screen applicants may assume that this selection procedure provides an adequate safeguard against employee theft and will discontinue use of other theft-deterrent methods (e.g., video surveillance). In such an instance, employee theft might actually increase after the integrity test is introduced and other organisational procedures are eliminated. Thus, the decisions subsequent to the introduction of the test may have had an unanticipated, negative consequence on the organisation. Such consequences may lead to policy or practice decisions to reduce the negative impact. However, such consequences do not threaten the validity of inferences that can be drawn from the integrity tests, as the consequences are not a function of the test itself.

2.1.2 PLANNING THE VALIDATION EFFORT

Before an assessment procedure is considered or a validation effort is planned, the proposed uses of the selection procedures being considered must be based on an understanding of the work performed, and the needs and rights of the organisation and its present and prospective employees. These proposed uses should be consistent with professional, ethical, and legal responsibilities. Validation begins with a clear statement of the proposed uses as well as the intended interpretations and outcomes of the instrument, and should be designed to determine how well the proposed uses will be achieved. All aspects of the decision-making process should make a valid contribution to the achievement of these objectives, however, the validity of the final assessment decision is of primary importance.

Selection procedures used in the overall selection process should be supported by validity evidence. When a selection decision is based on multiple components combined into a composite, evidence for the final decision has primary importance. The validation effort should accumulate evidence that generalises to the selection procedure and work behaviour in the operational setting. The design of this effort may take many forms such as single local studies, consortium studies, meta-analyses, transportability studies, or synthetic validity/job component studies. More than one source of evidence or validation strategy may be valuable in any particular validation effort.

In planning a validation effort for personnel decisions, three sources of evidence are most likely to be relevant: relationships to measures of other variables, content-related evidence, and internal structure evidence. Under some circumstances evidence based on response processes and evidence based on consequences may be important to consider. The decision to pursue one or more of these sources of evidence is based on many considerations including proposed uses, types of desired selection procedures, availability and relevance of existing information and resources, and strength and relevance of an existing professional knowledge base. Where the proposed uses rely on complex, novel, or unique conclusions, multiple lines of converging evidence may be important.

The design of the validation effort is the result of professional judgment, balancing considerations that affect the strength of the intended validity inference with practical limitations. Important aspects to be considered include (a) existing evidence, (b) design features required by the proposed uses, (c) design features necessary to satisfy the general requirements of sound inference, (d) feasibility of particular design features, and (e) whether the validity needs to be explored for different groups (e.g. culture, language, gender) and how this will impact on the design features.

2.1.2.1 Existing Evidence

An important consideration in many validation efforts is whether sufficient validity evidence already exists to support the proposed uses. The availability and relevance of existing evidence and the potential information value of new evidence should be carefully weighed in designing the validation effort. All validity conclusions are generalisations from the results in the validation setting to selection procedures and work behaviour in the operational setting. The information value of existing and possible new evidence is based on the many factors that affect the strength of this generalisation.

Existing evidence provides information value where it establishes statistical dependability and supports the generalisation from the validation setting(s) to the operational settings. Where such evidence has been accumulated, it may provide a sufficient rationale for inferring validity in the operational setting and may support a decision not to gather additional evidence. Such inferences depend on evidence of validity rather than mere claims of validity. Advances in meta-analysis methods and a growing knowledge base of meta-analysis results have established considerable validation evidence for cognitive ability measures, and evidence is accruing for noncognitive measures such as personality and physical abilities. However, existing evidence alone may not be sufficient to support inferences of validity in a given situation.

Validity conclusions based on existing evidence may be strengthened by evidence from more than one method especially where the validity inference depends heavily on some underlying or theoretical explanatory concept or construct. In such cases, different methods may not support the same conclusions about the underlying explanatory concepts or constructs. For example, factor analyses of test scores may not replicate factor analyses of ratings of the same attributes. In these situations, convergent and discriminant evidence across multiple methods may be important.

2.1.2.2 Proposed Uses

In designing a validation effort, whether based on existing evidence, new evidence, or both, primary consideration should be given to the design features necessary to support the proposed uses. Examples of such features include the work to be targeted (e.g., one job title or a family of related work), the relevant candidate pool (e.g., experienced or nonexperienced candidates, candidates from diverse cultural and language backgrounds), the uniqueness of the operational setting (e.g., one homogeneous organisation or many different organisations), and relevant criterion measures (e.g., productivity or turnover).

2.1.2.3 Requirements of Sound Inference

Primary consideration should also be given to the general requirements of sound validity inferences. These include measurement reliability and validity, representative samples, appropriate analysis techniques, and controls over plausible confounding factors. People who provide information in the validation effort should be knowledgeable and qualified for the tasks they are asked to perform and content they are asked to contribute.

2.1.2.4 Feasibility

Validation planning must consider the feasibility of the design requirements necessary to support an inference of validity. Unfortunately, it is not always possible to conduct a well-designed or even an acceptable study. Validation efforts may be limited by time, resource availability, sample size, or other organisation constraints including cost. In some situations these limits may narrow the scope of appropriate generalisations, but in other situations they may cause design flaws leading to inaccurate generalisations. While validation efforts with a narrow focus may have value, poorly executed validation efforts may lead the employer to reject beneficial selection procedures or accept invalid ones. Misleading, poorly designed validation efforts should not be undertaken. A poor study is not better than no study at all.

2.1.3 ANALYSIS OF WORK

Historically, selection procedures were developed for specific jobs or job families. This often remains the case today, and traditional job analysis methods are still relevant and appropriate in those situations. However, organisations that experience rapid changes in the external environment, the nature of work, or processes for accomplishing work may find that traditional jobs no longer exist. In such cases, considering the competencies or broad requirements for a wider range or type of work activity may be more appropriate. Competency models are often used by organisations for many

different purposes (Schippmann et al., 2000). When they are intended to support the underlying validity or use of a selection procedure, these *Guidelines* apply.

The term "analysis of work" is used throughout this document and includes information that traditionally has been collected through job analysis methods as well as other information about the work, worker, organisation, and work environment. The focus for conducting an analysis of work may include different dimensions or characteristics of work including work complexity, work environment, work context, work tasks, behaviours and activities performed, or worker requirements (e.g., knowledge, skills, abilities, and other personal characteristics [KSAOs]).

2.1.3.1 Purposes for Conducting an Analysis of Work

Two major purposes can be identified for an analysis of work: to develop selection procedures and to develop or identify criterion measures. Part of the process of developing selection procedures is an analysis of work that identifies worker requirements including a description of the general level of ability, skill, knowledge, or other characteristics needed. Such an analysis of work would determine the characteristics workers need to be successful in a specific work setting, or the degree to which the work requirements are similar to requirements for work performed elsewhere. The other purpose of developing or identifying criterion measures is done by assembling the information needed to understand the work performed, the setting in which the work is accomplished, and the organisation's goals.

In conducting an analysis of work, the important consideration is building an understanding of the organisation's needs as they relate to the assessment problem so that the researcher can formulate sound hypotheses about the relationships between predictors and criteria. Such analysis is essential for justifying a construct as being important for job success. A number of job analysis procedures exist, each with differing contributions to the objectives of a validity study. Several books provide summaries of various job analysis procedures and discussions of their relative utility in various situations (Brannick & Levine, 2002; Gael, 1983; McCormick, 1979).

There is no single approach that is the preferred method for the analysis of work. The analyses used in a specific study of work are a function of the nature of work, current information about the work, the organisational setting, the workers themselves, and the purpose of the study. Understanding the organisation's requirements or objectives is important when selecting an appropriate method for conducting an analysis of work. The choice of method and the identification of the information to be gathered by that method should include the relevant research literature.

2.1.3.2 Level of Detail

The level of detail required of an analysis of work is directly related to its intended use and the availability of information about the work.

Situations where a less detailed analysis may be sufficient is when there is already information descriptive of the work, or when prior research about the job requirements allows the generation of sound hypotheses concerning the predictors or criteria across job families or organisations. When a detailed analysis of work is not required, the researcher should compile reasonable evidence establishing that the job(s) in question are similar in terms of work behaviour and/or required knowledge, skills, abilities, and/or other characteristics, or falls into a group of jobs for which validity can be generalised. Situations that require a more detailed analysis of work may include those in which there is little existing work information available, or the organisation intends to develop predictors of specific job knowledge.

Any methods used to obtain information about work or workers should have reasonable psychometric characteristics and should be understood by the participants. The sources of job information should also be credible. Lack of consensus about the information contained in the analysis of work should be

noted and considered further. Current job descriptions or other documents may or may not serve the immediate research purpose. Such information needs to be evaluated to determine its relevance and usefulness.

In some instances, an analysis of work may be the basis for assigning individuals to or selecting individuals for future jobs that do not exist at present. In other instances, an analysis of work may be used for transitioning workers from current to future work behaviours and activities. In both instances, the future work behaviours and activities, as well as the worker requirements may differ markedly from those that exist at present.

Similarly, the work environment in which an organisation operates may also change over time. For example, technology has permitted many individuals to work from virtual offices and replaced many functions that were previously conducted by individuals. Further, the global environment has expanded geographical boundaries and markets for many organisations. Procedures similar to those used to analyse current work requirements may be applicable for conducting an analysis of work in environments of rapid change. However, other approaches that may be more responsive to the complexities of the emerging work environments are more appropriate (Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1999; Schneider & Konz, 1989). The central point in such instances is the need to obtain reliable and relevant job information that addresses anticipated behaviours, activities, or KSAOs.

If there is reason to question whether people with similar job titles or work families are doing similar work, or if there is a problem of grouping jobs with similar complexity, attributes, behaviours, activities, or worker KSAOs, inclusion of multiple perspectives and incumbents in an analysis of work may be necessary. Even when incumbents are in positions with similar job titles or work families, studying multiple incumbents may be necessary to understand differences in work complexity, work context, work environment, job behaviours, or worker KSAOs as a function of shift, location, variations in how work is performed, and other factors that may create differences in similar job titles or worker families.

2.2 SOURCES OF VALIDITY EVIDENCE

Inferences made from the results of a selection procedure to the performance of subsequent work behaviour or outcomes need to be based on evidence that supports those inferences. Three sources of evidence will be described: namely, evidence of validity based on relationships with measures of other variables, evidence based on content, and evidence based on the internal structure of the selection procedure. The generalisation of validity evidence accumulated from existing research to the current employment situation is discussed in the "Generalising Validity Evidence" section.

2.2.1 EVIDENCE OF VALIDITY BASED ON RELATIONSHIPS WITH MEASURES OF OTHER VARIABLES

The *Guidelines* and the *Standards* view a construct as the concept a selection procedure is intended to measure. At times the construct is not fully understood or well articulated. However, relationships among variables reflect their underlying constructs. For example, a predictor generally cannot correlate with a criterion unless to some extent one or more of the same constructs underlie both variables. Consequently, validation efforts based on constructs apply to all investigations of validity.

Principles for using a criterion-related strategy to accumulate validity evidence in employment settings are discussed below. While not explicitly addressed, the following principles also apply to research using variables other than job performance criteria (e.g., convergent and discriminant evidence). Some theory or rationale should guide the selection of these other variables as well as the interpretation of the study results.

2.2.2 CRITERION-RELATED EVIDENCE OF VALIDITY

Personnel selection procedures are used to predict future performance or other work behaviour. This would suggest that assessment results should be interpreted in terms of expected job performance and not in terms of the construct being assessed. Evidence for criterion-related validity typically consists of a demonstration of a relationship (via statistical significance testing or establishing confidence intervals) between the results of a selection procedure (predictor) and one or more measures of work-relevant behaviour or work outcomes (criteria). The choice of predictors and criteria should be based on an understanding of the objectives for test use, job information, and existing knowledge regarding test validity.

A predictor is an aid to decision-making applied in the context of selection or other personnel decisions. Predictors include, but are not limited to, standardised ability tests, personality inventories, biographical data forms, situational tests, assessment centre evaluations, simulations, ratings based on interviews, performance ratings, and evaluations of training or experience. A standardised procedure is one that presents and uses consistent directions and procedures for administration, scoring, and interpretation. Standardised predictors and criterion measures are preferred. The discussion in this section, however, applies to all predictors and criteria, whether standardised or unstandardised.

2.2.2.1 Feasibility of a Criterion-Related Validation Study

The availability of appropriate criterion measures, the representativeness of the research sample, and the adequacy of statistical power are very important in determining the feasibility of conducting a criterion-related study. Depending on their magnitude, deficiencies in any of these considerations can significantly weaken a criterion-related validation study.

One of the considerations that must be established is that the job is reasonably stable and not in a period of rapid evolution. Although validity coefficients seem to be robust across tasks and situations (Schmidt, Hunter & Pearlman, 1981) for some predictor-criterion relationships, the traditional logic of validation research is that it is undertaken under conditions similar to those that are expected to exist when the results are made operational. If this assumption is not tenable, the researcher should either modify the validation strategy appropriately or postpone the study until the jobs and their settings are reasonably stable.

It must also be possible to obtain or develop a relevant, reliable, and uncontaminated criterion measure. Ideally these aspects should be investigated empirically. Of these characteristics, the most important is relevance. A relevant criterion is one that reflects the relative standing of employees with respect to important work behaviour(s) or outcome measure(s). If such a criterion measure does not exist or cannot be developed, use of a criterion-related validation strategy is not feasible. It is inappropriate to conduct a criterion-related study using criteria that are irrelevant or contaminated, even if they are reliable and available.

A competent criterion-related validation study should be based on a sample that is reasonably representative of the work and candidate pool. Restriction of range in the predictor, criterion or both may distort an estimate obtained from a particular sample. Differences between the sample used for validation and a candidate pool on a given variable merit attention when credible research evidence exists demonstrating that the variable affects validity.

A criterion-related validity study should have adequate statistical power otherwise the issue of validity may be left unresolved. Statistical power refers to the probability of detecting a relationship between predictor and criterion in a sample if such a relationship exists in the population. A number of factors related to statistical power can influence the feasibility of a criterion-related study. Among these factors are the degree (and type) of range restriction in the predictor or the criterion, reliability of the criterion, and statistical power. Sample size, the statistic computed, the probability level chosen for the confidence interval, and the size of the predictor-criterion relationship determine the confidence

interval around the validity estimate. In practice, these threats and factors occur in varying levels that, in combination, affect power and the precision of estimation. Therefore, statistical power and precision of estimation should be carefully considered before undertaking a criterion-related validity study and, if a study is conducted, the report should include information relevant to power estimation.

2.2.2.2 Design and Conduct of Criterion-Related Studies

If a criterion-related strategy is feasible, attention is then directed to the design and conduct of the study. A variety of designs can be identified. The traditional classification of predictive and concurrent criterion-related validity evidence is based on the presence or absence of a time lapse between the collection of predictor and criterion data. The employment status of the sample (incumbents or applicants) also may differentiate the designs. In predictive designs, data on the selection procedure are typically collected at or about the time individuals are selected. After a specified period of time (for survival criteria) or after employees' relative performance levels have stabilised (for performance criteria), criterion data are collected. In concurrent designs, the predictor and criterion data are collected, usually on incumbents, at approximately the same time.

There are, however, other differences between and within predictive and concurrent designs that can affect the interpretation of the results of criterion-related validation studies. Designs may differ in the time of predictor data collection relative to a selection decision or the time at which employees start in a job—before, simultaneously, shortly after, or after a substantial time period in the job. Designs may differ with respect to the basis for the selection decision for participants in the research sample; they may have been selected using the predictor under study, an "existing" in-use predictor, a random procedure, or some combination of these. Designs also may differ with respect to the population sampled. For example, the design may use an applicant population or a population of recently hired employees, recent employees not yet fully trained, or employees with the full range of individual differences in experience.

The effect of the predictive or concurrent nature of the design may depend upon the predictor construct. For tests of cognitive abilities, estimates of validity obtained from predictive and concurrent designs may be expected to be comparable (Barrett, Phillips, & Alexander, 1981; Bemis, 1968; Pearlman, Schmidt, & Hunter, 1980). Findings regarding the comparability of predictive and concurrent designs cannot be generalised automatically to all situations and to other types of predictors and criteria.

Occasionally, a selection procedure is designed for predicting higher-level work than that for which candidates are initially selected. Such higher-level work may be considered a target job or work in a criterion-related study if a substantial number of individuals who remain employed and available for advancement progress to the higher level within a reasonable period of time. Where employees do not advance to the higher level in sufficient numbers, assessment of candidates for such work still may be acceptable if the validity study is conducted using criteria that reflect performance at both the level of work that the candidate will be hired to perform and the higher level. The same logic may apply to situations in which people are rotated among jobs.

In some organisations, work changes so rapidly or is so fluid that validation with regard to performance in one or more "target" job(s) is impossible; successful performance is more closely related to abilities that contribute broadly to organisational effectiveness. In such instances, the researcher may accumulate evidence in support of the relationship between predictor constructs (e.g., flexibility, adaptability, team orientation, learning speed, and capacity) and organisation-wide criteria (such as working effectively under very tight deadlines).

In planning a validation study, the researcher should identify the design characteristics (nature of appropriate samples; need for and feasibility of specific statistical controls, corrections or analyses;

number of cases etc.) required for a professionally acceptable validity study. The researcher should then determine how closely the design can approximate that ideal within situational constraints and decide whether criterion-related validation is feasible.

2.2.2.3 Criterion Development

In general, if criteria are chosen to represent work-related activities, behaviours or outcomes, the results of an analysis of work are helpful in criterion construction. If the goal of a given study is the prediction of organisational criteria such as tenure, absenteeism, or other types of organisation-wide criteria, an in-depth analysis is usually not necessary, though an understanding of the work and its context is beneficial. Some considerations in criterion development follow.

Criteria should be chosen on the basis of work relevance, freedom from contamination, and reliability rather than availability. This implies that the purposes of the validation study are (a) clearly stated, (b) supportive of the organisation's needs and purposes, and (c) acceptable in the social and legal context of the organisation. The researcher should not use criterion measures that are unrelated to the purposes of the study to achieve the appearance of broad coverage.

Criterion relevance. Criteria should represent important organisational, team, and individual outcomes such as work-related behaviours, outputs, attitudes, or performance in training, as indicated by a review of information about the work. Criteria need not be all-inclusive, but there should be clear rationale linking the criteria to the proposed uses of the selection procedure. Criteria can be measures of overall or task-specific work performance, work behaviours, or work outcomes. Depending upon the work being studied and the purposes of the validation study, various criteria such as a standard work sample, behavioural and performance ratings, success in work-relevant training, turnover, contextual performance/organisational citizenship, or rate of advancement may be appropriate. Regardless of the measure used as a criterion, it is necessary to ensure its relevance to work.

Criterion contamination. A criterion measure is contaminated to the extent that it includes extraneous, systematic variance. Examples of possible contaminating factors include differences in the quality of machinery, unequal sales territories, raters' knowledge of predictor scores, job tenure, and shift, location of the job, and attitudes of raters. While avoiding completely (or even knowing) all sources of contamination is impossible, efforts should be made to minimise their effects. For instance, standardising the administration of the criterion measure minimises one source of possible contamination. Measurement of some contaminating variables might enable the researcher to control statistically for them; in other cases, special diligence in the construction of the measurement procedure and in its use may be all that can be done.

Criterion deficiency. A criterion measure is deficient to the extent that it excludes relevant, systematic variance. For example, a criterion measure intended as a measure of overall work performance would be deficient if it did not include work behaviours or outcomes critical to job performance. For example, Rotundo and Sackett (2002) suggested overall work performance to be a function of work core tasks (+), citizenship (+) and counterproductive behaviour (-).

Criterion bias. Criterion bias is systematic error resulting from criterion contamination or deficiency that differentially affects the criterion performance of different subgroups. The presence or absence of criterion bias cannot be detected from knowledge of criterion scores alone. A difference in criterion scores of older and younger employees or day and night shift workers could reflect bias in raters or differences in equipment or conditions, or the difference might reflect genuine differences in performance. The possibility of criterion bias must be anticipated. The researcher should protect against bias insofar as is feasible and use professional judgment when evaluating the data.

Criterion reliability. When estimated by appropriate measures, criterion measures should exhibit reliability. For examples of appropriate and inappropriate uses of a variety of reliability estimates see Hunter and Schmidt (1996). Criterion reliability places a ceiling on validity estimates. In other words,

the effect of criterion unreliability is to underestimate criterion-related validity in the population of interest.

Ratings as criteria. Among the most commonly used and generally appropriate measures of performance are ratings. If raters (supervisors, peers, self, clients, or others) are expected to evaluate several different aspects of performance, the development of rating factors is ordinarily guided by an analysis of the work. Further, raters should be sufficiently familiar with the relevant demands of the work as well as the individual to be rated to effectively evaluate performance. Raters should also be trained in the observation and evaluation of work performance. Research suggests that performance ratings collected for research purposes can be preferable for use in validation studies to those routinely collected for administrative use (Jawahar & Williams, 1997).

2.2.2.4 Choice of Predictor

Many factors, including professional judgment and the proposed use of the selection procedure, influence the choice of the predictor(s).

Selecting predictors. Variables chosen as predictors should have an empirical, logical, or theoretical foundation. The rationale for a choice of predictor(s) should be specified. A predictor is more likely to provide evidence of validity if there is good reason or theory to suppose that a relationship exists between it and the behaviour it is designed to predict. For example, Lopes, Roodt and Mauer (2001) provide a clear rationale for the use of learning potential batteries as a predictor of job performance in job applicants within the South African context. A clear understanding of the work, the research literature, or the logic of predictor development provides this rationale. This principle is not intended to rule out the application of fortuitous findings, but such findings, especially if based on small research samples, should be verified through replication with an independent sample.

Preliminary choices among predictors should be based on the researcher's scientific knowledge without regard for personal bias or prejudice. Therefore, the researcher's choice of specific predictors should be based on theory and the findings of relevant research rather than personal interest or mere familiarity.

Predictor contamination. As with criteria, a predictor measure is contaminated to the extent that it includes extraneous, systematic variance. A number of factors can contribute to predictor contamination including unstandardised administrative procedures and irrelevant content. Some procedures, such as unstructured interviews, may be more susceptible than others to predictor contamination. Efforts should be made to minimise predictor contamination.

Predictors and selection decision strategies. Outcomes of decision strategies should be recognised as predictors. Decision makers who interpret and act upon predictor data interject something of themselves into the interpretive or decision-making process. Judgments or decisions thus may become at least an additional predictor, or, in some instances, the only predictor. For example, if the decision strategy uses judgment to combine multiple predictors (e.g., tests, reference checks, interview results) into a final selection decision, the actual predictor is the judgment reached by the person who weights and summarises all the information. Ideally, it is this judgment that should be the focus of the validation effort. If this is not feasible, support for the judgment should be based on validity evidence for the specific components.

Predictor reliability. Predictor reliability, like criterion reliability, should be estimated whenever feasible. Predictor reliability should be estimated through appropriate methods and should be sufficiently high enough to warrant use. Again the reliability of the predictor, like with the criterion, places a ceiling on any validity estimate.

2.2.2.5 Choice of Participants

Samples should be chosen with the intent to generalise to the selection situations of interest. The generalisability of the research results depends in part on the sample chosen. The nature of the sample will also be affected by the method of validation used.

The sample for a validation study should be chosen carefully. Whether the study is predictive or concurrent, a sample of incumbents is unlikely to be representative of the applicant group on all variables. The impact of characteristics such as demographics, motivation, ability, and experience on predictor-criterion relationships, and hence on the generalisation, is an empirical question, and the researcher should therefore rely on the research literature to make professional judgments about their possible relevance. Because many characteristics studied to date appear to have little or no effect on predictor-criterion relationships, no variable should be assumed to moderate validity coefficients in the absence of explicit evidence for such an effect. Research in the USA and Europe shows that validities across races (black vs. white) are usually comparable on cognitive assessment tests (Linn, 1978; Anderson, 2005).

The sample should be large enough to provide adequate statistical power. A study that has only a low probability of detecting the true validity of the predictor provides little information. Statistical power may be increased to acceptable levels in a number of ways, the most obvious of which is to increase sample size by the addition of persons sampled from the same population. (See Cohen, 1988, for more information on power analysis).

When combining data from separate samples, both jobs and workers should be similar on any variables that may affect validity. If samples are comparable on these variables, pooling the samples would provide increased statistical power.

Dropouts and/or exclusion of outliers should be explained. Occasionally, information and all variables for sample participants are not available or it was impossible to collect information on some participants. Information on the representativeness of the available sample should be provided, perhaps including items such as their job level, tenure, ethnic status or gender. Exclusion of sample members because of extreme values on some variables should be noted and justified.

2.2.2.6 Procedural Considerations

The test user should consider the probable use of any end products prior to the collection and analysis of data.

The test user may consider alternative criterion-related research methods that offer a sound rationale. Examples include co-operative research on an industry-wide basis, consortia of small users, or gathering data for validity generalisation. Such approaches generally require considerable effort in planning and in data analysis, and considerable care in ascertaining the appropriateness of the data included from the different sources.

Procedures for test administration and scoring in validation research should be clearly set out and should be consistent with the standardisation plan for operational use. Any specified operational characteristics (such as time limits, oral instructions, practice problems, answer sheets and scoring formulas) should be clearly set out and followed in the search for validation. Failure to do this essentially prohibits generalisations from the search being applied to the operational context. The point of this principle is that for a search to enhance the general body of knowledge, the critical research procedures must be consistent with those that are to be utilised in practice.

There should be at least presumptive evidence for the validity of a predictor prior to its operational use. If possible, predictors should be validated prior to operational use. Some test users find this principle difficult to follow because of the need for the organisation to make employment decisions.

Where there is evidence from other studies or situations that makes valid prediction likely, it may be feasible to use the predictors immediately. The researcher should avoid situations that make it impossible or difficult to establish validity. For example, assessment decisions based on an unvalidated test should not be so highly selective that severe restriction of range results. Data required for correction of restriction of range should be collected and maintained. If there is no firm basis for the presumption of validity, the test user should judge whether the cost of postponing the use of the predictor is greater or less than the dangers of using it prematurely while collecting data.

Predictor data and criterion measures should be independent. If criterion ratings are collected from supervisors who know assessment test scores, the two sets of data are not independent. The resulting validity coefficient depends on both the true relationship and the manipulation of ratings (consciously or unconsciously) to conform to scores. Such contamination should be avoided.

2.2.2.7 Data Analysis for Criterion-Related Validity

The quality of the validation study depends as much on the appropriateness of the data analysis as on the data collected during the research. Researchers need to ensure that the statistics used are appropriate. Moreover, as with the choice of criterion or predictor variables, the researcher should not choose a data analysis method simply because the computer package for it is readily available. Researchers who delegate data analyses to others still retain responsibility for ensuring the suitability and accuracy of the analyses.

Strength of the predictor-criterion relationship. The analysis should provide information about effect sizes and the statistical significance or confidence associated with predictor-criterion relationships. Effect size estimates and confidence intervals can be useful in making professional judgments about the strength of predictor-criterion relationships (Schmidt, 1996). Other approaches such as expectancy tables are also useful in many situations, particularly if the assumptions of a correlational analysis are not met. Traditionally, a validity coefficient or similar statistic that has a probability of less than one in twenty of having occurred by chance if the true relationship is zero (Type I error) may be considered as establishing significant validity.

Research on the power of criterion-related validation studies and meta-analytic research suggests that achieving adequate power while simultaneously controlling Type I error rates can be problematic in a local validation study and may require sample sizes that are difficult to obtain. Researchers should give at least equal attention to the risks of Type II error (Type II errors being the failure to detect validity where with greater power in the analysis, correlations would be viewed as significant). When multiple tests of significance are conducted for predictors with no cumulative literature indicating evidence of validity, the test user should check for the possibility of observing some apparently significant relationships on a chance basis. When multivariate techniques are used, the number of cases should be large relative to the number of variables. The analysis should provide information about the strength of the relationship, usually a coefficient of correlation.

Reports of any analysis should provide information about the nature of the predictor-criterion relationship and how it might be used in prediction. The information should include number of cases, measures of central tendency, characteristics of distributions, and variability for both the predictor and criterion variables, as well as the interrelationships among all variables studied.

Adjustments to validity estimates. Researchers should obtain as unbiased an estimate as possible of the validity of the predictor in the population in which it is used. Observed validity coefficients may underestimate the predictor-criterion relationship due to the effects of range restriction and unreliability in the predictors or criteria. When range restriction causes underestimation of the validity coefficient, a suitable bivariate or multivariate adjustment should be made when the necessary information is available, (e.g. Schepers, 1995). Adjustment of the validity coefficient for criterion unreliability should be made if an appropriate estimate of criterion reliability can be obtained. Researchers should make sure that reliability estimates used in making corrections are appropriate to

avoid under- or overestimating validity coefficients. For example, in a study utilising a criterion-related strategy in which the criteria are performance ratings, differences between raters and differences across time may be considered in estimating criterion reliability because internal consistency estimates, by themselves, may be inadequate.

When adjustments are made, both unadjusted and adjusted validity coefficients should be reported. Researchers should be aware that the usual tests of statistical significance do not apply to adjusted coefficients such as those adjusted for restriction of range and/or criterion unreliability (Bobko & Riecke, 1980; Raju & Brand, in press; Raju, Burke, Normand, & Langlois, 1991). The adjusted coefficient is generally the best point estimate of the population validity coefficient; confidence intervals around it should be computed. No adjustment of a validity coefficient for unreliability of the predictor should be made or reported unless it is clearly stated that the coefficient is theoretical and cannot be interpreted as reflecting the actual operational validity of the selection procedure.

Combining predictors and criteria. Where predictors are used in combination, researchers should consider and document the method of combination. Predictors can be combined using weights derived from a multiple regression analysis (or another appropriate multivariate technique), unit weights, unequal weights that approximate regression weights, weights that are determined from work-analytic procedures, or weights based on professional judgment. Generally, after cross-validation, the more complex weighting procedures offer no or only a slight improvement over simple weighting techniques (Aamodt & Kimbrough, 1985). When combining scores, care must be taken to ensure that differences in the variability of different predictors do not lead to over- or under-weighting of one or more predictors.

Selection procedures that have linear relationships with work performance can be combined for use in either a linear manner (e.g., by summing scores on different selection procedures) or in a configural manner (e.g., by using multiple cutoffs). The researcher should be aware of the administrative, legal, and other implications of each choice. When configural selection rules are used, a clear rationale for their use should be provided (e.g., meeting larger organisational goals or needs, administrative convenience, or reduced testing costs).

Similarly, if the researcher combines scores from several criteria into a composite score, there should be a rationale to support the rules of combination and the rules of combination should be described. Usually, it is better to assign unit or equal weights to the several criterion components than to attempt to develop precise empirical weights. When measures are combined, researchers should recognise that effective weights (i.e., the contributions of the various components to the variance of the composite) are a function of a variable's standard deviation and are unlikely to be the same as the nominal weights. Effective weights depend not only on the nominal weights assigned to the components but also on their standard deviations and intercorrelations.

Cross-validation. Researchers should guard against overestimates of validity resulting from capitalisation on chance. Especially when the research sample is small, estimates of the validity of a composite battery developed on the basis of a regression equation should be adjusted using the appropriate shrinkage formula or be cross-validated on another sample. The assignment of either rational or unit weights to predictors does not result in shrinkage in the usual sense. Where a smaller number of predictors is selected for use based on sample validity coefficients from a larger number included in the study, shrinkage formulas can be used only if the larger number is entered into the formula as the number of predictors, though this will produce a slightly conservative estimate of the cross-validated multiple correlation.

Documenting and interpreting validation analyses. The results obtained using a criterion-related strategy should be interpreted against the background of the relevant research literature. Cumulative research knowledge plays an important role in any validation effort. A large body of research regarding relationships between many predictors and work performance currently exists (Schmidt & Hunter, 1998). Indeed, after consulting the literature a test user may conclude that the existing research

base is sufficient to support the use of certain instruments without any additional criterion-related research.

An extremely large sample or replication is required to give full credence to unusual findings. Such findings include, but are not limited to, suppressor or moderator effects, non-linear regression, and benefits of configural scoring. *Post hoc* hypotheses in multivariate studies and differential weightings of highly correlated predictors are particularly suspect and should be replicated before they are accepted and results implemented.

Establishing the practical value of the assessment. The judgment of the practical value of an assessment procedure is based on the evidence of validity, the assessment ratio, base rate/criterion variance, the number to be selected and the nature of the job. Expectancy tables may also be useful for this purpose, as can the Taylor-Russell and Naylor-Shine tables. The literature on the impact of assessment tests on the productivity of employees has provided estimates of utility that allow the use of regression equations (Brogden, 1949; Cronbach & Gleser, 1965; Dunnet, Hough & Triandis, 1991; Schmidt, Hunter, McKenzie & Muldrow, 1979; Boubreau, 1991). Both projected productivity gains per employee and projected total productivity gains resulting from the use of the assessment procedures are relevant in assessing their practical value (Boudreau & Berger, 1985; Cascio, 2000; Schmidt, Mack & Hunter, 1984). Utility analysis is discussed further in the section "Assessment Utility".

Data correctness. Data should be free from clerical error – data entry, coding, computational work and reports should be checked carefully and thoroughly.

2.2.3 EVIDENCE FOR VALIDITY BASED ON CONTENT

Evidence for validity based on content typically consists of a demonstration of a strong linkage between the content of the selection procedure and important work behaviours, activities, worker requirements, or outcomes on the job. This linkage also supports construct interpretation. When the selection procedure is designed explicitly as a sample of important elements in the work domain, the validation study should provide evidence that the selection procedure samples the important work behaviours, activities, and/or worker KSAOs necessary for performance on the job, in job training, or on specified aspects of either. This provides the rationale for the generalisation of the results from the validation study to prediction of work behaviours (Goldstein, Zedeck, & Schneider, 1993).

The content-based selection procedures discussed here are those designed as representative samples of the most important work behaviours, activities, and/or worker KSAOs drawn from the work domain and defined by the analysis of work. The content of the selection procedure includes the questions, tasks, themes, format, wording, and meaning of items, response formats, and guidelines regarding the administration and scoring of the selection procedure. The following provides guidance for the development or choice of procedures based primarily on content.

2.2.3.1 Feasibility of a Content-Based Validation Study

A number of issues may affect the feasibility of a content-based validation study and should be evaluated before beginning such a study. Among these issues are the stability of the work and the worker requirements, the interference of irrelevant content, the availability of qualified and unbiased subject matter experts, and cost and time constraints.

The researcher should consider whether the work and the worker requirements are reasonably stable. When feasible, a content-based selection procedure should remove or minimise content that is irrelevant to the domain sampled. Virtually any content-based procedure includes some elements that are not part of the work domain (e.g., standardisation of the selection procedure or use of response formats that are not part of the job content, such as multiple choice formats or written responses when the job does not require writing).

The success of the content-based validation study is closely related to the qualifications of the subject matter experts (SMEs). SMEs define the work domain and participate in the analysis of work by identifying the important work behaviours, activities, and worker KSAOs. The experts should have thorough knowledge of the work behaviours and activities, responsibilities of the job incumbents, and the KSAOs prerequisite to effective performance on the job. The SMEs should include persons who are fully knowledgeable about relevant organisational characteristics such as shift, location, type of equipment used, and so forth. A method for translating subject matter expert judgments into the selection procedure should be selected or developed and documented. If SME ratings are used to evaluate the match of the content-based procedure to the work and worker requirements, procedures and criteria for rating each aspect should be standardised and delineated.

Cost and time constraints can affect the feasibility of some content-based procedures. In some situations, designing and implementing a simulation that replicates the work setting or type of work may be too costly. In others, developing and assessing the reliability of the procedure may take too long because samples are too small or the behaviour is not easily measured using this strategy.

2.2.3.2 Design and Conduct of Content-Based Strategies

The content-based validation study specifically demonstrates that the content of the selection procedure represents an adequate sample of the important work behaviours, activities, and/or worker KSAOs defined by the analysis of work. This involves choosing SMEs, defining the content to be included in the selection procedure, developing the selection procedure, establishing the guidelines for administration and scoring, and evaluating the effectiveness of the validation effort.

2.2.3.3 Defining the Content Domain

The characterisation of the work domain should be based on accurate and thorough information about the work including analysis of work behaviours and activities, responsibilities of the job incumbents, and/or the KSAOs prerequisite to effective performance on the job. In addition, definition of the content to be included in the domain is based on an understanding of the work, and may consider organisational needs, labour markets, and other factors that are relevant to personnel specifications and relevant to the organisation's purposes. The domain need not include everything that is done on the job. The researcher should indicate what important work behaviours, activities, and worker KSAOs are included in the domain, describe how the content of the work domain is linked to the selection procedure, and explain why certain parts of the domain were or were not included in the selection procedure. A procedure may sample a given job performance domain, but if that domain is not an important part of the job, the value of the procedure for employment purposes is negligible.

The fact that the construct assessed by a selection procedure is labelled an ability does not *per se* preclude the reliance on a content-oriented strategy. When selection procedure content is linked to job content, content-oriented strategies are useful. When selection procedure content is less clearly linked to job content, other sources of validity evidence take precedence.

The selection procedure content should be based on an analysis of work that specifies whether the employee is expected to have all the important work behaviours, activities, and/or KSAOs before selection into the job or whether basic or advanced training will be provided after selection. If the intended purpose of the selection procedure is to hire or promote individuals into jobs for which no advanced training is provided, the researcher should define the selection procedure in terms of the work behaviours, activities, and/or KSAOs an employee is expected to have before placement on the job. If the intent of the content-based procedure is to select individuals for a training program, the work behaviours, activities, and/or worker KSAOs would be those needed to succeed in a training program. Because the intended purpose is to hire or promote individuals who have the prerequisite work behaviours, activities, and/or KSAOs to learn the work as well as to perform the work, the selection procedure should be based on an analysis of work that defines the balance between the work behaviours, activities, and/or KSAOs the applicant is expected to have before placement on the job

and the amount of training the organisation will provide. For example, the fact that an employee will be taught to interpret company technical manuals may mean that the job applicant should be evaluated for reading ability. A selection procedure that assesses the individual's ability to read at a level required for understanding the technical manuals would likely be predictive of work performance.

A content-based selection procedure may also include evidence of specific prior training, experience, or achievement. Recognition of prior learning – as referred to in the South African Qualifications Authority Act (No. 58 of 1995) – comes to mind here. This evidence is judged on the basis of the relationship between the content of the experience and the content of the work requiring that experience. To justify such relationships, more than a superficial resemblance between the content of the experience variables and the content of the work is required. For example, course titles and job titles may not give an adequate indication of the content of the course or the job or the level of proficiency an applicant has developed in some important area. What should be evaluated is the similarity between the behaviours, activities, processes performed, or the KSAOs required by the work.

2.2.3.4 Choosing the Selection Procedure

The development or choice of a selection procedure is usually restricted to important or frequent behaviours and activities or to prerequisite KSAOs. The researcher should have adequate coverage of work behaviours and activities and/or worker requirements from this restricted domain to provide sufficient evidence to support the validity of the inference. The fidelity of the selection procedure content to important work behaviours forms the basis for the inference.

Sampling the content domain. The process of constructing or choosing the selection procedure requires sampling the work content domain. Not every element of the work domain needs to be assessed. Rather, a sample of the work behaviours, activities, and worker KSAOs can provide a good estimate of the predicted work performance. Sampling should have a rationale based on the professional judgment of the researcher and an analysis of work that details important work behaviours and activities, important components of the work context, and KSAOs needed to perform the work. Random sampling of the content of the work domain is usually not feasible or appropriate. The rationale underlying the sampling should be documented.

Describing the level of specificity. In defining the work content domain, the degree of specificity needed in a work analysis and a selection procedure should be described in advance. The more a selection procedure has fidelity to exact job components, the more likely it is that the content-based evidence will be demonstrated. However, when the work changes and fidelity drops, the selection procedure is less likely to remain appropriate. Thus, considering the extent to which the work is likely to change is important. If changes are likely to be frequent, the researcher may wish to develop a selection procedure that has less specificity. For example, in developing a selection procedure for the job of word processor, the procedure may exclude content such as "demonstrates proficiency with a particular word processing program" and instead include content that is less specific, such as "demonstrates proficiency with word processing principles and techniques."

The degree to which the results of validation studies can be generalised depends in part on the specificity of the selection procedure and its applicability across settings, time, and jobs. While general measures may be more resilient to work changes and more transferable to other, similar work, they also may be subject to more scrutiny because the correspondence between the measure and the work content is less detailed.

2.2.3.5 Procedural Considerations

The researcher needs to establish the guidelines for administering and scoring the content-based procedure. Typically, defining the administration and scoring guidelines for a paper-based procedure that measures job-related knowledge or cognitive skills is relatively uncomplicated. On the other hand,

including a work behaviour or activity in the content-based selection procedure may introduce administration and scoring challenges, which should be evaluated in advance. Generally, the more closely a selection procedure replicates a specific work behaviour, the more accurate the content-based inference. Yet, the more closely a selection procedure replicates a work behaviour, the more difficult the procedure may be to administer and score.

For example, troubleshooting multistep computer problems may be an important part of a technical support person's work. It may be difficult, however, to develop and score a multistep troubleshooting simulation or work sample, because examinees may not use the same steps or strategy when attempting to solve the problem. A lower fidelity alternative such as single-step problems could be used so that important aspects of the work domain are still included in the selection procedure. In all cases, the researcher should ensure that the procedures are measuring skills and knowledge that are important in the work rather than irrelevant content.

2.2.3.6 Evaluating Content-Related Evidence

Evidence for validity based on content rests on demonstrating that the selection procedure adequately samples and is linked to the important work behaviours, activities, and/or worker KSAOs defined by the analysis of work. The documented methods used in developing the selection procedure constitute the primary evidence for the inference that scores from the selection procedure can be generalised to the work behaviours and can be interpreted in terms of predicted work performance. The sufficiency of the match between selection procedure and work domain is a matter of professional judgment based on evidence collected in the validation effort (Goldstein et al., 1993).

Reliability of performance on content-based selection procedures should be determined when feasible. If ratings from more than one rater are used to evaluate performance on a simulation or work sample, the researcher should evaluate inter-rater agreement in operational use.

2.2.4 EVIDENCE OF VALIDITY BASED ON INTERNAL STRUCTURE

Information about the internal structure of any selection procedure can also support validation arguments. Internal structure evidence alone is not sufficient evidence to establish the usefulness of a selection procedure in predicting future work performance. However, internal structure is important in planning the development of a selection procedure. The specific analyses that are relevant depend on the conceptual framework of the selection procedure, which in turn is typically established by the proposed use of the procedure.

When evidence of validity is based on internal structure, the researcher may consider the relationships among items, components of the selection procedures, or scales measuring constructs. Inclusion of items in a selection procedure should be based primarily on their relevance to a construct or content domain on their intercorrelations. Well-constructed components or scales that have near-zero correlations with other components or scales, or a total score, should not necessarily be eliminated. For example, if the selection procedure purposely contains components relevant to different construct or content domains (e.g., a selection battery composed of a reading test, an in-basket, and an interview), the scores on these components may not be highly correlated.

However, if the conceptual framework posits a single dimension or construct, one should strive for a high level of homogeneity among the components, which can be evaluated in terms of various internal consistency estimates of reliability. If the intent of the conceptual framework requires a more complex internal structure, overall internal consistency might not be an appropriate measure. For example, the internal consistency reliability estimate for a performance rating form involving several supposedly unrelated scales might only represent halo effect.

When scoring involves a high level of judgment on the part of those doing the scoring, indices of inter-rater or scorer consistency, such as generalisability coefficients or measures of inter-rater agreement, may be more appropriate than internal consistency estimates.

2.3 GENERALISING VALIDITY EVIDENCE

At times, sufficient accumulated validity evidence is available for a selection procedure to justify its use in a new situation without conducting a local validation research study. In these instances, use of the selection procedure may be based on demonstration of the generalised validity inferences from that selection procedure, coupled with a compelling argument for its applicability to the current situation. Although neither mutually exclusive nor exhaustive, several strategies for generalising validity evidence have been delineated, namely: (a) transportability, (b) synthetic validity/job component validity, and (c) meta-analytic validity generalisation.

2.3.1 TRANSPORTABILITY

One approach to generalising the validity of inferences from scores on a selection procedure involves the use of a specific selection procedure in a new situation based on results of a validation research study conducted elsewhere. This is referred to as demonstrating the "transportability" of validity evidence for the selection procedure. When proposing to "transport" use of a procedure, a careful review of the original validation study is warranted to ensure acceptability of the technical soundness of that study and to determine its relevance to the new situation. Key points for consideration when establishing the appropriateness of transportability are, most prominently, job comparability in terms of content or requirements, as well as, possibly, similarity of job context and candidate group.

2.3.2 SYNTHETIC VALIDITY/JOB COMPONENT VALIDITY

A second approach to establishing generalised validity of inferences based on scores from a selection procedure is referred to as synthetic validity or job component validity. While some researchers distinguish these terms, others do not, and in either case several variations on each exist. A defining feature of synthetic validity/job component validity is the justification of the use of a selection procedure based upon the demonstrated validity of inferences from scores on the selection procedure with respect to one or more domains of work (job components). Thus, establishing synthetic validity/job component validity requires documentation of the relationship between the selection procedure and one or more specific domains of work (job components) within a single job or across different jobs. If the relationship between the selection procedure and the job component(s) is established, then the validity of the selection procedure for that job component may be generalisable to other situations in which the job components are comparable.

The validity of a selection procedure may be established with respect to different domains (components) of work, then "synthesised" (combined) for use based on the domains (or components) of work relevant for a given job or job family. In some instances, this may involve conducting a research study designed to demonstrate evidence for the generalised validity of inferences from scores on a set of selection procedures, and then using various subsets of these procedures for selection into both jobs or job families in the original study as well as into other jobs or job families. In other cases, it may involve generalising the validity of inferences based on scores on selection procedures examined in one or more research studies conducted elsewhere to the new situation. In both cases, detailed analysis of the work is required for use of this strategy of generalising validity evidence.

2.3.3 META-ANALYSIS

Meta-analysis is a third procedure and strategy that can be used to determine the degree to which predictor-criterion relationships are specific to the situations in which the validity data have been gathered or are generalisable to other situations, as well as to determine the sources of cross-situation variability (Aguinis & Pierce, 1998). Meta-analysis requires the accumulation of findings from a number of validity studies to determine the best estimates of the predictor-criterion relationship for the kinds of work domains and settings included in the studies.

While transportability and synthetic validity/job component validity efforts may be based on an original study or studies that establish the validity of inferences based on scores from the selection procedure through a content-based and/or a criterion-related strategy**,** meta-analysis is a strategy that can only be applied in cases in which the original studies relied upon criterion-related evidence of validity**.** The question to be answered using a meta-analytic strategy is whether the valid inferences about work behaviour or job performance can be drawn from predictor scores across given jobs or job families in different settings. (Note that the focus here is on using meta-analysis to examine predictor-criterion relationships. Meta-analysis also can be used to examine other issues, such as convergence among instruments intended to measure the same construct.)

Meta-analysis is the basis for the technique that is often referred to as "validity generalisation." In general**,** research has shown much of the variation in observed differences in obtained validity coefficients in different situations can be attributed to sampling error and other statistical artifacts (Ackerman & Humphreys, 1990; Barrick & Mount, 1991; Callender & Osburn, 1980; 1981; Hartigan & Wigdor, 1989; Hunter & Hunter, 1984; Schmidt, Hunter, & Pearlman, 1981). These findings are particularly well-established for cognitive ability tests; additional recent research results also are accruing that indicate the generalisability of predictor-criterion relationships for noncognitive constructs in employment settings (e.g. Barrick et al, 2001; Van der Walt et al, 2002).

Professional judgment in interpreting and applying the results of meta-analytic research is important. Researchers should consider the meta-analytic methods used and their underlying assumptions, the tenability of the assumptions, and artifacts that may influence the results (Bobko & Stone-Romero, 1998; Raju, Anselmi, Goodman, & Thomas, 1998; Raju et al., 1991; Raju, Pappas, & Williams, 1989). In evaluating meta-analytic evidence, the researcher should be concerned with potential moderators to the extent that such moderators would affect conclusions about the presence and generalisability of validity. In such cases, researchers should consider both statistical power to detect such moderators and/or the precision of estimation with respect to such moderators. In addition, the researcher should consider the probabilities of both Type I and Type II decision errors (Oswald & Johnson, 1998; Sackett, Harris, & Orr, 1986).

Reports that contribute to the meta-analytic research results should be clearly identified and available. Researchers should consult the relevant literature to ensure that the meta-analytic strategies used are sound and have been properly applied, that the appropriate procedures for estimating predictor-criterion relationships on the basis of cumulative evidence have been followed, that the conditions for the application of meta-analytic results have been met, and that the application of meta-analytic conclusions is appropriate for the work and settings studied. The rules by which the researchers categorised the work and jobs studied, the selection procedures used, the definitions of what the selection procedure is measuring, the job performance criteria used, and other study characteristics that were hypothesised to impact the study results should be fully reported. The quality of the individual research studies and their impact, if any, on the meta-analytic conclusions and their use also should be informed by good professional judgment (Guion, 1998; Law, Schmidt, & Hunter, 1994a, 1994b).

Note that sole reliance upon available cumulative evidence may not be sufficient to meet specific employer operational needs such as for the placement of employees or for the optimal combination of procedures. Consequently, additional studies and data may be required to meet these specific needs. If such studies are not feasible in an organisation, researchers and employers may engage in cooperative studies.

Meta-analytic methods for demonstrating generalised validity are still evolving. Researchers should be aware of continuing research and critiques that may provide further refinement of the techniques as well as a broader range of predictor-criterion relationships to which meta-analysis has been applied.

Generalising validity evidence from meta-analytic results is often more useful than a single study. However, if important conditions in the operational setting are not represented in the meta-analysis

(e.g., the local setting involves a managerial job and the meta-analytic data base is limited to entry-level jobs), a local individual study may be more accurate than the average predictor-criterion relationship reported in a meta-analytic study. A competently conducted study, with a large sample using the same test, for the same kind of work activities, may be more accurate, informative, and useful than a cumulation of small validation studies that are not representative of the setting to which one wants to generalise validity.

Reliance on meta-analytic results is more straightforward when they are organised around a construct or set of constructs. When different predictors (as well as different criteria) intended to measure the same construct are combined in a meta-analysis, findings are meaningful to the extent that there is evidence that they do indeed reflect the same construct (e.g., convergent validity evidence). If, for example, meta-analytic evidence relies on data from five highly correlated, published measures of a predictor construct, these findings cannot be assumed to generalise to other measures using the same construct label without evidence that those other measures indeed reflect the same construct.

When studies are cumulated on the basis of common methods (e.g., interviews, biodata) instead of constructs, a different set of interpretational difficulties arises. Generalisation is straightforward when, for example, an empirical biodata scale has been developed for a specific occupation, multiple validity studies have been conducted using that scale in that occupation, and the intent is to generalise to another setting that employs individuals in that same occupation. However, researchers may have difficulty when they attempt to generalise about a method in general, rather than about a specific application of the method. Because methods such as the interview can be designed to assess widely varying constructs (from job knowledge to integrity), generalising from cumulative findings is only possible if the features of the method that result in positive method-criterion relationships are clearly understood, if the content of the procedures and meaning of the scores are relevant for the intended purpose, and if generalisation is limited to other applications of the method that include those features. Consider the accumulation of validity findings from various interview methods, where in the population of settings in which interviews are used, the interview development process, content, and scoring vary (e.g., some knowledge-focused and some value-focused; some structured and some unstructured). Now consider a setting in which these features have not been coded, and thus it is unclear whether these features vary in the sample of studies available for meta-analysis. Generalising from a meta-analysis of such data to a new similarly unspecified interview, to a different interview method, or to a different or new situation, is not warranted. For example, it may be the case that all studies in the database involve knowledge-focused interviews, and consistency in validity across knowledge-focused interviews offers no grounds for inferring that validity will generalise to value-focused interviews. In contrast, a cumulative database on interviews where content, structure, and scoring are coded could support generalisation to an interview meeting the same specifications.

2.4 FAIRNESS AND BIAS

Fairness is a social rather than a psychometric concept. Its definition depends on what one considers to be fair. Fairness has no single meaning and, therefore, no single definition, whether statistical, psychometric, or social. Fairness or the lack thereof is not the result of the assessment instrument or predictor, nor is it the property of the assessment procedure used. Fairness is the total of all the variables that play a role or influence the final personnel decision. This may include the test, predictor, and integration of data, recommendations based on these data or the final decision made by line management.

2.4.1 FAIRNESS DEFINED

The *Standards* notes four possible meanings of "fairness."

The first meaning of fairness as described by the *Standards* views fairness as requiring equal group outcomes (e.g., equal passing rates for subgroups of interest). The *Standards* rejects this definition, noting that it has been almost entirely repudiated in the professional testing literature. It notes that

while group differences should trigger heightened scrutiny for possible sources of bias (i.e., a systematic error that differentially affects the performance of different groups of test takers), outcome differences in and of themselves do not indicate bias. It further notes that there is broad agreement that examinees with equal standing on the construct of interest should, on average, earn the same score regardless of group membership.

The second meaning views fairness in terms of the equitable treatment of all examinees. Equitable treatment in terms of testing conditions, access to practice materials, performance feedback, retest opportunities, and other features of test administration, including providing reasonable accommodation for test takers with disabilities when appropriate, are important aspects of fairness under this perspective. There is consensus on a need for equitable treatment in test administration (although not necessarily on what constitutes equitable treatment).

The third meaning views fairness as requiring that examinees have a comparable opportunity to learn the subject matter covered by the test. However, the *Standards* notes that this perspective is most prevalent in the domain of educational achievement testing and that opportunity to learn ordinarily plays no role in determining the fairness of employee selection procedures. One exception would be settings where the organisation using the tests purposely limits access to information needed to perform well on the tests on the basis of group membership. In such cases, while the test itself may be unbiased in its coverage of job content, the use of the test would be viewed as unfair under this perspective.

The fourth meaning views fairness as a lack of predictive bias. This perspective views predictor use as fair if a common regression line can be used to describe the predictor-criterion relationship for all subgroups of interest; subgroup differences in regression slopes or intercepts signal predictive bias. There is broad scientific agreement on this definition of predictive bias, but there is no similar broad agreement that the lack of predictive bias can be equated with fairness. For example, a selection system might exhibit no predictive bias by race or gender, but still be viewed as unfair if equitable treatment (e.g., access to practice materials) was not provided to all examinees.

An alternative form of fairness may exist as a result of the perceptions of the participants being rated. Indeed, since South African job applicants are regarded as employees with certain legal rights, de Jong and Visser (2000) point out that it is essential that applicants' perceptions of fairness of selection procedures be studied. A study by Rademan and Vos (2001) on performance appraisals in the public sector demonstrated a significant difference between supervisors and subordinates regarding their perceptions of the appraisal process. This refers to whether subordinates felt that they were being treated fairly through the system and whether supervisors applied the given system in a manner perceived to be fair. In another study conducted by de Jong and Visser (2000) in terms of black and white fairness perceptions of ten selection techniques, it was found that while some group differences exist, the preference of both the black and white groups is for objective rather than subjective selection techniques.

There are multiple perspectives on fairness. There is agreement that issues of equitable treatment, predictive bias, and scrutiny for possible bias when subgroup differences are observed, are important concerns in personnel selection; there is not, however, agreement that the term “fairness” can be uniquely defined in terms of any of these issues.

2.4.2 BIAS

The *Standards* notes that bias refers to any construct-irrelevant source of variance that results in systematically higher or lower scores for identifiable groups of examinees. The effect of such irrelevant sources of variance on scores on a given variable is referred to as measurement bias. The effects of such sources of variance on predictor-criterion relationships, such that slope or intercepts of

the regression line relating the predictor to the criterion are different for one group than for another, is referred to as predictive bias. The *Standards* notes that, in the employment context, evidence of bias or lack of bias generally relies on the analysis of predictive bias. Both forms of bias are discussed below.

2.4.2.1 Predictive Bias

While fairness has no single accepted meaning, there is agreement as to the meaning of predictive bias. Predictive bias refers to the usefulness, validity and fairness of the test for the purpose for which it was designed, and is found when for a given subgroup, consistent nonzero errors of prediction are made for members of the subgroup (Cleary, 1968; Humphreys, 1952). (Another term used to describe this phenomenon is differential prediction. The term "differential prediction" is sometimes used in the classification and placement literature to refer to differences in predicted performance when an individual is classified into one condition rather than into another; this usage should not be confused with the use of the term here to refer to predictive bias.) Although other definitions of bias have been introduced, such models have been critiqued and found wanting on grounds such as lack of internal consistency (Petersen & Novick, 1976).

Testing for predictive bias involves using moderated multiple regression, where the criterion measure is regressed on the predictor score, subgroup membership, and an interaction term between the two. Slope and/or intercept differences between subgroups indicate predictive bias (Gulliksen & Wilks, 1950; Lautenschlager & Mendoza, 1986; Nunnally & Bernstein, 1994).

American literature has extensively examined predictive bias in the cognitive ability domain. For White–African American and White–Hispanic comparisons, slope differences are rarely found; while intercept differences are not uncommon, they typically take the form of overprediction of minority group performance (Bartlett, Bobko, Mosier, & Hannan, 1978; Hunter, Schmidt, & Rauschenberger, 1984; Schmidt, Pearlman, & Hunter, 1980). In some other domains, there has been little to no published research on predictive bias, though work in the personality domain is now beginning to appear. Saad and Sackett (2002) report findings parallel to those in the ability domain in examining predictive bias by gender using personality measures (i.e., little evidence of slope differences and intercept differences in the form of over-prediction of female performance). Given the limited research to date, broad conclusions about the prevalence of predictive bias for many constructs are premature at this time.

There is insufficient literature in the South African context to support these American findings, and more research is needed on the different ethnic groups. It has been noted that the cultural distance between American whites and blacks is very different to that between South African whites and blacks, which might be used as an argument for why differential prediction may exist or manifest itself differently in the South African context.

Owen (1989), however, found that the Junior Aptitude Test measures the same constructs in white, Indian and black pupils, but that item bias might exist (de Villiers, 1997; Wheeler, 1993). A study conducted by Meiring et al (2005) examined bias at the level of constructs (structural equivalence) and items (item bias) in two cognitive tests and a personality questionnaire. While the researchers found good construct equivalence and low item bias for the cognitive instruments, various scales of the personality questionnaire revealed construct bias. The item bias for the personality questionnaire was low. In another study conducted by Kriek, Hurst and Charoux (1994), the fairness of assessment centre technology across race was explored. The researchers found the assessment centre predicted job performance for both the white and black groups, and there was no evidence of bias in the predictive validity of the assessment centre.

From the above it can be concluded that there is still a need to research issues of bias among the different ethnic groups in South Africa.

Several important technical concerns with the analysis of predictive bias are noted here. The first is that an analysis of predictive bias requires an unbiased criterion. Confidence in the criterion measure is a prerequisite for an analysis of predictive bias. It is important to note, though, that while researchers should exercise great care in the development and collection of criterion data, investigations of criterion bias are limited by the lack of a true score against which criterion measures can be compared. The second is the issue of statistical power to detect slope and intercept differences. Small total or subgroup sample sizes, unequal subgroup sample sizes, range restriction, and predictor unreliability are factors contributing to low power (Aguinis, 1995; Aguinis & Stone-Romero, 1997). A third is the assumption of homogeneity of error variances (Aguinis, Peterson, & Pierce, 1999); alternative statistical tests may be preferable when this assumption is violated (Alexander & DeShon, 1994; DeShon & Alexander, 1996; Oswald, Saad, & Sackett, 2000).

Some perspectives view the analysis of predictive bias as an activity contingent on a finding of mean subgroup differences. In fact, however, subgroup differences and predictive bias can exist independently of one another. Thus, whether or not subgroup differences on the predictor are found, predictive bias analysis should be undertaken when there are compelling reasons to question whether a predictor and a criterion are related in a comparable fashion for specific subgroups, given the availability of appropriate data. In domains where relevant research exists, generalised evidence can be appropriate for examining predictive bias.

2.4.2.2 Measurement Bias

Measurement bias, namely, sources of irrelevant variance that result in systematically higher or lower scores for members of particular groups, is a potential concern for all variables, both predictors and criteria. Meiring et al (2005) note how studies in South Africa have reported race, education, language and understanding of English as the main reasons impacting on construct and item comparability (Abrahams, 1996, 2002; Abrahams & Mauer, 1999a, 1999b; Claasen, 1997; Meiring, 2000; Meiring, van De Vijver, Rothman & Barrick, 2005).

Determining whether measurement bias is present is often difficult, as this requires comparing an observed score to a true score. In many domains, such as performance appraisal, such a standard for comparison is generally unavailable.

An approach to examining measurement bias in the domain of multi-item tests is to perform a differential item functioning (DIF) analysis. DIF refers to analyses that identify items for which members of different subgroups with identical total test scores (or identical estimated true scores in item response theory [IRT] models) have differing item performance. Such analyses are uncommon in the employment domain. First, they require data on large research samples prior to operational use, as DIF analyses are often part of the predictor development process. Second, empirical research in domains where DIF analyses are common has rarely found sizable and replicable DIF effects (Sackett, Schmitt, Ellingson, & Kabin, 2001). Third, such analyses require unidimensional tests, and many employment tests are not factorially pure unidimensional tests, and the unidimensionality assumption is often untested in DIF research (Hunter & Schmidt, 2000). Fourth, for cognitive tests it is common to find roughly equal numbers of differentially functioning items favouring each subgroup, resulting in no systematic bias at the test level (Hunter & Schmidt, 2000). As a result of these factors, DIF findings should be viewed with caution. DIF analysis is not likely to become a routine or expected part of the test development and validation process in employment settings; however, researchers may choose to explore DIF when data sets appropriate for such analysis are available. In addition, when undertaking bias and equivalence investigations, it is recommended that a variety of data analysis techniques (e.g. DIF, factor analysis, weighted multidimensional scaling, structural equation modelling) be used before a decision is reached as to whether a measure is biased. Van de Vijver and Leung (1997) present a comprehensive discussion of definitions and other analysis techniques on bias.

Linked to the idea of measurement bias in terms of conducting analysis at the item level is the concept of an item sensitivity review, in which items are reviewed by individuals with diverse perspectives for

language or content that might have differing meaning for members of various subgroups and language that could be demeaning or offensive to members of various subgroups. Instructions to candidates and to scorers or assessors also may be reviewed in a similar manner. The value of such analysis will vary by test content, and the need for and use of such information is a matter of researcher judgment in a given situation.

2.4.2.3 Models of Test Fairness

A personnel decision may be considered fair in as far as it measures and predicts the criteria which it proposes to measure, and that it does so accurately and reliably. This refers specifically to the validity and reliability of the assessment instrument.

Arvey and Faley (1988) identify various models of test fairness that are aimed at identifying ways of determining the fairness of a given test, including the systematic mean difference between subgroups on tests model, differences in validity model, differences in regression lines model, conditional probability model and the equal risk model. Huysamen (1995) also provides a discussion reviewing several of these models. However, no agreement has been reached with regard to which is the correct or best model, as each researcher and test user will identify the model that best suits their objectives, and which fits their definition of fairness in assessment and in the use of psychological assessment.

Whatever the case may be, the need to correct past injustices remains a prevalent issue in South African society, and particularly in the workplace. Who, in fact, benefits from these efforts again depends on the principles and values adhered to by the individuals or companies making the decisions.

It is, however, recommended that only fairness models based on the regression model should be used in studies investigating the fairness of assessment procedures.

2.4.2.4 Ensuring Test Fairness

It is recommended that those professionals involved in personnel decisions should establish clear and job-related criteria for assessment prior to establishing the assessment procedure and tools to be used. It is also recommended that a model of fairness be established, according to which assessment procedures may be planned and carried out.

The essence of this is that the industrial psychologist or human resources practitioner should establish the job-relatedness of all the criteria that are used in any personnel decision. The only way that this can be achieved is by conducting proper and thorough job analysis to establish the required KSAOs for the job in question.

The goal is equity and fairness. This refers not only to being fair, but to be seen as being fair, as the perception of fairness is a subjective experience. The emphasis may therefore also fall on consultation with all stakeholders in developing decision-making policies and evaluating assessment materials and criteria for assessment before any personnel decisions are taken.

Arvey et al (1992), perhaps realistically, state that decision errors will always occur, as there is no such thing as a perfect assessment procedure. Occupational and job-related assessment methods do, however, provide a more objective means of assessing and selecting candidates in the workplace, as long as the assessment tool is valid and reliable for the purpose for which it is being used.

2.4.2.5 Adverse Impact

Adverse impact is also a related concept that will become more common in the South African context. Adverse impact refers to the possible impact of the assessment or test results on personnel decisions regarding different designated groups. The problem arises when a significant difference is found between the average test or assessment performance of different cultural or gender groups. In the

absence of validation evidence, there is likely to be a presumption that the group with the lower average performance was being indirectly discriminated against. That is, if the same entry standard were demanded of all applicants, the lower scoring group would find it harder to comply with the requirement. This is generally referred to as Adverse Impact.

Positive validation evidence for the test generally justifies the use of the test and rules out the possibility of unfair discrimination. By showing that those who perform poorly on the test also perform poorly on the job, a positive validation result confirms that rejecting low scorers is reasonable. This means that the differences between the groups will not lead to unfair discrimination but the differences can be justified on the basis of an inherent requirement of the job. This, however, means that you still have adverse impact but the resulting impact can be justified.

In general, however, the greater the degree of adverse impact resulting from the use of a test or assessment the higher the validity of the test should be to justify its use. The alternative would be to search for an alternative test with the same validity or accuracy of prediction, but with less adverse impact. Personality questionnaires for example demonstrate less adverse impact than ability tests and some organisations give more weight to personality measures in decision-making models to overcome the adverse impact problems associated with ability tests.

There remains the possibility that overall validity is masking cases where a test has poorer or no predictive validity for some groups, or that group differences in test scores are not reflected in job performance. Much research into these issues in other countries and in particular in the United States, covering many types of tests and a wide range of occupational fields has indicated that such scenarios are extremely rare, if they exist at all. This is especially so when good test practice has been followed. There is, however, a lack of published studies in this arena for South Africa and more work still needs to be done. The South African studies on this topic tend to find similar results to that of other countries.

2.5 ASSESSMENT UTILITY

The inclusion of the human resource function in the family of organisational functions is justified by its commitment to contribute towards the primary organisational objective of maximising the value of the organisation for its owners. It therefore follows logically that all interventions initiated by the human resource function should, in the final analysis, also be evaluated against the yardstick of profitability. The design, implementation and operation of human resource interventions thus only makes sense from an institutional perspective if a satisfactory return on the capital invested in the intervention is achieved over the period in which the intervention generates its effect. There thus rests an obligation on the human resource function to prove through appropriate financial indicators (Boudreau, 1991; Cronshaw & Alexander, 1985) that its interventions do add value to the organisation.

Assessment utility models serve as reality simplifying conceptual frameworks designed as aids for depicting, estimating and explaining the usefulness of assessment decision strategies, and suggesting how that information can be used to improve assessment strategies. When viewed from a historical perspective, the evolution of assessment utility models presents a fairly systematic progression from somewhat unsophisticated models to detailed, complex and rather daunting contemporary models (Boudreau, 1991). A number of utility models can be differentiated in terms of their interpretation of utility/payoff. All utility models should ideally be used to establish the business necessity of an assessment procedure.

The utility analysis model with the longest history defines payoff in terms of the validity coefficient. In terms of this classical model, the utility of an assessment strategy is solely a function of the correlation between a weighted, linear composite of predictors and a criterion measure. Assessment utility is thus equated with prediction accuracy defined in terms of the residual criterion variance. Over and under prediction are regarded as equally undesirable, irrespective of the position on the criterion

scale where they occur. Two translations of the validity coefficient are typically applied to convey its utility implications. The index of forecasting efficiency indicates the proportional reduction in the standard error of estimate of criterion scores predicted by the regression of the criterion on the weighted linear composite of predictors compared with the standard error of estimate of criterion scores predicted by the criterion mean. The coefficient of determination, or the square validity coefficient, reflects the proportion of variance in the criterion measure accounted for by the weighted linear composite of assessment predictors. Both these indexes lead to the rather disheartening conclusion that only assessment strategies with validities exceeding those normally obtained in validation studies will have substantial practical utility. The fundamental problem with this line of reasoning, however, lies in its complete disregard for the fact that criterion estimates are not desired as an end in themselves, but rather as necessary information required to arrive at a qualitative decision. Precision in criterion estimation is important in human resource assessment, but only in as far as it affects the quality of decision-making.

The following utility models could be considered in utility analysis. The Taylor-Russell utility model defines payoff in terms of the success ratio (Taylor & Russell, 1939). The Naylor-Shine utility model defines payoff in terms of the expected standardised criterion score of the selected group of applicants on the continuous criterion scale (Naylor & Shine, 1965). The Brogden-Cronbach-Gleser (B-C-G) assessment utility model defines payoff in terms of a monetary valued criterion (Brogden, 1946; Brogden, 1949; Cronbach & Gleser, 1965).

Utility analysis is a developing literature. Wisdom suggests that all estimates of values to be placed in equations should be chosen so as to lead to conservative estimates of utility. In addition to these conservative point estimates, minimum and maximum estimates should be presented when appropriate and needed.

Utility analysis should only be conducted once the fairness of an assessment procedure has been examined. Changes in the assessment decision rule to enhance the fairness of the assessment procedure will, in the case of the regression-based interpretations of fairness, result in increases in the various utility estimates; it pays to select fairly.

2.6 OPERATIONAL CONSIDERATIONS IN PERSONNEL SELECTION

This section of the *Guidelines* describes operational issues associated with the development or choice of a selection procedure, the conduct or accumulation of research to support the validity inferences made, documentation of the research effort in technical reports and administration manuals, and subsequent implementation and use. The need for sound professional judgment based on the extant scientific literature and the researcher's own experience will be required at every step of the process. In addition, all aspects of the research described in the *Guidelines* should be performed in compliance with the standards of the HPCSA's Ethical Code of Professional Conduct.

Topics are introduced in an order that generally corresponds to the temporal progression of the validation effort. For example, the section on understanding work and worker requirements precedes decisions regarding the selection procedure. In other cases, the placement is based on the logical relationship among the topics. Therefore, the order in which steps are taken in practice is ultimately a matter of professional and scientific judgment based on the given situation. It is recognised that in some instances a selection procedure may be implemented at the same time the validation process is underway.

2.6.1 INITIATING A VALIDATION EFFORT

The researcher works collaboratively with representatives of the organisation to define its needs and objectives, identify organisational constraints, plan the research, and communicate with major stakeholders regarding aspects of the process that will involve or influence them.

2.6.1.1 Defining the Organisation's Needs, Objectives, and Constraints

Researchers use their expertise and experience to assist the organisation in refining its goals and objectives. Different departments of the organisation may have different and sometimes competing and conflicting objectives. For instance, one department may prefer rigorous selection standards even though they create hardships for the staffing department responsible for recruiting qualified applicants. As another example, one organisation may need a large number of entry-level workers where there is minimal opportunity to move upward. In another organisation, the focus may be on hiring a few individuals with the capacity to move upward in a relatively short period of time. In all situations, the researcher and the organisation's representatives should factor in the desires of the various stakeholders and determine the relative weights to be given to each point of view.

The researcher provides accurate information regarding the benefits and limitations of various strategies in meeting the organisation's goals based on past experience and the extant scientific research. The researcher is encouraged to work with all departments (e.g., human resources, labour relations, legal) that may have an effect on or be affected by the selection procedure and stakeholders (e.g., internal or external individuals and groups such as labour organisations, advocacy groups, customers).

Climate and culture. Researchers face the challenge of ensuring high quality selection procedures in the context of the organisation's history and current environment regarding employment-related strategies and practices as well as the cultural setting in which it operates. Organisations operate in complex environments that sometimes place extreme and conflicting pressures on the management team. Researchers must consider the attitudes and commitments of organisation leaders and employees who are faced with intense competition, mergers, and other corporate events that may influence the relative importance of selection research in their view. Researchers also may need to take into account the legal and labour environment when deciding on validation approaches or selection instruments. In addition, global selection systems should take into consideration locally accepted practices.

Workforce size and availability. The number of individuals who currently perform the work and their similarity to the applicant population can be important considerations when designing the validation strategy. The number of workers may shape the validation strategy pursued (e.g., validity generalisation, content-oriented strategy) as well as affect the feasibility and method for pilot testing procedures.

Even when the number of workers is sufficient, their availability and willingness to participate in a validation study may be limited. For example, organisational needs may require that a core group of workers be present on the job at all times; labour organisations may influence the number and type of persons willing to participate in the research; and workers who have experienced organisational restructuring may be sceptical about the purpose of the research and its effect on their own positions.

Large discrepancies in the capabilities of incumbents and the available applicant pool also present challenges, particularly in establishing norms and setting cutoff scores. For example, organisations that have a more capable work force than applicant pool may find cutoff scores based on the performance of incumbents on the selection procedure inappropriate for applicants. Similarly, organisations seeking to upgrade the skills of their current workforce may need other sources of information for setting cutoff scores.

Sources of information. Sources of information needed for the validation and implementation efforts include, but are not limited to, the workers themselves, managers, supervisors, trainers, customers, archival records, databases, and research reports internal and external to the organisation. Based on the complexity of the work, the climate, and organisational constraints, some sources of information may be preferred over others. In some situations, the preferred source of information may not be available. Depending on the organisational constraints, alternatives to the researcher's preferred source of information may be required. Alternative sources also may be used to supplement information gathered from the preferred source.

Acceptability of selection procedures. Most organisations want selection procedures that are predictive, easy to use, cost effective, and legally defensible. However, there are often additional considerations. For example, an organisation's past experiences with respect to certain types of selection procedures may influence its decisions. Selection procedures that have been challenged in the past may not be acceptable to organisations, particularly if the organisation was not successful in defending them. In addition, selection procedures that are viewed as controversial by individuals, labour organisations, or other stakeholders may not be acceptable.

Some organisations find certain types of selection procedure questions unacceptable. For example, some biodata and personality inventory items (e.g., childhood experiences, personal interests) may be viewed as an invasion of privacy, even if they can be shown to be related to the criterion measures or the requirements of the job.

Some organisations prefer selection procedures that provide information regarding the strengths and developmental needs of the test taker. Procedures that measure constructs that can be learned (e.g., keyboarding or word processing) may be preferred over procedures that elicit information concerning previous life experiences or stable personality traits. Procedures that appear more relevant or face valid to the organisation may be more acceptable to the stakeholders than other procedures that relate to a less obvious construct regardless of any empirical evidence of validity. However, face validity is not an acceptable substitute for other forms of validity evidence as treated in the *Guidelines*. Although acceptability is important, it is just one of many factors to consider when selecting or designing a selection procedure. Nevertheless, the researcher should explain to decision-makers issues underlying selection procedure acceptability as part of the initial planning effort.

2.6.1.2 Communicating the Validation Plan

Both management and workers need to understand in general terms the purpose of the research, the plan for conducting the research, and their respective roles in the development and validation of the selection procedure. The researcher must use professional judgment in determining the appropriate information to provide and the communication format and style that will be most effective.

Researchers encourage organisations to consider the effects of participation in the validation effort on employees, departments, and business units. Typically, organisations decide that data from validation studies will be kept confidential and not used for subsequent employment-related decisions.

2.6.2 UNDERSTANDING WORK AND WORKER REQUIREMENTS

Historically, selection procedures were developed for specific jobs or job families. This remains the case in many situations. However, industries that experience rapid technological development or institute other strategies for accomplishing work may find that traditional jobs no longer exist. In such cases, considering important job requirements for a wider range or type of work activity may be more appropriate.

2.6.2.1 Strategies for Analysing the Work Domain and Defining Worker Requirements

The approach, method, and analyses used in a specific study of work is a function of the nature of the work itself, those who perform the work, and the organisational setting in which the work is accomplished. There is no single strategy that must be carried out, and multiple strategies may be appropriate.

There are situations where the importance or relevance of a construct is self-evident and does not require extensive work analysis. For example, absenteeism and turnover and their underlying constructs may be relevant to all work activities in an organisation. Therefore, demonstration of their relevance is not typically necessary.

2.6.2.2 Considerations in Specifying the Sampling Plan

The sampling plan for data collection should take into account the number of workers and their work locations, their characteristics (e.g., amount of experience, training, proficiency, etc.), shift or other work cycles, and other variables that might influence the work analysis.

2.6.2.3 Documentation of the Results

The methodology, data collection methods, analyses, results, and another implications for the validation effort should be documented. Frequently, this documentation will include a description of the major work activities, important worker requirements and their relationships to selection procedure content, and scoring when appropriate. The documentation should provide sufficient detail for another researcher to replicate the work analysis process. The documentation should also help the researcher understand the role of the work analysis as the foundation for any validation efforts.

2.6.3 SELECTING ASSESSMENT PROCEDURES FOR THE VALIDATION EFFORT

The researcher exercises professional judgment to determine those selection procedures that should be included in the validation effort. This judgment takes into consideration the organisational needs as well as the issues discussed in this section.

2.6.3.1 Review of Research Literature

Researchers should become familiar with research related to the organisation's objectives. The research literature is used to inform choices about selection procedures and the validation strategy to be employed.

2.6.3.2 Psychometric Considerations

When selecting one or more predictors, a number of psychometric characteristics of each instrument should be considered. Some of the more important psychometric considerations include reliability, evidence supporting the validity of the intended inferences, and differences among subgroups.

When choosing components of a selection battery, the researcher should consider the overall contribution of each component, its relative contribution, and potential construct redundancy and decide how much construct redundancy is desirable given the instruments and the situation.

2.6.3.3 Administration and Scoring Considerations

There are practical considerations regarding the consistent administration and scoring of a selection procedure. For example, the researcher must ensure that administration and scoring tasks can be completed consistently across all locations and administrators. To the extent that the selection procedure (e.g., work samples) requires subjective judgments in scoring, issues of rater training and inter-rater reliability become especially important. If standardised conditions are violated in the administration or scoring of a selection procedure, the generalisability of findings may be compromised. Regarding Internet administration, the ITC's *International Guidelines on Computer-Based and Internet Delivered Testing* makes reference to proctored and unproctored assessment. Practitioners should take cognisance of this document as it relates to the administration of tests. A South African study conducted by Holtzhausen (2004) compared a controlled Internet administration of a personality questionnaire with a supervised paper-and-pencil administration of the same. Comparable psychometric properties were found between the proctored and unproctored samples in terms of reliability and covariance structures, indicating that there was no distortion to the instrument itself.

2.6.3.4 Format and Medium

Format refers to the design of response requirements for selection procedure items (e.g., multiple-choice, essay). The choice of format may be influenced by the resources available to administer and score the selection procedure. For example, objectively scored items with established correct responses may be administered and scored in less time than selection procedures that require the individual to respond in more complex ways or that use individually judged responses.

Medium refers to the method of delivery of the selection procedure content. For example, a measure of cognitive ability could be presented via paper-and-pencil, computer, video, or orally.

There are advantages and disadvantages in selecting or adapting existing selection procedures from one medium to another. Computer-administered procedures may reduce the demands on administrators and enhance standardisation. Computer-administered tests also provide opportunities to measure constructs that do not lend themselves to testing by paper-and-pencil (e.g., use of spreadsheets and database management). Research has found that carefully developed computerised versions of cognitive ability power tests assess the same construct as the paper-and-pencil versions (Mead & Drasgow, 1993).

Changing the medium may also change the construct being measured. For example, converting a paper-and-pencil situational judgment test to a video where the situations will be acted out will reduce the reading component of the test. Also, adapting speeded tests of cognitive ability to a computerised version has been found to alter the construct being measured (Mead & Drasgow, 1993).

A number of considerations are important when evaluating the format and medium options. Cost and efficiency of operation may be the primary concern to the organisation. In addition, security, standardisation of testing conditions, candidate authentication, and accessibility of testing opportunities are all important considerations. Developers of selection systems should be cognizant that format and medium can affect mean score differences among subgroups (Hough, Oswald, & Ployhart, 2001). When possible, the structural equivalence of different media, for example Internet, computer and paper-and-pencil testing should be investigated (See the ITC's *International Guidelines on Computer-Based and Internet Delivered Testing* for pointers on adapting tests for computer-based or Internet delivery).

2.6.3.5 Acceptability to the Candidate

In addition to the organisation's needs and objectives, researchers also need to consider the acceptability of the selection procedure to candidates. A number of factors influence candidates' reactions to a selection procedure, including individual characteristics (e.g., work experiences, demographics, and cultural backgrounds), the role of the individual (e.g., applicant, incumbent, manager), the extent to which the content of the selection procedure resembles the work, the individual's capability with respect to the constructs measured, and the perceived passing or selection rate. Generally, the greater the similarity between the selection procedure and the work performed, the greater the acceptability to candidates, management, and other stakeholders. However, selection procedures that closely resemble the work may be perceived as obsolete when the work changes.

Some selection procedures may appear less face valid than other procedures. For example, the value of information collected on biodata forms and personality inventories in predicting job performance may not be obvious to many. Communications regarding the selection procedure, the constructs measured, and the role of incumbents and managers in developing the procedure may improve understanding and acceptance of a selection procedure.

There are situations where some candidates refuse to participate in certain types of selection procedures. It may be useful to consider whether desirable candidates remove themselves from consideration because of the selection procedure. In addition, recruiters sometimes resist or attempt to

circumvent the use of selection procedures because it increases the need for additional candidates. Therefore, researchers should consider approaches designed to minimise any negative perception of a selection procedure.

2.6.3.6 Alternate Forms

Alternate forms of a selection procedure may be needed to reduce practice effects and enhance security. Alternate forms may allow the organisation to continue assessment after a security breach; however, researchers may provide information to organisations to help them balance these advantages with the increased costs for development and validation of alternate forms. If alternate forms (including adaptive tests) are developed, care must be taken to ensure that candidate scores are comparable across forms. If alternate forms are used, establishing the equivalence of scores on the different forms is usually necessary. The statistical procedures used in equating studies typically take into account the size and relevant characteristics of the samples, the use of an anchor test or linking test items, and the feasibility of determining equating functions within subgroups.

2.6.4 SELECTING THE VALIDATION STRATEGY

Once researchers have worked with the organisation to define its objectives for developing a selection procedure, understand the requirements of the work, and reach agreement on the type of selection procedure, researchers must decide what validation strategy or strategies will be pursued to accumulate evidence to support the intended inference. In addition, the strategy selected must be feasible in the organisational context and meet the project goals within the constraints imposed by the situation.

2.6.4.1 Fit to Objectives, Constraints, and Selection Procedures

In choosing an initial validation strategy, the researcher should consider the fit of the strategy to the organisation's objectives and constraints, as well as its fit to the selection procedures planned and the criterion measures. Three examples are provided below to describe possible ways in which validation strategies may be matched to organisational objectives and constraints. In the first scenario, an organisation wanting to assemble validity evidence for a small population position may rely upon a validity generalisation strategy because extensive cumulative evidence exists for the predictor-criterion relationship in similar situations. In contrast, another organisation facing a similar problem that wants to extend a selection procedure from one business unit to another may use a transportability study to establish the validity of the employee selection procedure in another business unit with the same job**.** Neither option may be available when a position is unique to the organisation. Thus, in the third situation, the organisation may rely on a content-based validity strategy.

2.6.4.2 Individual Assessments

Individual assessment refers to one-on-one evaluations on the basis of a wide range of cognitive and noncognitive measures that are integrated by the assessor, often resulting in a recommendation rather than a selection decision or prediction of a specific level of job performance (Jeanneret & Silzer, 1998). The assessor should have a rationale for the determination and use of the selection procedures. In such instances, the validity of the assessor's clinical judgments is most important to the evaluation of the assessment process. If there are multiple assessors, the consistency of their assessment findings can be valuable to understanding validity and making accurate judgments about the relevant KSAOs. Validation research studies of clinical judgments are clearly an exception rather than the rule (Ryan & Sackett, 1998). However, both validity generalisation and content-oriented validation strategies may be appropriate. For example, there may be a wide range of generalisable evidence that has been accumulated by a test publisher or the assessing psychologist demonstrating that a personality scale (e.g., conscientiousness) is predictive of successful managerial performance. Therefore, such a selection procedure would be appropriate for use in an executive assessment protocol. An example of a content-oriented validation approach would be demonstrating the relationship of an in-basket

selection procedure that measures planning capability to the planning requirements of an executive position.

2.6.5 SELECTING CRITERION MEASURES

When the source of validity evidence is based on the relationships between predictor scores and other variables (criteria), the nature of the criteria is determined by the proposed uses of the selection procedures and outcomes from the analysis of work and worker requirements. Professional judgment should be exercised in selecting the most appropriate criteria given known organisational constraints and climate.

2.6.5.1 Performance-Oriented Criteria

Criteria that are representative of work activities, behaviours, or outcomes usually focus on the job performance of incumbents. Supervisory ratings are the most frequently used criteria, and often they are designed specifically for use in the research study. Other performance information also may be useful (e.g., training program scores, sales, error rates, and productivity indices). Consideration should be given to psychometric factors for all criteria whenever feasible.

2.6.5.2 Other Indices

Depending on the objective of the validation effort, indices other than those directly related to task performance may be most appropriate. Examples include absenteeism, turnover, and other organisational citizenship behaviours (e.g. Rotundo & Sackett, 2002). Again, the researcher should be cautious about deficiencies or contaminating factors in such indices.

2.6.5.3 Relevance and Psychometric Considerations

Criteria are typically expected to represent some construct (often work performance), and the quality of that representation should be established. For example, the fidelity of a work sample used as a criterion should be documented on the basis of the work analysis. Supervisory ratings should be defined and scaled in terms of relevant work activities or situations. All criteria should be representative of important work behaviours, outcomes, or relevant organisational expectations regarding individual employee behaviour or team performance.

Criteria should be reliable, and the determination of that reliability may be influenced by the study parameters and organisational constraints. For example, while it may be desirable to have two raters independently evaluate the performance of an employee to determine the inter-rater reliability of the ratings, the work situation and supervisory relationships may preclude such an effort (e.g., there may not be two supervisors knowledgeable about an employee's work). In any circumstance, the researcher should determine what reliability estimates are calculated, how they are obtained, and what levels of reliability are acceptable. Please take cognisance of the critical value of the criterion (refer back to page 14 for a discussion on criterion development).

2.6.6 DATA COLLECTION

Collection of both predictor and criterion data in a validation study requires careful planning and organising to ensure complete and accurate data. The standardised conditions under which the validation research is conducted are normally replicated to the extent possible during actual use of the selection procedure. In order to collect accurate and complete information, the test user should consider the following activities.

2.6.6.1 Communications

Relevant information about the data collection effort should be communicated to all those affected by the effort including management, those who take the test for research purposes, persons who provide criterion data, and those who will use the test. Appropriate communications will facilitate the data collection and encourage all involved to provide accurate and complete information. The kind of information shared depends on the needs of the organisation and the individuals involved. For example, participants in the validation research will want to know how their test results will be used and who will have access to the results. Supervisors who provide criterion ratings and others who provide archival criterion data will want to know the logistics of data collection, ultimate use, and provisions for confidentiality.

End users, such as the staffing organisation or the client organisation employing individuals who were screened with the selection procedures, should have an overview of the study. When feasible, anticipated uses of job analysis, test, and criterion data should be shared with those who generated it.

2.6.6.2 Pilot Testing

The researcher should determine the extent to which pilot testing is necessary or useful. Previous experience with specific selection procedures may reduce or eliminate this need. Availability of test takers and opportunities to conduct pilot testing may be influenced by various organisational constraints.

2.6.6.3 Match Between Data Collection and Implementation Expectations

Selection procedures should be administered in the same way that they will be administered in actual use. For example, if interviewers are provided face-to-face training in the validation study, similar training should be provided in actual use. Instructions and answers to candidate questions should be as similar as possible during validation and implementation.

2.6.6.4 Confidentiality

Confidentiality is an ethical responsibility of the researcher. It is also a major concern to all those involved in the research. Those who provide information, performance ratings, or content validity linkages may be more willing to provide accurate information if they are assured of the confidentiality of their individual contributions. Participants in validation research studies should be given confidentiality unless there are persuasive reasons to proceed otherwise.

The researcher should carefully decide what level of anonymity or confidentiality can be established and maintain it throughout the study. The researcher provides the maximum confidentiality feasible in the collection and storage of data, recognising that identifying information of some type is often required to link data collected at different times or by different individuals. Web-based data collection presents additional confidentiality challenges. See the HPCSA's Ethical Code for Professional Conduct regarding confidentiality.

2.6.6.5 Quality Control and Security

The test user should employ data collection techniques that are designed to enhance the accuracy and security of the data and test materials. Public disclosure of the content and scoring of most selection procedures should be recognised as a potentially serious threat to their reliability, validity, and subsequent use. All data should be retained at a level of security that permits access only for those with a need to know.

2.6.7 DATA ANALYSES

A wide variety of data may be collected and analysed throughout the course of a validation study. The responsibilities and supervision of the people who conduct data analyses should be commensurate with their capabilities and relevant experience.

2.6.7.1 Data Accuracy

All data should be checked for accuracy. Checks for data accuracy typically include verification that scores are within the possible ranges and that no apparent falsification of responses occurred.

2.6.7.2 Missing Data/Outliers

Often, one or more data points are missing, and/or outliers exist in the data set. Because these circumstances are typically unique to the validation effort underway, establishing hard-and-fast rules is not possible. Instead, the researcher must examine each situation on its own merits and follow a strategy based on professional judgment. Researchers should document the rationale for treating missing data and/or outliers so their work can be replicated. If imputation techniques are used to estimate missing data, such techniques and their underlying assumptions should be documented clearly.

2.6.7.3 Descriptive Statistics

Most data analyses will begin with descriptive statistics that present analyses of frequencies, central tendencies, and variances. Such descriptions should be provided for the total group and for relevant subgroups if they are large enough to yield reasonably reliable statistics.

2.6.7.4 Appropriate Analyses

Data analyses should be appropriate for the method or strategy undertaken. Data are frequently collected as part of the analysis of work and during the validation effort itself. Data analytic methods used also should be appropriate for the nature of the data (e.g., nominal, ordinal, interval, ratio), sample sizes, and other considerations that will lead to correct inferences regarding the data sets.

2.6.7.5 Differential Prediction

Organisations vary in their goals, and competing interests within the organisation are not unusual. Efforts to reduce differences for one subgroup may increase differences for another. Given the difficulty in reconciling different interests in the case of substantial over- or underprediction, researchers oftentimes consider the effects of the prediction errors and their relationship to organisational goals.

A finding of predictive bias does not necessarily prevent the operational use of a selection procedure. For example, if the study is based upon an extremely large sample, a finding of a small but statistically significant differential prediction may have little practical effect. In general, the finding of concern would be evidence of substantial underprediction of performance in the subgroup of interest. Such a finding would generally preclude operational use of the predictor and would likely lead to additional research and considerations of modifying or replacing the selection procedure.

Absent a finding of substantial underprediction, a reasonable course of action for some organisations would be to recommend uniform operational use of the predictor for all groups. However, a large amount of overprediction may also lead to a consideration of alternate selection procedures.

2.6.7.6 Combining Selection Procedures Into an Assessment Battery

The researcher must exercise professional judgment regarding the outcomes of the validation effort to determine those predictors that should be included in the final selection procedure and the method of combination (including predictor weighting) that will meet the goals of the organisation. The algorithm for combining the selection procedures and the rationale for the algorithm should be described. When combining predictors, the validity of the inferences resulting from the composite is of primary importance.

2.6.7.7 Multiple Hurdles Versus Compensatory Models

Taking into account the purpose of the assessment and the outcomes of the validity study, the researcher must decide whether candidates are required to score above a specific level on each of several assessments (multiple hurdles) or achieve a specific total score across all assessments (compensatory model). There are no absolutes regarding which model should be implemented, and at times a hurdle may be most appropriate for one predictor, while a compensatory model may be best for other predictors within the overall selection procedure. The rationale and supporting evidence should be presented for the model recommended for assessment scoring and interpretation. Researchers should be aware that the method of combining test scores might affect the overall reliability of the entire selection process and the subgroup passing rates (Sackett & Roth, 1996).

2.6.7.8 Cutoff Scores Versus Rank Orders

Two frequently implemented selection decision strategies are selection based on a cutoff score or selection of candidates in a top-down order. There is no single method for establishing cutoff scores. If based on valid predictors demonstrating linearity or monotonicity throughout the range of prediction, cutoff scores may be set as high or as low as needed to meet the requirements of the organisation. Additionally, given monotonicity, selecting the top scorers in top-down order maximises estimated performance on the criterion measure if there is an appropriate amount of variance in the predictor. Where there is an indication of nonmonotonicity, this finding should be taken into consideration in determining how to use selection procedure scores.

Given the unitary concept of validity and the underlying premise (based on empirical evidence) that inferences regarding predictors of a cognitive nature and performance criteria are linear (Coward & Sackett, 1990), cognitively based selection techniques developed by content-oriented procedures and differentiating adequately within the range of interest can usually be assumed to have a linear relationship to job behaviour. Such content-oriented procedures support rank ordering and setting the cutoff score as high or as low as necessary. Research has not yet established whether this same set of premises holds true for other types of predictors (e.g., personality inventories, interest inventories, indices of values).

Professional judgment is necessary in setting any cutoff score and typically is based on a rationale that may include such factors as estimated cost-benefit ratio, number of vacancies and selection ratio, expectancy of success versus failure, the consequences of failure on the job, performance and diversity goals of the organisation, or judgments as to the knowledge, skill, ability, or other characteristics required by the work. When cutoff scores are used as a basis for rejecting applicants, the researcher should document their rationale or justification.

Cutoff scores are different from critical scores in that a cutoff score defines a point on a selection procedure score distribution below which candidates are rejected, while a critical score defines a specified point on a distribution of selection procedure scores above which candidates are considered successful. Critical scores are criterion referenced and may be useful for implementing a certain type of selection procedure (e.g., certification exam), but are not appropriate when no absolute minimum on the selection procedure score distribution can be discerned (e.g., cognitive ability or aptitude test).

When researchers make recommendations concerning the use of a rank ordering method or a cutoff score, the recommendation often takes into account the labour market, the consequences of errors in prediction, the level of a KSAO represented by a chosen cutoff score, the utility of the selection procedure, resources needed to monitor and maintain a list of qualified candidates, and other relevant factors. The goals of the organisation may favour a particular alternative. For example, some organisations decide to use a cutoff score rather than rank ordering to increase workforce diversity, recognising that a reduction also may occur in job performance and utility. Whatever the decision, the researcher should document the rationale for it.

In South Africa, when companies set recruitment targets to increase workforce diversity, top-down selection within groups is a viable alternative to increase diversity and maintain high utility. In terms of labour legislation in South Africa, utilising within-group top-down selection is legal. This does, however, become a policy issue that needs to be determined before the implementation of selection decisions.

2.6.7.9 Bands

Bands are ranges of selection procedure scores in which candidates are treated alike. The implementation of a banding procedure makes use of cutoff scores, and there are a variety of methods for determining bands (Cascio, Outtz, Zedeck, & Goldstein, 1991; Campion et al., 2001).

Bands may be created for a variety of administrative or organisational purposes; they also may be formed to take into account the imprecision of selection procedure scores and their inferences. However, because bands group candidates who have different selection procedure scores, predictions of expected criterion outcomes are less precise. Thus, banding will necessarily yield lower expected criterion outcomes and selection utility (with regard to the criterion outcomes predicted by the selection procedure) than will top-down, rank order selection. On the other hand, the lowered expected criterion outcomes and selection utility may be balanced by benefits such as administrative ease and the possibility of increased workforce diversity, depending on how within-band selection decisions are made. If a banding procedure is implemented, the basis for its development and the decision rules to be followed in its administration should be clearly documented.

2.6.7.10 Norms

Normative information relevant to the applicant pool and the incumbent population should be presented when appropriate. The normative group should be described in terms of its relevant demographic and occupational characteristics and presented for subgroups with adequate sample sizes. The time frame in which the normative results were established should be stated.

The choice of norm group used in evaluating candidates' results plays a key role in the assessment process, and should be both as representative of the applicant group and incumbent population as possible, and appropriate for the position the assessment was conducted for. In practical terms, when determining whether a norm group is representative of the applicant group and incumbent population, similarities are considered in aspects such as ethnicity, gender, age and educational background.

2.6.7.11 Communicating the Effectiveness of Selection Procedures

Two potentially useful methods for communicating the effectiveness of selection procedures are expectancy analyses and utility estimates.

Expectancies and practical value. Expectancy charts may assist in understanding the relationship between a selection procedure score and work performance. Further, information in the Taylor-Russell Tables (Taylor & Russell, 1939) identifies what proportions of hired candidates will be successful under different combinations of test validity (expressed as correlation coefficients), selection ratios, and percentages of current employees that are satisfactory performers.

Utility. Projected productivity gains or utility estimates for each employee and the organisation due to use of the selection procedure may be relevant in assessing its practical value. Utility estimates also may be used to compare the relative value of alternative selection procedures. The literature regarding the impact of selection tests on employee productivity has provided several means to estimate utility (Brogden, 1949; Cascio, 2000**;** Cronbach & Gleser, 1965; Hunter, Schmidt, & Judiesch, 1990; Naylor & Shine, 1965; Raju, Burke, & Normand, 1990**;** Schmidt, Hunter, McKenzie, & Muldrow, 1979). Some of these express utility in terms of reductions in some outcome of interest (e.g., reduction in accidents, reduction in person-hours needed to accomplish a body of work). Others express utility in Rand terms, with the Rand value obtained via a regression equation incorporating a number of parameters, such as the increment in validity over current practices and the Rand value of a standard deviation of performance. Still others express utility in terms of percentage increases in output due to improved selection. The values for terms in these models are often estimated with some uncertainty, and thus the result is a projection of gains to be realised if all of the model assumptions hold true. Often researchers do not conduct follow-up studies to determine whether projected gains are, in fact, realised. Under such circumstances, the results of utility analyses should be identified as estimates based on a set of assumptions, and minimal and maximal point estimates of utility should be presented when appropriate to reflect the uncertainty in estimating various parameters in the utility model. Practitioners are recommended to embark on criterion-referenced approaches.

2.6.8 APPROPRIATE USE OF SELECTION PROCEDURES

Inferences from selection procedure scores are validated for use in a prescribed manner for specific purposes. To the extent that a use deviates from either the prescribed procedures or the intended purpose, the inference of validity for the selection procedure is likely to be affected.

2.6.8.1 Combining Selection Procedures

Personnel decisions are often made on the basis of information from a combination of selection procedures. The individual components as well as the combination should be based upon evidence of validity. Changes in the components or the mix of components typically require the accumulation of additional evidence to support the validity of inferences for the altered procedure. When a multiple hurdle approach is employed, the original validation data may be relied on for those components that remain intact. However, the effectiveness of the selection procedure as a whole may be reduced as a result of the introduction of a predictor of unknown quality.

When a compensatory approach is used, the addition or deletion of a selection procedure component can fundamentally change the inferences that may be supported. Under these circumstances, the original validation evidence may not be sufficient when there are alterations to the selection procedures that are not supported by a follow-up validation effort.

2.6.8.2 Using Selection Procedures for Other Purposes

The selection procedure should be used only for the purposes for which there is validity evidence. For example, diagnostic use of a selection procedure that has not been validated in a way to yield such information should be avoided. Likewise, the use of a selection procedure designed for an educational environment cannot be justified for the purpose of predicting success in employment settings unless the education tasks and the work performed in the validation research or their underlying requirements are closely related, or unless the relevant research literature supports this generalisation.

2.6.9 RECOMMENDATIONS

The recommendations based on the results of a validation effort should be consistent with the objectives of the research, the data analyses performed, and the researcher's professional judgment and ethical responsibilities. The recommended use should be consistent with the procedures used in, and the outcomes from, the validation research including the validity evidence for each selection

procedure or composite score and the integration of information from multiple sources. In addition, the researcher typically considers the cost, labour market, and performance expectations of the organisation, particularly when choosing a strategy to determine who is selected by the procedure. Tight labour markets may necessitate acceptance of a lower score on the selection procedure. Also, the organisation's expectations regarding work force diversity may influence the use of test information for that organisation.

2.6.10 TECHNICAL VALIDATION REPORT

Reports of validation efforts should include enough detail to enable a researcher competent in personnel selection to know what was done, to draw independent conclusions in evaluating the research, and to replicate the study. The reports must accurately portray the findings, as well as the interpretations of and decisions based on the results. Research findings that may qualify the conclusions or the generalisability of results should be reported. The following information should be included:

2.6.10.1 Identifying Information

The report should include the author(s), dates of the study, and other information that would permit another researcher to understand who conducted the original research.

2.6.10.2 Statement of Purpose

The purpose of the validation research should be stated in the report.

2.6.10.3 Analysis of Work

The report should contain a description of the analysis of work, including procedures employed, the participants in the process, data analyses, and results.

2.6.10.4 Search for Alternative Selection Procedures

The researcher should document any search for selection procedures (including alternate combinations of these procedures) that are substantially equally valid and reduce subgroup differences.

2.6.10.5 Selection Procedures

Names, editions, and forms of selection procedures purchased from publishers should be provided as well as descriptions and, if appropriate, sample items. When proprietary tests are developed, the researcher should include a description of the items, the construct(s) that are measured, and sample items, if appropriate. Typically, copies of tests or scoring procedures should not be included in technical reports or administration manuals in order to protect the confidentiality of operational items. The rationale for the use of each procedure and basic descriptive statistics, including appropriate reliability estimates for the sample in the research study when feasible, also should be included. If raters are an integral part of the selection procedure (as in some work samples), the reliability of their ratings should be determined and documented.

2.6.10.6 Relationship to Work Requirements

The report should provide a description of the methods used by the researcher to determine that the selection procedure is significantly related to a criterion measure or representative of a job content domain. Establishing the relationship of a selection procedure to job content is particularly important when conducting a job content validation study.

2.6.10.7 Criterion Measures (When Applicable)

A description of the criterion measures, the rationale for their use, the data collection procedures, and a discussion of their relevance, reliability, possible deficiency, freedom from contamination, and freedom from or control of bias should be provided in detail.

2.6.10.8 Research Sample

The sampling procedure and the characteristics of the research sample relative to the interpretation of the results should be described. The description should include a definition of the population that the sample is designed to represent, sampling biases that may detract from the representativeness of the sample, and the significance of any deviations from representativeness for the interpretation of the results. Data regarding restriction in the range of scores on predictors or criterion measures are especially important. When a transportability study is conducted to support the use of a particular selection procedure, the relationship between the original validation research sample and the population for which the use of the selection procedure is proposed should be included in the technical report.

2.6.10.9 Results

All summary statistics that relate to the conclusions drawn by the researcher and the recommendations for use should be included. Tables should present complete data, not just significant or positive results. The sample size, means, standard deviations, and intercorrelations of variables measured and other information useful to the interpretation of the results should be presented and clearly labelled. Both uncorrected and corrected values should be presented when corrections are made for statistical artifacts such as restriction of range or unreliability of the criterion.

2.6.10.10 Scoring and Transformation of Raw Scores

Methods used to score items and tasks should be fully described. When performance tasks, work samples, or other methods requiring some element of judgement are used, a description of the type of rater training conducted and scoring criteria should be provided.

Derived scales used for reporting scores and their rationale should be described in detail in the research report or administration manual. Whether using derived scores or locally produced labels (such as "qualified", "marginal", or "unqualified"), the researcher should clearly describe the logical and psychometric foundations.

2.6.10.11 Normative Information

Parameters for normative data provide researchers and users with information that guides relevant interpretations. Such parameters often include demographic and occupational characteristics of the normative sample, time frame of the data collection, and status of test takers (e.g., applicants, incumbents, college students).

When normative information is presented, it should include measures of central tendency and variability and should clearly describe the normative data (e.g., percentiles, standard scores). Norm tables usually report the percent passing at specific scores and may be useful in determining the effects of a cutoff score. Expectancy tables indicate the proportion of a specific sample of candidates who reach a specified level of success and are often used to guide implementation decisions.

2.6.10.12 Recommendations

The recommendations for implementation and the rationale supporting them (e.g., the use of rank ordering, score bands, or cutoff scores, and the means of combining information in making personnel

decisions) should be provided. Because implementation rules like those applied to cutoff scores sometimes do change, subsequent modifications should be documented and placed in an addendum to the research report or administration manual.

2.6.10.13 Caution Regarding Interpretations

Research reports and/or administration manuals should help readers make appropriate interpretations of data and should warn them against common misuses of information.

2.6.10.14 References

There should be complete references for all published literature and available technical reports cited in the report. Technical reports completed for private organisations are often considered proprietary and confidential, and the researcher cannot violate the limitations imposed by the organisation. Consequently, some technical reports that may have been used by the researcher are not generally available.

2.6.11 ADMINISTRATION GUIDE

The term "test administrator" refers to those individuals responsible for day-to-day activities such as scheduling testing sessions, administering the selection procedure, scoring the procedure, and maintaining the databases.

An administration guide should document completely the information needed to administer the selection procedure, score it, and interpret the score. When the selection procedure is computer-based or in a format other than paper-and-pencil, the administration guide should also include detailed instructions on the special conditions of administration. While this document is sometimes a part of a technical report, it is often separate so that confidential information in the validity study is protected and administrators are provided with only the information needed to administer the selection procedure. In other situations, the test user will write parts of the administration guide since the researcher may not know the organisation's specific policies or the details of its implementation strategies. In deciding whether two separate documents are needed, the researcher should consider access to each document, the sensitivity of information to be included, the purpose of each document, and the intended audience.

Administration guides developed by a publisher are often supplemented with addenda that cover local decisions made by the user organisation. Consequently, not all the information listed below will be found in every administration guide from a publisher. However, the researcher in the user organisation should try to provide answers or guidance for the issues raised.

The information developed for users or examinees should be accurate and complete for its purposes and should not be misleading. Communications regarding selection procedures should be stated as clearly and accurately as possible so that readers know how to carry out their responsibilities competently. The writing style of all informational material should be written to meet the needs of the likely audience. Normally, the following information should be included in an administration guide.

2.6.11.1 Introduction and Overview

This section of the report should inform the reader of the purpose of the assessment procedure and provide an overview of the research that supports the use of the procedure. The introduction should explain why the organisation uses formal, validated selection procedures, the benefits of professionally developed selection procedures, the importance of security, and the organisation's expectations regarding the consistency of their use. Care must be taken in preparing such documents to avoid giving the reader an impression that an assessment program is more useful than is really the case.

2.6.11.2 Contact Information

The administration guide should provide information about whom to contact if there are questions or unanticipated problems associated with the selection procedure.

2.6.11.3 Selection Procedures

The selection procedure should be thoroughly described. Names, editions, and forms of published procedures as well as information for ordering materials and ensuring their security should be provided. Although entire tests are not usually included in administration guides for security reasons, sample items are helpful. When proprietary tests are developed, the researcher should include a description of the items, the construct(s) that are measured, and sample items.

2.6.11.4 Applicability

The description of the selection procedure should indicate to whom the procedure is applicable (e.g., candidates for "x" job) and state any exceptions (e.g., exemptions for job incumbents) to test requirements. It may also be useful to indicate that use of a selection procedure is based on one or more validation efforts that focused on specific jobs/job requirements and that these efforts define the boundaries of any test applications. If the organisation has rules about when tests are administered, these rules must be clearly stated in the administration guide used by the organisation. For example, some organisations only administer a selection procedure when there is a job vacancy. Other organisations may administer selection procedures periodically in order to build pools of qualified candidates.

2.6.11.5 Administrators

The administration guide should state the necessary qualifications of administrators and the training required to administer selection procedures in general, as well as the specific selection procedure described in the administration guide. Administrators should receive training in the administration of selection procedures. Administrators must understand that failure to follow standardised protocol may render the research results and operational scores irrelevant to some degree. The researcher must be both insistent and persuasive to gain understanding with regard to both the nature of and the need for standardised administration of tests or other procedures. Periodic retraining may be needed to reinforce the administration rules. Observational checks or other quality control mechanisms should be built into the test administration system to ensure accurate and consistent administration.

2.6.11.6 Information Provided to Candidates

Many organisations use test brochures or test orientation materials to inform candidates about the employee selection process. Some organisations also provide informational sessions prior to the administration of a selection procedure. When appropriate, the researcher should consider providing candidates with clearly written, uniform information about the selection procedure such as the purpose, administrative procedures, completion strategies, time management, feedback, confidentiality, process for requesting accommodation for disability, and other relevant user policies. Whenever possible, both the content and the process for orienting candidates should be standardised. The administration guide should describe these materials and provide information on how the administrator may obtain them. The rules for distribution should be explicitly stated in order to facilitate consistent treatment of candidates.

Some practitioners recommend that informed consent be obtained from an individual undergoing assessment. In the work context, however, the HPCSA's Ethical Code of Professional Conduct notes that written informed consent is not necessary when consent is implied, as testing is conducted as a routine educational, institutional or organisational activity (as in job interview testing).

2.6.11.7 Guidelines for Administration of Selection Procedures

The researcher should use the administration guide as an opportunity to convey the organisation's requirements for selection procedure administration. In addition to detailed instructions regarding the actual administration of the selection procedure, the administration guide often includes rules and tips for providing an appropriate testing environment as well as ensuring the candidate's identity. Deviating from the procedures and instructions for administration detailed in the instrument manual may result in unstandardised administration for some or all candidates sitting the assessments, which could lead to possible unfair and unlawful discrimination.

For security reasons, the identity of all candidates should be confirmed prior to administration. Administrators should monitor the administration to control possible disruptions, protect the security of the test materials, and prevent collaborative efforts by candidates. The security provisions, like other aspects of the *Guidelines*, apply equally to computer and Internet-administered sessions. (See the ITC's *International Guidelines on Computer-Based and Internet Delivered Testing* regarding confidentiality and verification of candidate identity.)

2.6.11.8 Administration Environment

There are a number of factors that potentially affect test administration: appropriate workspace, adequate lighting, and a quiet, comfortable setting, free of distractions. The researcher should consider these conditions and their potential effects on test performance. At a minimum, selection procedure administration should be in an environment that is responsive to candidates' concerns about the selection procedures and maintains their dignity.

In a country as diverse as South Africa the choice of language for administration is an important consideration, as well as the impact it may have on test performance if a decision is made to test some people in their first language and others in their second. For example, a test assessing English reading skills may be appropriate for those individuals whose first language is English, but could present problems for people who are not as comfortable with English. The key determination of whether use of the test is fair in this situation would be how it relates to the requirements of the position. For a position where the ability to read and understand material written in English is critical and an inherent requirement of the job, the use of such a test would be considered fair, justifiable and legally defensible. (See the *Standards* for additional information regarding testing individuals of diverse linguistic backgrounds.)

As far as possible, the conditions of the administration should be standard for all candidates, which includes the language the instructions are given in. When conducting test administration with candidates whose first language is not the language of administration, additional care should be taken to ensure that the candidates understand the instructions properly.

2.6.11.9 Scoring Instructions and Interpretation Guidelines

The researcher should provide the selection procedure administrator or user with details on how the selection procedure is to be scored and how results should be interpreted. The administration guide should help readers make appropriate interpretations of data and warn them against common misuses of information.

Processes that will ensure accuracy in scoring, checking, and recording results should be used. This principle applies to the researcher and to any agent to whom this responsibility has been delegated. The responsibility cannot be abrogated by purchasing services from an outside scoring service. Quality control checks should be implemented to ensure accurate scoring and recording.

Instructions for scoring by the user should be presented in the administration guide in detail to reduce clerical errors in scoring and to increase the reliability of any judgments required. Distinctions among

constructs should be described to support the accuracy of scoring judgments. Scoring keys should not be included in technical reports or administration manuals and should be made available only to persons who score or scale responses.

If Computer-Based Test Interpretation (CBTI) is used to process responses to a selection procedure, the researcher should provide detailed instructions on how the CBTI is to be used in decision-making. See the test review guidelines of the European Federation of Psychologists Association for more information on quality standards for CBTI.

2.6.11.10 Test Score Databases

Organisations should decide what records of assessment administrations and scores are to be maintained and should provide detailed information (or reference detailed information) regarding record keeping and databases. In addition, policies on the retention of records (e.g., duration, security, accessibility, etc.) and the use of archival data over time should be established and communicated. Raw scores should be kept because data reported in derived scales may limit further research. Databases should be maintained for sufficient time periods to support periodic audits of the selection process.

2.6.11.11 Reporting and Using Selection Procedure Scores

The researcher must communicate how selection procedure scores are to be reported and used. Results should be reported in language likely to be interpreted correctly by persons who receive them. The administration guide should also indicate who has access to selection procedure scores.

Administrators should be cautioned about using selection procedure information for uses other than those intended. For example, although selection procedure data may have some validity in determining later retention decisions, more potentially relevant measures such as performance ratings may be available. Furthermore, if the pattern of selection procedure scores is used to make differential job assignments, evidence is required demonstrating that the scores are linked to, or predictive of, different performance levels in the various assignments of job groupings.

2.6.11.12 Candidate Feedback

In addition to reporting selection procedure scores to others within the organisation, the researcher should include information on how to provide feedback to candidates, if such feedback is feasible or appropriate. Feedback should be provided in clear language that is understandable by candidates receiving the feedback, and should not violate the security of the test or its scoring.

2.6.11.13 Non-standard Administrations (See Also Candidates With Disabilities)

The administration guide should cover non-standard selection procedure administrations. Such administrations encompass not only accommodated selection procedure sessions, but also sessions that were disrupted (e.g., power failures, local emergency, and illness of a candidate), involved errors (e.g., questions and answer sheet did not match, timing mistake), or were non-standard in some other way.

The administration guide should establish a clear process to document and explain any modifications of selection procedures, disruptions in administration, or any other deviation from established procedures in the administration, scoring, or handling of scores. While it is impossible to predict all possible occurrences, the researcher should communicate general principles for how deviations from normal procedures are to be handled.

2.6.11.14 Reassessing Candidates

Generally, employers should provide opportunities for reassessment and reconsidering candidates whenever technically and administratively feasible. In some situations, as in one-time examinations, reassessment may not be a viable option. In order to facilitate consistency of treatment, the administration guide should clearly explain whether candidates may be reassessed and how reassessment will take place. In some organisations, specific time intervals must elapse. In others, although difficult to evaluate, significant developmental activities must have occurred prior to reassessment.

2.6.11.15 Corrective Reassessment

Users in conjunction with researchers should consider when corrective reassessment is appropriate. Critical errors on the part of the administrator (e.g., timing mistakes, use of nonmatching selection procedure booklet and answer sheet) and extraordinary disturbances (e.g., fire alarm, acutely ill assessee) are usually justifications for reassessment. The administration guide should cover procedures and guidelines for granting corrective reassessment and documenting all requests.

2.6.11.16 Security of the Selection Procedure

Selection procedure items that are widely known (through study, coaching, or other means) in an organisation are usually less effective in differentiating among candidates on relevant constructs. Maintenance of test security therefore limits the type and amount of feedback provided to candidates. The more detail on candidate responses provided, the greater the security risk. The administration guide should emphasise the importance of safeguarding the content, scoring, and validity of the selection procedure.

Selection procedures usually represent a significant investment on the part of the organisation for development and validation. The administration guide should point out the value of the selection procedure itself and the cost of compromised selection procedures in terms of the additional research required and the possibility of a less capable candidate being hired.

Practices for ensuring the security of selection procedure documents (e.g., numbering test booklets and maintaining records of the numbers; keeping used and unused selection procedures in a secure, locked facility; collecting scratch paper after administration sessions) and selection procedure scoring should be communicated.

Selection procedure scores must be kept secure and should be released only to those who have a need to know and who are qualified to interpret them. Special practices may be required to protect confidential materials and selection procedure information that exist in electronic forms. Although security practices may be difficult to apply in the case of employment interviews, the importance of security as a means of preserving their standardisation and validity should be considered. Organisations are encouraged to develop policies that specify the length of time that confidential information is retained. When confidential information is destroyed, the user should consider ways of maintaining its security such as having selection personnel supervise the destruction of the documents.

2.6.11.17 References

When other useful documents are mentioned, they should be referenced fully. When the documents are internal publications, the means of acquiring those documents should be described.

2.6.12 OTHER CIRCUMSTANCES REGARDING THE VALIDATION EFFORT AND USE OF SELECTION PROCEDURES

2.6.12.1 Influence of Changes in Organisational Demands

Because organisations and their work forces are dynamic in nature, changes in organisational functioning may occur and subsequent selection procedure modifications may be necessary. Changing work requirements may lead to adjustments in cutoff scores or the introduction of a new assessment, both of which would require further study of the existing selection procedure. If advised of such circumstances, the researcher should examine each situation on its own merits and make recommendations regarding the impact of organisational change on the validation and use of any selection procedure.

2.6.12.2 Review of Validation and Need for Updating the Validation Effort

Researchers should develop strategies to anticipate that the validity of inferences for a selection procedure used in a particular situation may change over time. Such changes may occur because of changes in the work itself, worker requirements, or work setting. Users (either on their own or with researcher assistance) of a selection procedure should periodically review the operational use of the assessment instrument and the data at hand (including timeliness of normative data if appropriate) to determine if additional research is needed to support the continued use of the selection procedure. When needed, the research should be brought up to date and reported. The technical or administration guides should be revised (or an addendum added) if changes in research data or use of procedures make any statement or instruction incorrect or misleading.

2.6.13 CANDIDATES WITH DISABILITIES

Assessing candidates with disabilities may require special accommodations that deviate from standardised procedures. Accommodations are made to minimise the impact of a known disability that is not relevant to the construct being assessed. For example, an individual's upper extremity motor impairment may affect a score on a measure of cognitive ability although the motor impairment is not related to the individual's cognitive ability. Accommodations may include, but are not limited to, modifications to the environment (e.g., high desks), medium (e.g., Braille, reader), time limit, or content. Combinations of accommodations may be required to make valid inferences regarding the candidate's ability on the construct(s) of interest. For more information on accommodations and their implications, consult the International Guidelines for Test Use of the International Test Commission.

Professional judgment is required on the part of the user and the developer regarding the type or types of accommodations that have the least negative impact on the validity of the inferences made from the selection procedure scores. Empirical research is usually lacking on the effect of given accommodations on selection procedure performance for candidates with different disabilities or varying magnitudes of the same disability.

2.6.13.1 Responsibilities of the Selection Procedure Developers, Researchers, and Users

Researchers and individuals charged with approving the accommodation for an organisation should be knowledgeable about the availability of modified forms of the selection procedure, psychometric theory, and the likely effect of the disability on selection procedure performance. Users may choose to modify the original selection procedure, develop a modified procedure for candidates with disabilities, or waive the selection procedure altogether and use other information regarding the candidate's job-related knowledge, skills, abilities or other characteristics. While empirical research to demonstrate comparability between the original procedure and the modified procedure may not be feasible in most instances, the individuals developing the modifications should make attempts when possible to limit the modifications, consistent with legal responsibility, to those that allow, insofar as is possible, the comparability of procedures.

Development and validation. Although most employers have too few cases for extensive research, the principles set forth in this document in the preparation of modified selection procedures for candidates with disabilities should be followed to the extent possible. Modified procedures should be pilot-tested with candidates whose disabilities resemble those of the target population when possible and feasible. Practical limitations such as small sample size often restrict the ability of the researcher to statistically equate modified versions of the selection procedure to the original form. These considerations also limit efforts to establish the reliability of the scores and the validity of the inferences made from these scores. Nevertheless, the reliability of selection procedure scores and the validity of inferences based on these scores should be determined whenever possible. In the rare case when it is possible, the effects of administration of the original form of the selection procedure to candidates with disabilities also should be explored.

Documentation and communications regarding accommodations. Descriptions of the modifications made, the psychometric characteristics of the modified selection procedures, and the performance of candidates with disabilities on the original form of the procedure, if available, should be included in the documentation. In addition, selection procedure users should always document the modifications and conditions under which the procedure was administered and keep that information secure and separate from the assessment data in the organisation's records. Legal considerations may prohibit giving decision makers information on whether a candidate's score was earned with a selection procedure accommodation and the nature of the modification. However, users may designate those scores earned with an accommodation in such a way to permit special handling in data analysis.

Selection procedure modification. The test user should take steps to ensure that a candidate's score on the selection procedure accurately reflects the candidate's ability rather than construct-irrelevant disabilities. One of these steps is a dialogue with the candidate with the disability about the accommodations possible. In some cases, the construct being assessed cannot be differentiated from the disability (e.g., proofreading test taken by a sight-impaired candidate). Other times, the disability does not affect selection procedure performance and no accommodation is necessary. Components of a selection procedure battery should be considered separately in determinations of modifications. To the extent possible, standardised features of administration should be retained in order to maximise comparability among scores. Approval of prespecified, routine accommodations not expected to affect the psychometric interpretation of the selection procedure scores (e.g., adjusting table height) may be delegated to administrators.

Maintaining consistency with assessment use in the organisation. The selection procedures used when assessing candidates with disabilities should resemble as closely as possible the selection procedures used for other candidates. The selection procedures are developed for the purpose of making selection decisions, not for the purpose of assessing the extent of a candidate's disability. The addition of a procedure designed to assess the existence or degree of a disability is inappropriate as a selection tool.

SECTION 3
SUMMARY AND CHECKLIST

These *Guidelines* are meant to specify good practice in the choice, development, evaluation and use of personnel assessment procedures. They represent currently accepted practice in personnel assessment and, as such, are not mandates, nor do they specify minimum standards.

A review of the changes that have taken place in the American *Principles* since the first edition in 1975 should assure the reader that the field of personnel assessment is far from static. Certain issues for which the research base is still evolving have not been treated in complete detail in the guidelines. The guidelines primarily focus on the establishment of the validity of a single assessment procedure. In most practical situations, multiple procedures are used and must be combined to yield a decision. The fundamental issue is not how the data is combined, but the validity of the decision.

Although the guidelines are not designed to generate debate, SIOPSA welcomes useful commentary that may aid in the development of future editions.

When any assessment procedure is used, it is used at least with the implicit assumption that some important aspect of behaviour on the job (including performance in training, advancement or other organisationally pertinent behaviour, as well as quality or quantity of job performance) can be predicted from numerical scores on that assessment procedure. The essential principle in the evaluation of any assessment procedure is that evidence should be accumulated to support an inference of job relatedness.

3.1 PLANNING AND ANALYSIS OF WORK

1. Is there a clear statement of the proposed uses of the selection procedures being considered, based on an understanding of the organisation's needs and rights and of its present and prospective employees?
2. Has the user identified the sources of evidence most likely to be relevant for the validation effort, that is, relationships to measures of other variables, content-related evidence, and evidence based on the internal structure of the test?
3. Has the design of the validation effort considered (a) existing evidence, (b) design features required by the proposed uses, (c) design features necessary to satisfy the general requirements of sound inference, and (d) the feasibility of particular design features?
4. Has there been a systematic analysis of work that considers, for example, work complexity; work environment; work context; work tasks, behaviours, and activities performed; or worker requirements (e.g. knowledge, abilities, skills and other personal characteristics [KSAOs])?
5. Does the analysis of work identify worker requirements, as well as criterion measures, by assembling information needed to understand the work performed, the setting in which the work is accomplished, and the organisation's goals?
6. In the analysis of work, is the level of detail appropriate for the intended use and the availability of information about the work?

3.2 SOURCES OF VALIDITY EVIDENCE

1. Does the user understand the construct the selection procedure is intended to measure?
2. If criteria other than job performance are used, is there a theory or rationale to guide the choice of these variables?

3.2.1 Criterion-Related Evidence of Validity

1. Is the choice of predictors and criteria based on an understanding of the objectives for test use, job information, and existing knowledge regarding test validity?
2. Are standardised procedures used? That is, are there consistent directions and procedures for administration, scoring and interpretation?

3.2.1.1 Feasibility

1. Is it possible to obtain or develop a relevant, reliable and uncontaminated criterion measure(s)?
2. Is it possible to do research on a sample that is reasonably representative of the population of people and jobs to which the results are to be generalised?
3. Does the study have adequate statistical power, that is, is there a probability of detecting a significant predictor-criterion relationship in a sample if such a relationship does, in fact, exist?
4. Has the researcher identified design characteristics that might affect the precision of the estimate of predictor-criterion relationship (e.g. sample size, the statistic computed, the probability level chosen for the confidence interval, the size of the relationship)?
5. Is the design, predictive or concurrent, appropriate for the population and purpose of the study?

3.2.2 Design and Conduct of Criterion-Related Studies

3.2.2.1 Criterion Development

1. Are criteria chosen on the basis of work relevance, freedom from contamination and reliability rather than availability?
2. Do all criteria represent important organisational, team and individual outcomes, such as work-related behaviours, outputs, attitudes, or performance in training, as indicated by a review of information about the work?
3. Do adequate safeguards exist to reduce the possibility of criterion contamination, deficiency or bias?
4. Has criterion reliability been estimated?
5. If ratings are used as measures of performance, is the development of rating factors guided by an analysis of the work?
6. Are raters familiar with the demands of the work, as well as the individual to be rated? Are raters trained in the observation and evaluation of work performance?

3.2.2.2 Choice of Predictors

1. Is there an empirical, logical or theoretical foundation for each predictor variable chosen?
2. Is the preliminary choice among predictors based on the researcher's scientific knowledge rather than on personal interest or mere familiarity?
3. Have steps been taken to minimise predictor contamination (e.g. by using standardised procedures, such as structured interviews)?
4. If judgement is used in weighting and summarising predictor data, is the judgement itself recognised as an additional predictor?
5. Has predictor reliability been estimated?

3.2.2.3 Choice of Participants

1. Is the sample for a validation study representative of the selection situation of interest?
2. If a researcher concludes that a variable moderates validity coefficients, is there explicit evidence for such an effect?

3.2.2.4 Procedural Considerations

1. Is validation research directed at entry-level jobs, immediate promotions or jobs likely to be attained?
2. Have alternate criterion-related research methods been considered if they offer a sound rationale (e.g. co-operative research on an industry-wide basis)?
3. Have procedures for test administration and scoring in validation research been set out clearly? Are they consistent with the standardisation plan for operational use?
4. Is there at least presumptive evidence for the validity of a predictor prior to its operational use?
5. Operationally, are predictor data collected independently of criterion measures?

3.2.2.5 Data Analysis for Criterion-Related Validity

1. Has the method of analysis been chosen with due consideration for the characteristics of the data and the assumptions involved in the development of the method?
2. Has the type of statistical analysis to be used been considered during the planning of the research?
3. Does the data analysis provide information about effect sizes and the statistical significance or confidence associated with predictor-criterion relationships?
4. Have the relative risks of Type I and Type II errors been considered?
5. Does the analysis provide information about the nature of the predictor-criterion relationship and how it might be used in prediction (e.g. number of cases, measures of central tendency, characteristics of distributions, variability for both predictor and criterion variables, and interrelationships among all variables studied)?
6. Have adjustments been made for range restriction and/or criterion unreliability, if appropriate, in order to obtain an unbiased estimate of the validity of the predictor in the population in which it will be used?
7. If adjustments are made, have both adjusted and unadjusted validity coefficients been reported?
8. If predictors are to be used in combination, has careful consideration been given to the method used to combine them (e.g. in a linear manner, by summing of scores on different tests; or in a configural manner, by using multiple cut-offs)?
9. If a researcher combines scores from several criteria into a composite score, is there a rationale to support the rules of combination, and are the rules described?
10. Have appropriate safeguards been applied (e.g. use of cross-validation or shrinkage formulas) to guard against overestimates of validity resulting from capitalisation on chance?
11. Have the results of the present criterion-related validity study been interpreted against the background of previous relevant research literature?
12. Are unusual findings, such as suppressor or moderator effects, nonlinear regression, or the benefits of configural scoring, supported by an extremely large sample or replication?
13. Has there been an assessment of the practical value or utility of the selection procedure?
14. Has all keypunching, coding, and computational work been checked carefully and thoroughly to ensure that data are free from clerical error?

3.2.3 Evidence for Validity Based On Content

1. If a selection procedure has been designed explicitly as a sample of important elements in the work domain, does the validation study provide evidence that the selection procedure samples the important work behaviours, activities or worker KSAOs necessary for performance on the job or in training?
2. Are the work and worker requirements reasonably stable?
3. Are qualified and unbiased subject matter experts available?
4. Does the content-based procedure minimise elements that are not part of the work domain (e.g. multiple-choice formats or written content when the job does not require writing)?
5. Has each job content domain been defined completely and described thoroughly in terms of what it does and does not include, based on, for example, an analysis of work behaviours and activities, responsibilities of job incumbents or KSAOs required for effective performance on the job?
6. Has the researcher described the rationale underlying the sampling of the content domain?
7. Is the selection procedure based on an analysis of work that defines the balance between work behaviours, activities or KSAOs the applicant is expected to have before placement on the job and the amount of training the organisation will provide?
8. Does the specificity-generality of the content of the selection procedure reflect the extent to which the job is likely to change as a result of organisational needs, technology or equipment?
9. Has the researcher established guidelines for administering and scoring the content-based procedure?
10. Has the reliability of performance on content-based selection procedures been determined?
11. Is the job content domain restricted to critical or frequent activities or to prerequisite knowledge, skills, or abilities?
12. Where appropriate, have special circumstances been considered in defining job content domains (e.g. different locations of seldom-used symbols on different keyboards)?
13. Have scoring keys for content-oriented tests been checked for accuracy? For answers keyed as correct, are they correct under all reasonable-expected job-relevant circumstances?

3.2.4 Evidence of Validity Based on Internal Structure

1. Does the researcher recognise that evidence of internal structure, by itself, is insufficient to establish the usefulness of a selection procedure in predicting future work performance?
2. Are relevant analyses based on the conceptual framework of the selection procedure (typically established by the proposed use of the procedure)?
3. If evidence of validity is based on internal structure, did the researcher consider the relationship between items, components of the selection procedures or scales measuring constructs?
4. Is the inclusion of items in a selection procedure based primarily on their relevance to a construct or content domain and secondarily on their intercorrelations?
5. If scoring involves a high level of judgement, does the researcher recognise that indices of interrater or scorer consistency, such as generalisability coefficients or measures of interrater agreement, may be more appropriate than internal consistency estimates?

3.3 GENERALISING VALIDITY EVIDENCE

1. If a researcher wishes to generalise the validity of inferences from scores on a selection procedure to a new situation, based on validation research conducted elsewhere, is such transportability based on job comparability (in content or requirements) or similarity of job context and candidate group?

2. If synthetic or job component validity is used as a basis for generalising the validity of inferences from scores on a selection procedure, has the researcher documented the relationship between the selection procedure and one or more specific domains of work (job components) within a single job or across different jobs?
3. If meta-analysis is used as a basis for generalising research findings across settings, has the researcher considered the meta-analytic methods used, their underlying assumptions, the tenability of the assumptions and artifacts that may influence the results?
4. Are reports that contribute to the meta-analytic research results clearly identified and available?
5. Have researchers fully reported the rules they used to categorise jobs, tests, criteria and other characteristics of their studies? Have they reported the reliability of the coding schemes used to categorise these variables?
6. Are there important conditions in the operational setting that are not represented in the meta-analysis (e.g. the local setting involves a managerial job and the meta-analytic database is limited to entry-level jobs)?
7. If the cumulative validity evidence in a meta-analysis is relied on for jobs in new settings or organisations, are the following conditions met?
    a. Is the selection procedure to be used a measure of the trait, ability or construct studied? Is it a representative sample of the type of selection procedure included in the meta-analysis?
    b. Is the job in the new setting similar to, or a member of, the same job family as the job included in the validity generalisation study?
8. Is the researcher attempting to generalise on the basis of a method in general (e.g. interviews, biodata) rather than on the basis of a specific application of the method?

3.4 FAIRNESS AND BIAS

1. Does the researcher recognise that fairness has no single definition, whether statistical, psychometric or social?
2. Has the researcher tested for predictive bias (consistent nonzero errors of prediction for members of a subgroup) when there are compelling reasons to question whether a predictor and a criterion are related in a comparable fashion for specific subgroups, given the availability of appropriate data?
3. If a test of predictive bias is warranted, has the researcher tested for it using moderated multiple regression?
4. Do tests for predictive bias meet the following conditions: use of an unbiased criterion, sufficient statistical power, and homogeneity of error variances?
5. Has the researcher conducted an item sensitivity review, in which items are reviewed by individuals with diverse perspectives for language or content that might have differing meaning for members of various subgroups and for language that could be demeaning or offensive to members of various subgroups?

3.5 OPERATIONAL CONSIDERATIONS

3.5.1 Initiating a Validation Effort

1. Have all aspects of the research been performed in compliance with the ethical standards of the Health Professions Council of South Africa?
2. In defining an organisation's needs, objectives and constraints, have the researcher and the organisation's representative taken into account the wishes of various stakeholders and determined the relative weights to be given to each point of view?

3. Have researchers considered the legal and labour environments when deciding on validation approaches or selection instruments?
4. In choosing a validation strategy, has the researcher considered the number of individuals who currently perform the work and their similarity to the applicant population?
5. Has the researcher considered alternative sources of information for the validation effort, such as workers, managers, supervisors, trainers, customers, archival records, databases, and internal and external reports?
6. Has the researcher explained to decision-makers the issues underlying the acceptability of a selection procedure as part of the initial planning effort?
7. Do managers and workers understand in general terms the purpose of the research, the plan for conducting the research, and their respective roles in the development and validation of the selection procedure?

3.5.2 Understanding Work and Worker Requirements

1. In cases where traditional jobs no longer exist, has the researcher considered important requirements for a wider range or type of work activity?
2. Does the sampling plan for data collection take into account the number of workers and their locations, their characteristics (experience, training, proficiency), their shift or other work cycles, and other variables that might influence the analysis of work?
3. In documenting the work-analysis process, has the researcher described the data-collection methods, analyses, results, and implications for the validation effort?

3.6 REQUIREMENTS

3.6.1 Selecting Assessment Procedures for the Validation Effort

1. Is the researcher familiar with research related to the organisation's objectives?
2. In choosing components of a selection battery, has the researcher considered the overall contribution of each component, its relative contribution and potential construct redundancy?
3. Has the researcher ensured that administration and scoring tasks can be completed consistently across all locations and administrators?
4. Has the researcher carefully considered the format (e.g. multiple-choice, essay) and medium (i.e. the method of delivery) of the content of the selection procedure?
5. Have researchers considered approaches designed to minimise negative perceptions of a selection procedure and to enhance its acceptability to candidates?
6. If alternative forms of a selection procedure are developed, has the researcher taken steps to ensure that candidates' scores are comparable across forms?

3.6.2 Selecting the Validation Strategy

1. Is the strategy selected feasible in the organisational context, and does it meet project goals within the constraints imposed by the situation?
2. When individual assessment is used (one-on-one evaluations), does the assessor have a rationale for determining and using selection procedures?

3.6.3 Selection Criterion Measures

1. Has the researcher considered the psychometric characteristics of performance-oriented criteria (those that represent work activities, behaviours or outcomes, such as supervisory ratings)?

2. Are all criteria representative of important work behaviours, outcomes or relevant organisational expectations regarding individual behaviour or team performance?

3.6.4 Data Collection

1. Has the researcher communicated relevant information about the data-collection effort to all those affected, including management, test takers, those who provide criterion data and those who will use the test?
2. Has the researcher determined the extent to which pilot testing is necessary or useful?
3. Have participants in the validation research been given confidentiality unless there are persuasive reasons to proceed otherwise?
4. Have all data been retained at a level of security that permits access only for those with a need to know?

3.6.5 Data Analyses

1. Have all data been checked for accuracy?
2. Is there a documented rationale for handling missing data or outliers?
3. Are data analyses appropriate for the method or strategy undertaken, the nature of the data (nominal, ordinal, interval, ratio), the sample sizes, and other considerations that will lead to correct inferences from the data?
4. If selection procedures are combined, have the algorithm for combination and the rationale for the algorithm been described?
5. Have the rationale and supporting evidence for the use of multiple hurdles or a compensatory model been presented?
6. In recommending the use of a rank-ordering method or a cutoff score, does the recommendation take into account labour-market conditions, the consequences of errors in prediction, the level of a KSAO represented by a chosen cutoff score, and the utility of the selection procedure?
7. If test-score banding is used, has the researcher documented the basis for its development and the decision rules to be followed in its administration?
8. Has the researcher presented normative information relevant to the applicant pool and the incumbent population?

3.7 COMMUNICATING THE EFFECTIVENESS OF SELECTION PROCEDURES

1. Has the researcher used expectancy or utility analyses to communicate the effectiveness of selection procedures?
2. Has the researcher identified the results of utility analyses as estimates based on a set of assumptions?
3. Have minimal and maximal point estimates of utility been presented to reflect the uncertainty in estimating various parameters of the utility model?

3.8 APPROPRIATE USE OF SELECTION PROCEDURES

1. Has the researcher produced evidence of validity to support individual components as well as the combination of selection procedures?
2. Are selection procedures used only for the purposes for which there is validity evidence?
3. Are the recommendations based on the results of a validation effort consistent with the objectives of the research, the data analyses performed, and the researcher's professional judgement and ethical responsibilities?

3.8.1 Technical Validation Report

1. Do all reports of validation research include the name of the author and date of the study, a statement of the purpose of the research, a description of the analysis of work, and documentation of any search for alternative selection procedures?
2. Are the names, editions and forms of commercially available selection instruments described? For proprietary instruments, has the researcher described the items, the construct(s) that are measured, and sample items, if appropriate?
3. Does the report describe the methods used by the researcher to determine that the selection procedure is significantly related to a criterion measure or representative of a job content domain?
4. Does the report provide a detailed description of criterion measures; the rationale for their use; data-collection procedures; and a discussion of their relevance, reliability, and freedom from bias?
5. Does the report describe the research sample and the sampling procedure relative to the interpretation of results? Does it provide data regarding restriction in the range of scores on predictors or criteria?
6. Are all summary data available that pertain to the conclusions drawn by the researcher and to his or her recommendations?
7. Are the methods used to score items and tasks described fully?
8. Are norm or expectancy tables presented to help guide relevant interpretations?
9. Does the report give recommendations for implementation and the rationale supporting them (e.g. rank-ordering, score bands, cutoff scores)?
10. Have all research findings that might qualify the conclusions or the generalisability of results been reported?
11. Are complete references given for all published literature and available technical reports (some of which may be proprietary and confidential)?

3.8.2 Administration Guide

1. Does the administration guide document completely the information needed to administer the selection procedure, score it, and interpret the score?
2. If the selection procedure is computer based or in a form other than paper and pencil, does the guide include detailed instructions on the special conditions of administration?
3. Is the information developed for users or examinees accurate and complete for its purposes and not misleading?
4. Does the writing style meet the needs of the likely audience?
5. Does the guide include an introduction to inform the reader of the purpose of the assessment procedure and to give an overview of the research that supports the procedure?
6. Does the guide include contact information, a thorough description of the selection procedures, and an indication of persons to whom the procedure is applicable, and does it state any exceptions to test requirements?
7. Does the administration guide state the necessary qualifications of administrators and the training required to administer the procedures described in the guide?
8. Does the guide give detailed instructions for the actual implementation of the selection procedures, as well as rules and tips for providing an appropriate testing environment and for ensuring the candidate's identity?
9. Does the guide include detailed instructions for scoring and interpreting the results of the selection procedure?
10. Have quality control checks been implemented to ensure accurate scoring and recording?

11. If computer-based test interpretation (CBTI) is used to process responses to a selection procedure, did the researcher provide detailed instructions on how CBTI is to be used in decision-making?
12. Does the guide provide detailed information on recordkeeping and test-score databases?
13. Does the guide communicate the way in which selection procedure scores are to be reported and used and who has access to them?
14. Does the guide include information about how to provide feedback to candidates?
15. Does the guide communicate general principles about how persons with disabilities or how deviations from normal procedures (e.g. sessions disrupted by power failures or illness of a candidate) are to be handled?
16. Does the guide explain whether candidates may be reassessed and how reassessment will take place?
17. Does the administration guide emphasise the importance of safeguarding the content, scoring and validity of the selection procedure, and does it identify practices for ensuring the security of selection-procedure documents?

3.8.3 Other Circumstances Regarding the Validation Effort and Use of Selection Procedures

1. If advised of changes in organisational functioning, does the researcher examine each situation on its own merits and make recommendations regarding the impact of the change on the validation and use of any selection procedure?
2. Does the researcher periodically review and, if necessary, update selection procedures and their technical or administration guides?
3. For candidates with disabilities, does the user make special accommodations to minimise the impact of a known disability that is not relevant to the construct being assessed?
4. Are researchers and individuals charged with approving accommodations knowledgeable about the availability of modified forms of the selection procedure, psychometric theory, and the likely effect of the disability on selection-procedure performance?
5. Although most employers have too few cases for extensive research, are the principles set out in this document followed to the extent possible in the preparation of modified selection procedures for candidates with disabilities?
6. Is there documentation of the modifications made, the psychometric characteristics of the modified selection procedures, and the performance of candidates with disabilities on the original form of the procedure (if available)?
7. Does the test user take steps to ensure that a candidate's score on the selection procedure accurately reflects his or her ability rather than construct-irrelevant disabilities?

USEFUL WEB ADDRESSES

The American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education's *Standards for Educational and Psychological Testing*. (Ordering information available from http://www.apa.org/science/standards.html).

The American Psychological Association's *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists*, available from http://www.apa.org/pi/ multiculturalguidelines/education.html.

The International Test Commission's *International Guidelines for Test Use*, available from www.intestcom.org.

The Health Professions Council of South Africa's *Ethical Code for Professional Conduct*, available from www.hpcsa.co.za.

The International Test Commission's *International Guidelines on Computer-Based and Internet Delivered Testing*, available from www.intestcom.org.

The European Federation of Psychologists Association's *Test Review Guidelines*, available from www.efpa.be.

The Society for Industrial and Organizational Psychology, Inc. (USA). www.siop.org

REFERENCES

Aamodt, M.G. & Kimbrough, W.W. (1985). Comparison of four methods for weighting multiple predictors. *Educational and Psychological Measurement, 45,* 477–482.

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF).* Unpublished doctoral thesis. Pretoria, South Africa: University of Pretoria.

Abrahams, F. (2002). Fair usage of the 16PF (SA 92) in South Africa: A response to C.H. Prinsloo & I. Ebersohn. *South African Journal of Psychology,* 32, 58-61.

Abrahams, F. & Mauer, K.F. (1999a). The comparability of the constructs of the 16PF in the South African context. *Journal of Industrial Psychology,* 25, 53-59.

Abrahams, F. & Mauer, K.F. (1999b) Qualitative and statistical impact of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in South African context. *South African Journal of Psychology*, 29, 76-86.

Ackerman, P.J. & Humphreys, L.G. (1990). Individual differences theory in industrial and organisational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organisational psychology* (Vol. 1). Palo Alto, CA: Consulting Psychologists Press.

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research*. Journal of Management, 21,* 1141–1158*.*

Aguinis, H., Petersen, S.A., & Pierce, C.A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organisational Research Methods, 2,* 315–339.

Aguinis, H., & Pierce, C.A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management*, *24*, 577–592.

Aguinis, H. & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192–206.

Alexander, R.A., & DeShon, R.P. (1994). The effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, *115*, 308–314.

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association (2003). Guidelines on multicultural education, training, research, practice, and organisational change for psychologists. *American Psychologist, 58*(5), 337–402.

Arvey, R.D. & Faley, R.H. (1988). *Fairness in selecting employees*. Reading, MA: Addison Wesley Publishing Company.

Anderson, N. (2005). *Future trends in employee assessment: towards an international science and practice.* Paper presented to the 25th Assessment Centre Study Group Conference in South Africa, Cape Town, 4 March 2005.

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

Barrick, M.R., Mount, M.K. & Judge, T.A. (2001). Personality and performance at the beginning of the new millennium: What do we know and what do we do next? *International Journal of Selection and Assessment*, 9,9-30.

Barrick, M.R. & Mount, M.K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26.

Bartlett, C.J., Bobko, P., Mosier, S.B. & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology*, *31*, 233–242.

Bedell, B., van Eeden, R & van Staden, F. (1999). Culture as moderator variable in psychological test performance: Issues and trends In South Africa. *South African Journal of Industrial Psychology*. 25(3). 1-7.

Bemis, S.E. (1968). Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology, 52,* 240–249.

Bobko, P. & Riecke, A. (1980). Large sample estimates for standard errors of functions of correlation coefficients. *Applied Psychological Measurement, 4*, 385–398.

Bobko, P. & Stone-Romero, E. (1998). Meta-analysis is another useful research tool but it is not a panacea. In G. Ferris (Ed.), *Research in personnel and human resources management*, Vol. 16, pp. 359–397, Greenwich, CT: JAI Press.

Boudreau, J.W. (1991). Utility analysis for decisions in human resource management. In M.D. Dunnette & L.M. Hough (Eds.). *Handbook of industrial or organizational psychology* (2nd edition; volume 2). Palo Alto, California: Consulting Psychologists Press, Inc.

Boudreau, J.W. & Berger, C.J. (1985). Decision-theoretic utility analysis applied to employee separations and acquisitions. *Journal of Applied Psychology,* 70, 581-612.

Brannick, M.T. & Levine, E.L. (2002). *Job analysis: Methods, Research, and Applications for Human Resource Management in the New Millennium.* Sage Publications.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology, 2*, 171–183.

Brogden, H.E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *The Journal of Educational Psychology,* 37, 65-76.

Callender, J.C., & Osburn, H.G. (1980). Development and testing of a new model of validity generalisation. *Journal of Applied Psychology*, *65,* 543–558.

Callender, J.C. & Osburn, H.G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance method estimate: Results for petroleum industry validation research. *Journal of Applied Psychology, 66,* 274–281.

Campion, M.A., Outtz, J.L., Zedeck, S., Schmidt, F., Kehoe, J. F., Murphy, K.R. & Guion, R.M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology, 54*(1), 149–185.

Cascio, W. F. (2000*). Costing human resources: The financial impact of behaviour in organisations* (4th ed.). Cincinnati, OH: Southwestern.

Cascio, W.F. & Aguinis, H. (2005). *Applied Psychology in Human Resource Management* (6th ed.). USA: Pearson Prentice Hall.

Cascio, W.F., Outtz, J., Zedeck, S. & Goldstein, I.L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance*, *4,* 233–264.

Claasen, N.C.W. (1997). Culture differences, politics and test bias in South Africa. *European Review of Applied Psychology,* 47, 297-307.

Cleary, T.A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

Coward, W.M. & Sackett, P.R. (1990). Linearity of ability-performance relationships: A reconfirmation. *Journal of Applied Psychology, 75,* 297–300.

Cronbach, L.J. & Gleser, G.C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Cronshaw, S.F. & Alexander, R.A. (1985). One answer to the demand for accountability: assessment utility as an investment decision. *Organizational Behavior and Human Decision Processes,* 35, 102-118.

De Jong, A. & Visser, D. (2000). Black and White employees' fairness perceptions of personnel selection techniques. *South African Journal of Psychology,* 30(4), 17–24.

De Villiers, K. (1997). *Die differensiele voorspelling van universiteitsprestasie by studente met diverse onderwysagterground.* Unpublished Masters Dissertation.

DeShon, R.P. & Alexander, R.A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal*. Psychological Methods, 1,* 261–277*.*

Dunnette, M.D., Hough, L.M. & Triandis, H.C. (1991). *Handbook of Industrial and Organisational Behaviour*. Palo Alto: Consulting Psychologist Press.

Employment Equity Act, No. 55 of 1998. *Government Gazette,* Vol. 400, No. 19370. Republic of South Africa, 19 October 1998.

Esterhuyse, K.G.F. & Van der Walt, H.S. (1995). A programme for the assessment of candidates for training as chartered accountants (Title translated)*. Acta Academia* 27(1): 129-142.

Gael, S. (1983). *Job analysis: A guide to assessing work activities.* San Francisco, CA: Jossey-Bass.

Geisinger, K.F. & Carlson, J.F. (1992). Assessing language-minority students. *Practical Assessment, Research & Evaluation*, 3(2).

Goldstein, I.L., Zedeck, S. & Schneider, B. (1993). An exploration of the job-analysis-content validity process. In N. Schmitt and W. Borman (Eds.), *Personnel selection in organisations*. San Francisco: Jossey-Bass.

Guion, R.M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum.

Gulliksen, H. & Wilks, S.S. (1950). Regression tests for several samples. *Psychometrika, 15,* 91–114. Hartigan, J. A., & Wigdor, A. K. (Eds.) (1989). *Fairness in employment testing*. Washington, DC: National Academy Press.

Holtzhausen, G. (2004). *Mode of administration and the stability of the OPQ32n: comparing internet (controlled) and paper-and-pencil (supervised) administration.* Unpublished Masters dissertation.

Hough, L.M., Oswald, F.L. & Ployhart, R.E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment*, *9,* 1–42.

Humphreys, L.G. (1952). Individual differences. *Annual Review of Psychology, 3,* 131–150.

Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–88.

Hunter, J. E., & Schmidt, F. L. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Hunter, J.E. & Schmidt, F.L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6*, 151–158.

Hunter, J. E., Schmidt, F. L., & Judiesch M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75,* 28–42.

Hunter, J. E., Schmidt, F.L. & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on mental testing.* New York: Plenum*.*

Huysamen, G.K. (1995). The applicability of fair selection models in the South African context. *Journal of Industrial Psychology,* 21(3), 1–6.

Huysamen, G.K. (2002). The relevance of the new APA standards for educational and psychological testing for employment testing In South Africa. *South African Journal of Psychology*, 32(2), 26-33.

Jawahar, I.M. & Williams, C.R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50,* 905–925.

Jeanneret, R. & Silzer, R. (Eds.). (1998). *Individual psychological assessment*. San Francisco: Jossey-Bass.

Kriek, H.J., Hurst, D.N. & Charoux, J.A.E. (1994). The assessment centre: testing the fairness hypothesis. *Journal of Industrial Psychology,* 20(2), 21-25.

Lautenschlager, G.J. & Mendoza, J.L. (1986). A step down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10*, 133–139.

Law, K.S., Schmidt, F.L. & Hunter, J.E. (1994a). Nonlinearity of range corrections in meta-analysis: Test of an improved procedure*. Journal of Applied Psychology, 79,* 425–438.

Law, K.S., Schmidt, F.L. & Hunter, J.E. (1994b). A test of two refinements in procedures in meta-analysis*. Journal of Applied Psychology, 79,* 978–986.

Linn, R.L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology,* 63, 507-512.

Lopes, A., Roodt, G., & Mauer, R. (2001). The predictive validity of the APIL-B In a financial Institution. *Journal of Industrial Psychology, 27*, 61 - 69.

McCormick, E.J. (1979). *Job analysis: Methods and applications.* New York: AMACOM.

Mead, A.D. & Drasgow, F. (1993). Equivalence of computerised and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449–458.

Meiring, D. (2000). *Revisiting the cross-cultural comparability of the 16 Personality Factor Inventory (16PF) in the South African context.* Paper presented at the Industrial Psychology Conference (incorporating the Psychometrics Conference), Pretoria, South Africa.

Meiring, D., Van de Vijver, A.J.R., Rothmann, S. & Sackett, P.R. (2005). *Differential prediction of cognitive and personality measures in South Africa.* In press.

Meiring, D., Van de Vijver, A.J.R., Rothmann, S. & Barrick, M.R. (2005). *Construct, item, and method bias of cognitive and personality tests in South Africa.* Manuscript submitted to the South African Journal of Industrial Psychology.

Naylor, J.C. & Shine, L.C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology, 3*, 33–42.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

Oswald, F.L. & Johnson, J.W. (1998). On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: Some initial Monte Carlo findings. *Journal of Applied Psychology, 83,* 164–178.

Oswald, F., Saad, S. & Sackett, P.R. (2000). The homogeneity assumption in differential prediction analysis: Does it really matter? *Journal of Applied Psychology, 85*, 536–541.

Owen, K. (1989). Test and item bias: The suitability of the Junior Aptitude Tests as a common test battery for white, Indian and black pupils in Standard 7. Report P-96. Human Sciences Research Council.

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

Petersen, N.S. & Novick, M.R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, *13*, 3–29.

Peterson, N.G., Mumford, M.D., Borman, W.C., Jeanneret, P.R. & Fleishman, E.A. (Eds). (1999). *An occupational information system for the 21st century: The development of O*Net.* American Psychological Association: Washington, DC.

Rademan, D.J. & Vos, H.D. (2001). Performance appraisals in the public sector: Are they accurate and fair? *Journal of Industrial Psychology, 27*, 54–60.

Raju, N.S., Anselmi, T.V., Goodman, J.S. & Thomas, A. (1998). The effect of correlated artefacts and true validity on the accuracy of parameter estimation in validity generalisation. *Personnel Psychology, 51,* 453–465.

Raju, N.S. & Brand, P.A. (in press). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement.*

Raju, N.S., Burke, M.J. & Normand, J. (1990). A new approach for utility analysis. *Journal of Applied Psychology, 75,* 3–12.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology, 76,* 432–446.

Raju, N.S., Pappas, S. & Williams, C.P. (1989). An empirical Monte Carlo test of the accuracy of the correlation, covariance, and regression slope models for assessing validity generalisation. *Journal of Applied Psychology, 74*, 901–911.

Richardson, M.W. (1941). Combination of measures. In P. Horst (Ed.). *The prediction of personal adjustment.* New York: Social Science Research Council.

Rotundo, M. & Sackett, P. (2002) The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy capturing approach. *Journal of Applied Psychology*, 87, 1: 66-80.

Ryan, A.M. & Sackett, P.R. (1998). Individual assessment: The research base. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment* (pp. 54–87). San Francisco: Jossey-Bass.

Saad, S. & Sackett, P.R. (2002). Examining differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology, 87,* 667–674.

Sackett, P.R., Harris, M.M. & Orr, J.M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology, 71,* 302–310.

Sackett, P.R. & Roth, L. (1996). Multi-stage selection strategies: a Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49*, 549–572.

Sackett, P.R., Schmitt, N., Ellingson, J.E. & Kabin, M.B. (2001). High stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302–318.

Schepers, J.M. (1995). *The development of a statistical procedure to correct the effects of restriction of range on validity coefficients*. Paper read at SIOPSA Conference on Psychometrics Pretoria , June 1995.

Schippmann, J.S., Ash, R.A., Battista, M., Carr, L., Eyde, L.D., Hesketh, B., Kehoe, J., Pearlman, K., Prien, E.P. & Sanchez, J.I. (2000). The practice of competency modeling. *Personnel Psychology, 53*, 703–740.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124,* 262–274.

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. & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology, 66,* 166–185.

Schmidt, F.L., Mack, M.J. & Hunter, J.E. (1984). Assessment utility in the occupation of US park rangers for three modes of test use. *Journal of Applied Psychology,* 69, 490–497.

Schmidt, F.L., Pearlman, K. & Hunter, J.E., (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33,* 705–724.

Schneider, B. & Konz, A. (1989). Strategic job analysis. *Human Resources Management, 28*, 51–63.

Society for Industrial and Organisational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures*. (3rd ed.). College Park, MD: Author.

Taylor, H.C. & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

Van der Merwe, R.P. (2002). Psychometric Testing and Human Resource Management. *South Journal of Industrial Psychology*, 28(2), 77–86.

Van de Vijver, F. & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. London: Sage.

Van der Walt, H.S., Meiring, D., Rothmann, S. & Barrick, M.R. (2002*). Meta-analysis of the relationship between personality measurements and job performance in South Africa.* Paper read at SIOPSA conference, Pretoria, June 2002.

Wheeler, H.L. (1993). *The fairness of an engineering selection battery in the mining industry*. Unpublished Masters Dissertation.

GLOSSARY OF TERMS

Ability A defined domain of cognitive, perceptual, psychomotor, or physical functioning.

Accommodation A change in the content, format, and/or administration of a selection procedure made to eliminate an irrelevant source of score variance resulting from a test taker's disability.

Adverse impact The use of a test or other selection procedure results in a substantially different rate of selection in hiring, promotion, or other employment decisions that works to the disadvantage of members of a designated group.

Adjusted validity/reliability coefficient A validity or reliability coefficient (most often a product-moment correlation) that has been adjusted to offset effects of differences in score variability, criterion variability, or unreliability of test and/or criterion. See *Restriction of range or variability*.

Alternate forms Two or more versions of a selection procedure that are considered interchangeable in that they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. *Alternate forms* is a generic term used to refer to either parallel forms or equivalent forms. *Parallel forms* have equal raw score means, equal standard deviations, equal error structures, and equal correlations with other measures for any given population. *Equivalent forms* do not have the statistical similarity of parallel forms, but the dissimilarities in raw score statistics are compensated for in the conversions to derived scores or in form-specific norm tables.

Analysis of work Any method used to gain an understanding of the work behaviours and activities required, or the worker requirements (e.g., knowledge, skills, abilities, and other personal characteristics), and the context or environment in which an organisation and individual may operate.

Assessment Any systematic method of obtaining information from tests and other sources; used to draw inferences about characteristics of people.

Assessment instrument Any method or device used to evaluate characteristics of a person.

Assessment procedure Process of arriving at an assessment decision.

Attribute A fundamental or characteristic property of people, situations, and things.

Battery A set of selection procedures administered as a unit.

Behaviour Observable aspects of a person's activities.

Bias In a statistical context, a systematic error in a score; a distorting factor or error in a set of data or in an experimental design. In discussing fairness, bias refers to variance due to contamination or deficiency that differentially affects the scores of different groups of individuals.

Compensatory model Two or more individual selection procedure component scores (often individual test scores) combined into a composite selection procedure according to some specified formula (including simple summation of scores and unit weighting). As a consequence of combining scores, some compensation for one or more of the constructs measured may occur due to differential performance on the individual selection procedures (i.e., a higher score on one test compensating for a lower score on another test).

Composite score A score that combines scores from several individual selection procedures according to a specified formula.

Concurrent validity evidence Demonstration of the relationship between job performance and other work outcomes, and scores on selection procedures obtained at approximately the same time.

Confidence interval An interval between two values on a score scale within which, with specified probability, a score or parameter of interest is expected to lie.

Configural scoring rule (Configural scoring) A rule for scoring a set of two or more elements (such as items or subtests) in which the score depends on a particular pattern of responses to the elements.

Consequence-based evidence Evidence that consequences of selection procedure use are consistent with the intended meaning or interpretation of the selection procedure.

Construct A concept or characteristic of individuals inferred from empirical evidence or theory.

Construct irrelevance The extent to which scores on a predictor are influenced by factors that are irrelevant to the construct. Such extraneous factors distort the meaning of scores from what is implied in the proposed interpretation.

Contamination Systematic variance that is irrelevant to the intended meaning of the measure.

Content domain A body of knowledge and/or set of tasks, activities, or other personal characteristics defined so that given knowledge, activities, or characteristics may be classified as included or excluded.

Content-based validity evidence Demonstration of the extent to which content on a selection procedure is a representative sample of work-related personal characteristics, work performance or other work activities or outcomes.

Convergent validity evidence Evidence of a relationship between measures intended to represent the same construct.

Correlation The degree to which two sets of measures vary together.

Criterion A measure of work performance or behaviour, such as productivity, accident rate, absenteeism, tenure, reject rate, training score, and supervisory ratings of job relevant behaviours, tasks or activities.

Criterion-related validity evidence Demonstration of a statistical relationship between scores on a predictor and scores on a criterion measure.

Criterion relevance The extent to which a criterion measure reflects important work performance dimensions or other work outcomes.

Critical score A specified point in a distribution of scores at or above which candidates are considered successful in the selection process. The critical score differs from cutoff score in that a critical score is by definition criterion referenced (i.e., the critical score is related to a minimally acceptable criterion) and is the same for all applicant groups.

Cross-validation The application of a scoring system or set of weights empirically derived in one sample to a different sample from the same population to investigate the stability of relationships based on the original weights.

Cutoff score A score at or above which applicants are selected for further consideration in the selection process. The cutoff score may be established on the basis of a number of considerations (e.g., labour market, organisational constraints, normative information). Cutoff scores are not necessarily criterion referenced, and different organisations may establish different cutoff scores on the same selection procedure based on their needs.

Derived score A score that results from a numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores) of the original selection procedure score.

Differential item functioning A statistical property of a test item in which different groups of test takers who have the same standing on the construct of measurement have different average item scores or, in some cases, different rates of endorsing various item options. Also known as DIF.

Differential prediction The case in which use of a common regression equation results in systematic nonzero errors of prediction for subgroups.

Discriminant validity evidence Evidence of a lack of relationship between measures intended to represent different constructs.

Expectancy table A table or chart used for making predictions of levels of criterion performance for specified intervals of predictor scores.

Fairness There are multiple perspectives on fairness. There is agreement that issues of equitable treatment, predictive bias, and scrutiny for possible bias when subgroup differences are observed are important concerns in personnel selection; there is not, however, agreement that the term "fairness" can be uniquely defined in terms of any of these issues.

Feasible Capable of being done successfully; in criterion-related research, economically practical and technically possible without misleading or uninterpretable results.

Generalised evidence of validity Evidence of validity that generalises to setting(s) other than the setting(s) in which the original validation evidence was documented. Generalised evidence of validity is accumulated through such strategies as transportability, synthetic validity/job component validity, and meta-analysis.

Informed consent The individual's freedom of choice in terms of what actions will take place and the right to be informed about assessment procedures.

Internal consistency reliability An indicator of the reliability of a score derived from the statistical interrelationships of responses among item responses or scores on different parts of an assessment.

Internal structure validity evidence Demonstration of the degree to which psychometric and statistical relationships among items, scales, or other components within a selection procedure are consistent with the intended meaning of scores on the selection procedure.

Inter-rater agreement The consistency with which two or more judges rate the work or performance of examinees.

Item A statement, question, exercise, or task on a selection procedure for which the test taker is to select or construct a response, or perform a task.

Item response theory (IRT) A mathematical model of the relationship between performance on a test item and the test taker's standing on a scale of the construct of measurement, usually denoted as θ. In the case of items scored 0/1 (incorrect/correct response) the model describes the relationship between θ and the item mean score (P) for test takers at level θ, over the range of permissible values of θ. In most applications, the mathematical function relating P to θ is assumed to be a logistic function that closely resembles the cumulative normal distribution.

Job component validity See *Synthetic validity evidence.*

Job description A statement of the work behaviours and activities required or the worker requirements (e.g., knowledge, skills, abilities, and other personal characteristics).

Job Knowledge Information (often technical in nature) needed to perform the work required by the job.

Job relatedness The inference that scores on a selection procedure are relevant to performance or other behaviour on the job; job relatedness may be demonstrated by appropriate criterion-related validity coefficients or by gathering evidence of the job relevance of the content of the selection instrument, or of the construct measured.

KSAOs Knowledge, skills, abilities, and other personal characteristics required in completing work in the context or environment in which an organisation and individual may operate.

Linear combination The sum of scores, whether weighted differentially or not, on different assessments to form a single composite score.

Local evidence Evidence (usually related to reliability or validity) collected in a single organisation or at a specific location.

Local study (local setting) See *Local evidence.*

Measurement bias See *Bias.*

Meta-analysis A statistical method of research in which the results from several independent studies of comparable phenomena are combined to estimate a parameter or the degree of relationship between variables.

Moderator variable A variable that affects the strength, form, or direction of a predictor-criterion relationship.

Multiple-hurdle model The implementation of a selection process whereby two or more separate procedures must be passed sequentially.

Multivariate Characterising a measure or study that incorporates several variables.

Non-linear combination A procedure for combing scores on tests by means other than their sums, such as logarithmic or exponential transformation.

Non-linear regression When, in the equation describing the prediction of a criterion variable (Y) from a predictor variable (X), namely Y - bX + K, b or X are not simple numbers (e.g. logarithms or powers other than 1).

Normative Pertaining to norm groups or the sample on which descriptive statistics (e.g., mean, standard deviation, etc.) or score interpretations (e.g., percentile, expectancy, etc.) are based.

Norms Statistics or tabular data (often raw and percentile scores) that summarize performance of a defined group on a selection procedure.

Objective Pertaining to scores obtained in a way that minimises bias or error due to different observers or scorers.

Operational independence Gathering of data by methods that are different in procedure or source so that measurement of one variable, such as a criterion, is not influenced by the process of measuring another variable.

Operational setting The specific organisation, work context, applicants, and employees to which a selection procedure is applied.

Performance The effectiveness and value of work behaviour and its outcomes.

Personal characteristics Traits or dispositions that describe individuals.

Population The universe of cases from which a sample is drawn and to which the sample results may be projected or generalised.

Personnel practitioner A professionally trained individual abiding by the external guidelines identified by the Board for Personnel Practice.

Power The probability that a statistical test will yield statistically significant results if an effect of specified magnitude exists in the population.

Practical significance The property of having important consequences for an action.

Predictive bias The systematic under- or overprediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance.

Predictive validity evidence Demonstration of the relationship between selection procedure scores and some future work behaviour or work outcomes.

Predictor A measure used to predict criterion performance.

Predictor-criterion relationship The relationship between a predictor and external criteria (e.g., job performance, tenure) or other predictors and measures of the same construct.

Professional judgment Evaluations and decisions that are informed by and representative of the profession's commonly accepted empirical, methodological, and experiential knowledge base.

Psychometric Pertaining to the measurement of psychological characteristics such as aptitudes, personality traits, achievement, skill, and knowledge.

Regression equation An algebraic equation used to predict criterion performance from predictor scores.

Relevance The extent to which a criterion measure reflects important job performance dimensions or behaviours.

Replication A repetition of a research study to investigate the generality or stability of results.

Reliability The degree to which scores for a group of assessees are consistent over one or more potential sources of error (e.g. time, raters, items, conditions of measurement, etc.) in the application of a measurement procedure.

Reliability estimate An indicator that reflects the degree to which scores are free of measurement error.

Response process A component, usually hypothetical, of a cognitive account of some behaviour, such as making an item response.

Restriction of range or variability Reduction in the observed score variance of a sample, compared to the variance of an entire population, as a consequence of constraints on the process of sampling.

Sample A selection of a specified number of entities called sampling units (test takers, items, etc.) from a large specified set of possible entities, called the population. A random sample is a selection according to a random process, with the selection of each entity in no way dependent on the selection of other entities. A stratified random sample is a set of random samples, each of a specified size, from several different sets, which are viewed as strata of the population.

Sampling bias The extent to which a sampling process introduces systematic misrepresentation of the intended population.

Score A number describing the assessment of an individual; a generic term applied for convenience to such diverse kinds of measurements as tests, production counts, absence records, course grades, ratings or other selection procedures or criterion measures.

Selection procedure An assessment instrument used to inform a personnel decision such as hiring, promotion or placement.

Selection procedure (test) user The individual(s) or organisation that selects, administers, and scores selection procedures (tests) and usually interprets scores that are obtained for a specified purpose in a defined organisational context.

Shrinkage Refers to the fact that a predictor equation based on a first sample will tend not to fit a second so well.

Shrinkage formula An adjustment to the multiple correlation coefficient for the fact that the beta weights in a prediction equation cannot be expected to fit a second sample as well as the original.

Skill Level of proficiency on a specific task or group of tasks.

Standard deviation A statistic used to describe the variability within a set of measurements.

Standard error of estimate Standard deviation of errors of prediction that is associated with using a regression equation to predict Y, given X.

Standard error of measurement Standard deviation of errors of measurement, that is, differences between true and obtained scores.

Standard score A derived score resulting in a distribution of scores for a specified population with specified values for the mean and standard deviation. The term is sometimes used to describe a distribution with a mean of 0.0 and a standard deviation of 1.0.

Standardisation (a) In test construction, the development of scoring norms or protocols based on the test performance of a sample of individuals selected to be representative of the candidates who will take the test for some defined use; (b) in selection procedure administration, the uniform administration and scoring of a selection procedure in a manner that is the same for all candidates.

Standardised predictor A test employed for estimating a criterion of job performance, the test having been developed and normative information produced according to professionally prescribed methods as described in standard reference works.

Statistical control A procedure for removing or attenuating a source of error or bias by mathematically suppressing its influence.

Statistical power See *Power*.

Statistical significance The finding that empirical data are inconsistent with a null hypothesis at some specified probability level.

Subject matter experts Individuals who have thorough knowledge of the work behaviours, activities, or responsibilities of job incumbents and the KSAOs needed for effective performance on the job.

Synthetic validity evidence Generalised evidence of validity based on previous demonstration of the validity of inferences from scores on the selection procedure or battery with respect to one or more domains of work (job components); also referred to as "job component validity evidence."

Systematic error A consistent score component (often observed indirectly), not related to the intended construct of measurement.

Test A measure or procedure in which a sample of an examinee's behaviour in a specified domain is obtained, evaluated, and scored using a standardised process.

Test development Process through which a test or other predictor is planned, constructed, evaluated, and modified, including consideration of content, format, administration, scoring, item properties, scaling, and technical quality for its intended purpose.

Trait An enduring characteristic of a person that is common to a number of that person's activities.

Transportability A strategy for generalising evidence of validity in which demonstration of important similarities between different work settings is used to infer that validation evidence for a selection procedure accumulated in one work setting generalises to another work setting.

True validity The validity coefficient calculated after the necessary correction have been made to account for statistical artefacts.

Type I and Type II errors Errors in hypothesis testing; a Type I error involves concluding that a significant relationship exists when it does not (rejecting H_0 and concluding that a phenomenon is true in the population when it is not); a Type II error involves concluding that no significant relationship exists when it does (accepting H_0 and concluding a phenomenon is not true in the population when it is).

Utility The practical usefulness of an assessment instrument that allows the user to make better hiring decisions, save money, improve efficiency etc.

Validation The process by which evidence of validity is gathered, analysed, and summarised. (Note: laypersons often misinterpret the term as if it implied giving a stamp of approval; the result of the research might be zero validity.)

Validity The degree to which accumulated evidence and theory support specific interpretations of scores from a selection procedure entailed by the proposed uses of that selection procedure.

Validity argument An explicit scientific rationale for the conclusion that accumulated evidence and theory support the proposed interpretation(s) of selection procedure scores entailed by the proposed uses.

Validity coefficient A measured coefficient reflecting the relationship between a selection procedure and a criterion that provides evidence about the validity of the selection variable.

Validity evidence Any research or theoretical evidence that pertains to the interpretation of predictor scores, or the rationale for the relevance of the interpretations, to the proposed use.

Validity generalisation Justification for the use of a selection procedure or battery in a new setting without conducting a local validation research study. See *Generalised evidence of validity.*

Variability the spread or scatter of scores.

Variable A quantity that may take on any one of a specified set of values.

Variance A measure of variability; the square of the standard deviation.