The Effective Use of Psychometric Assessments in Decision Making

JOHN EATWELL & IAN WILSON

In W.W. Waitoki, J.S. Feather, N.R. Robertson, & J.J. Rucklidge (Eds.). Professional Practice of Psychology in Aotearoa New Zealand. 3rd edition. Wellington, NZ: New Zealand Psychology Society.

Overview

This chapter aims to provide standards and address practice issues for all areas of psychology, and references relevant New Zealand research from all fields to support the standards. However, given the authors' backgrounds are both in Organisational Psychology the examples provided are typically from this area. The New Zealand Psychologists Board guidelines *The Use of Psychometrics Tests* (2015) reference the standards outlined in this chapter, and provide more clinical and educational examples. Therefore, it is recommended that practitioners from clinical and educational backgrounds reference both sets of guidelines.

Introduction

As scientists in the field of psychology we seek to understand and predict human behaviour and to use this information to make decisions and guide future actions. In clinical, industrial and organisational, and educational settings psychologists seek to understand the underlying cause, or likelihood, of behaviour based on evidence gathered from observational, interview, and assessment data.

Research into the process of building this understanding – making attributions, building theories or schema – of people's behaviours shows the process is not rational or logical, even with professionals involved (Meehl, 1954; Sawyer, 1966). Attributions are seldom made using the full range of available data as Heider (1958) or Kelly (1967) would have liked to believe. Difficulties include:

- characterising information on the basis of pre-existing theories (Jeng, 2006; Nisbett & Ross, 1980) and seeking information that confirms rather than disconfirms theories (Oswald & Grosjean, 2004);

- extreme examples overly influencing judgements (Rothbart, Fulero, Jensen, Howard, & Birrell, 1978; Schwarz et al. 1991);
- being unaware of the effects of small (Nisbett & Ross, 1980) or unrepresentative samples (Hamill, Wilson, & Nisbett, 1980; Kahneman, 2000);
- underutilising base rate information (Baron, 1994; Hamill et al., 1980).

Of all the problems in building our judgements about people the last point is arguably the easiest to control. Through using standardised assessment techniques (after evaluation of the psychometric qualities of the tools) we can make judgements confident that the results are going to give a measure of the consistency of an individual's behaviour (based on the reliability of the tool and validity of its predictions) and its uniqueness (based on the comparison of the result to appropriate groups).

Less structured information is, of course, still very useful in making decisions. However, base rate data should be an important element of our understanding, given the potential that as psychologists we too can fall into the trap of weighting anecdotal (such as people's description of events) information above more valid base rate data (Ginosar & Trope, 1980; Hamill et al., 1980; Taylor & Thompson, 1982). Standardised assessment provides a tool to compare a standardised sample of behaviour for an individual with a large sample of that behaviour – the base rate – to give a measure of relative propensity. The validity coefficient, as an index of the relationship between our sample of behaviour and the behaviour we are trying to predict, gives us the justification for this comparison.

These standards for practice take on special significance given the Treaty of Waitangi, and the Anglo-Saxon origin of testing and the extent to which assessment is used in New Zealand (Taylor, Keelty, & McDonnell, 2002). Although the origins of testing are European, Palmer (2005) argues that the measurement of skills and attributes is central to classical Māori culture. She advocates embracing psychometrics as a mechanism to measure these skills and attributes more

effectively to support a number of policy, health, and evaluation initiatives. Although it is widely accepted that well-constructed psychometric assessments provide standardised information about a candidate and have been shown in general to lead to better and fairer employment decisions, international research has found differences in scores, or the relationship between scores and job performance, which can lead to selection decisions that impact negatively on some cultural or gender groups.

It is pleasing to see a number of published studies now addressing whether tests or questionnaires are resulting in different (potentially discriminating) results for Māori (Barker-Collo, Bartle, Clarke, van Toledo, Vykopal and Willets, 2008; Fernando, Chard, Butcher & McKay, 2003; Guenole, Englert, & Taylor, 2003; Haitana, Pitama, & Rucklidge, 2010; Lichtwark, Starkey, & Barker-Collo, 2013; Sibley & Pirie, 2013; Starkey & Halliday, 2011). However, there are still a number of assessment localisation reviews being published that either do not include this important consideration or have such low sample sizes as to not provide any useful insights.

The highest standards of practice in the use of all psychometric tools are needed to maximise the benefit to organisations and the individuals assessed, to minimise risk to organisations and individuals, and to promote fairness and equality of opportunity for all. To facilitate this, the following guidelines have been produced for practitioners who use psychometric assessment instruments. The guidelines aim to cover:

- when to use and who should use psychometric tools;
- the principles involved in choosing appropriate tools;
- the preparation of clients and administration of tests;
- the use of comparison groups in the interpretation of results;
- the issues surrounding confidentiality and storage of materials and results.

Practitioners working in organisations that are International Standards Organisation (ISO) accredited should consider encouraging their employers to adopt the standard for psychological assessment (ISO10667) (Bartrum, 2013). Adopting the standard will

provide organisational support for practitioners in good practice when using psychometric tools.

When Should Assessment Instruments Be Used?

Assessment can be used in a number of situations where a fuller understanding of an individual's performance, preferences, or behaviours is needed. The key to effectively using psychometrics is to have a clear question to answer or criterion in mind to measure. In occupational settings the criteria can be competencies (or other constructs identified as being required for effective job performance) used in selection, placement or promotional decisions, or development. In diagnostic situations, such as clinical, educational or career or outplacement counselling, the criteria are typically from pre-existing frameworks or practitioner driven hypotheses. Having a pre-existing hypothesis (or competencies) focuses attention on the relevant information produced by the assessment and minimises the chance of extraneous information impacting on decisions. For example, in Robertson and Kinder's study (1993) people making selection decisions based on a whole personality profile were no more accurate than they would have been with much less reliable and valid data – for example, an unstructured interview. It was only when they narrowed what they were looking at in the profile and based their decisions on key criteria that their decisions improved. This can also be a problem with computer generated reports which provide ratings across a number of competencies or criteria. The practitioner needs to focus the report onto relevant criteria before end users sees it, otherwise decisions will be made in a similar way to that found by Robertson and Kinder. In each case a decision must be made as to whether using assessment instruments would be appropriate to help achieve the desired objectives.

Using a single assessment result alone should be avoided whenever possible. Confirmation of results through other information gathering techniques minimises the impact measurement error may have on our decisions.

On some occasions it is not appropriate to use psychometric assessment. For instance, it is inappropriate to use such tools for choosing which individuals are retained or released in a redundancy situation, since assessments are only an indicator of future potential and direct information on job performance should already be available. (However, such tools can be valuable in making re-deployment decisions where there are significantly different new roles available or in outplacement counselling.) The use of psychometrics in redundancy (rather than redeployment situations) has been successfully challenged in New Zealand Courts (Gilbert vs Transfield, 2013). Similarly, an organisation may want to restrict the use of some assessments to counselling or development applications.

Who Should Use Psychometric Instruments?

Knowledge and experience are required to use psychometric tools effectively. It is recognised throughout the world that the use of psychometric instruments by unsuitably qualified individuals can have very detrimental effects. To access materials requires both formal training in their use and, often, specialist training for the specific instruments in question. Qualified users should ensure that materials are only used appropriately and are not used by untrained people or for a purpose for which they were not intended. It is also their responsibility to work within the confines of their own expertise and to recognise when refresher training, skills updates or expert advice is needed.

Using Psychometric Instruments

Choice of Instrument

Strauss, Leathem, Humpries, and Podd (2012) found in their survey of New Zealand practitioners that availability of assessments was the primary criteria on which people were choosing and using tests, even when they knew the particular test they were using was inadequate. One cannot judge the quality of an instrument solely by how widely it is used, as instruments can become out-dated and, as Strauss et al. found, organisations use inappropriate and out-dated assessment tools.

It is important that whenever instruments are chosen there is written documentation of the reasoning behind the choice. This documentation may include items such as copies of research studies, hypotheses being tested, job analysis reports, job descriptions, person specifications, validation studies. If the relevance of the particular measure is challenged, such evidence supports the instrument choice, shows the care taken and helps ensure users do not take inappropriate short cuts. In occupational settings, the importance of job analysis in determining the knowledge, skills and abilities required for effective on the job performance is magnified by the Human Rights Act 1993, which makes it unlawful to make a selection decision on the basis of information that may be discriminatory or is not relevant to the job.

There are four main aspects to consider when choosing a psychometric instrument: The content of the instrument; the level the instrument is aimed at; its psychometric qualities (Cronbach, 1984); and any potential adverse impact.

Instrument Content

Whenever psychometric assessments are used it is vital that there is a match between the skills and characteristics measured and the job and organisational demands. This is reinforced in legislation and the Code of Ethics for Psychologists Working in Aotearoa New Zealand 2002 (the Code of Ethics):

- The Pre-employment Guidelines (Human Rights Commission, 2007) based on the Human Rights Act 1993, Section 3, states that employers should only request information that is clearly relevant to the requirements of the job an applicant is applying for or being developed into, and their ability to do the job. For instance, the instrument should not require understanding of complex vocabulary or performance at speed if these are not required on the job.
- The Privacy Act (2003), Principal 4, states that the manner of collection of personal Information should not be by means that are unfair or intrude unreasonably on the personal affairs of the individual concerned. For example, questions about personal preferences of taking baths or showers may not be intrusive in a clinical or education setting but would be viewed as irrelevant in an occupational setting.
- In addition to the legal requirements, the Code of Ethics stipulate that unnecessarily intrusive questions – for example, a personality questionnaire that asks if people have a “great fear of snakes” or “wish they were not bothered by thoughts about sex” to measure Tolerance and Social Presence respectively should not be asked if there is a less intrusive way of equal validity that could be used. The question has to be asked, is there another way of assessing these factors of equal validity without a person having to divulge their personal information?

In an occupational setting *Job Analysis* is the best way to determine the skills and attributes required for a particular job which can then be matched to appropriate assessment tools. The more detailed the analysis of the job and the closer the match between the attribute required on the job and that measured by the instrument, the higher the *content validity* of the tool. The content can be in terms of both the structure of the test (e.g., multi choice or written, whether it is given orally or it has to be read) as well as the substance of the actual items (e.g., the use of text from an operating manual to assess verbal comprehension in machine operators). Ultimately we want the content to be similar enough to predict future behaviour, but also generalisable enough so that it does not discriminate against those who have had less opportunity to work in the area.

The second aspect of instrument content to be aware of when matching assessment materials to jobs is the context in which the skill is measured. This should, as far as possible, reflect the type of content found in the job. For example, a typing test should require the typing of material similar to that required on the job and the numerical test should involve tasks similar to those to be completed on the job (do they have to complete number sequences in the test or interpret graphs and charts?) However, care must be taken not to include material requiring knowledge specific to the organisation that would, for instance, put external applicants at an unfair disadvantage. For instance, in the typing test, if a job relevant text is too technical for an external applicant to deal with before training, more general tests should be used. Content of an instrument that is of a more general nature should be equally accessible to all applicant groups: Men, women, Māori, and other cultural or ethnic groups.

Content Validity

- Content related to context (e.g., occupational settings).
- Format related to the way tasks or work is completed.
- Questions not unnecessarily intrusive.
- Comply with legislation.

Instrument Level

The level of difficulty at which the skill or attribute is measured should be appropriate to the job. An instrument that is too easy or difficult will not differentiate between individuals. In recruitment or development the level of the

instrument used should also be appropriate for the type of work being completed, or which will be completed in the future. If the general level of applicants is below standard required for the job, employers should consider what they can do to attract better applicants rather than changing the level of the tool. Training or job redesign options also need to be considered. If there is a tendency for individuals from one particular ethnic group or gender to fail to meet the required standard, Section 73 of the Human Rights Act 1993 sometimes allows positive action targeted at this group in order that they may "achieve an equal place with other members of the community". This may take the form of special training programmes, for example.

Psychometric Qualities

Assessment instruments should be psychometrically sound. The relevant information and statistics for judging the instruments should appear in the user manual. These should include:

- Information about the purpose and development of the test
- Specification of the skill or attribute the instrument measures – we need to know that the instrument we are using is actually assessing the type and level of skill or attribute required on the job.
- Description of groups for which the instrument is appropriate (educational background, work experience, etc.) – we need to ensure that the instruments we use are appropriate for the individuals being assessed. This can be done by referring to the biographical details of the comparison group chosen to benchmark against. For example, we would not give applicants to a clerical position a verbal reasoning test designed for university graduates.
- Details of the development process – it is important to find out whether the instruments we use have been developed in a rigorous manner. All instruments should be developed on the basis of a thorough job analysis to ensure that they are assessing skills, knowledge or abilities necessary for effective on-the-job performance in the occupational area for which they are designed.

- Information on the reliability, validity and norm groups of the instrument – given their importance, these are dealt with in more detail in the following sections.

Reliability

Reliability is important for two reasons. Firstly, it sets an upper limit on how valid the test can be as it tells us how much measurement error there is in the test. The more error in the test, the less accurate will be the measurement of what we are trying to understand (job performance in an occupational setting). Secondly, reliability gives us a margin of error for the interpretation of an individual's test score. This margin of error allows us to make accurate judgements about real differences between people based on their test score or answers to a questionnaire.

Reliability	
0.70	Acceptable
0.75	Good

Reliability is measured by a correlation coefficient, with 0.75 giving a margin of error of 0.5 a standard deviation around our test score, which is the equivalent of one Standard Ten (STEN) or 5 Transformed Scores (T Scores – see section under norming for more information on STENS and T Scores). This gives us a 68% probability that if we tested the individual again, their true performance would lie within 0.5 a standard deviation of their actual score.

Classical Test Reliability

In traditional paper and pencil tests, reliability is derived in three main ways:

1. *Internal consistency.* Having good internal consistency or split-half reliability is a start in understanding how consistent the responses have been throughout the questionnaire. Consistency of responses gives an indication of whether the items are actually measuring the same concept. Low consistency could be intentional and caused by having broad scales (e.g., a number of different concepts are being measured within the scale) or because the items are ambiguous and the candidate is interpreting items related to the same

construct differently as they go through the questionnaire. Good examples of the assessment and verification of internal consistency in New Zealand samples (Knight, McMahon, Skeaff, & Green 2008; Krynen, Osborne, Duck, Houkamau, & Sibley, 2013; Sibley & Pirie, 2013; Wright, Burt, & Strongman, 2006).

Although internal consistency gives an indication of the reliability of a test or questionnaire, it does not account for a number of other potential sources of error in testing – for example, temporary states within the individual, variations in the administration instructions, variations in scoring, or the conditions in which administration occurred. Therefore, it is a minimum standard for test development rather than the gold standard.

2. *Test-retest reliability* is likely to be the best indicator of the potential error in how tests are actually used. Because it is derived from two separate sittings of the test across time, it captures all potential sources of error – including administration, scoring, interpretation of the items, temporary states, and the content of the test. See Wright, Burt and Strongman (2006) for a New Zealand example.

3. *Parallel-forms reliability* captures some of the potential errors in tests, such as administration instructions and item clarity; however, as parallel form testing is usually done at the same point in time it does not allow estimates of change over time.

Item Response Theory

With tests developed using item response theory (IRT) some of the traditional classical test theory methods of calculating reliability can still apply (test-retest and parallel-forms). However, in IRT itself, the notion of reliability is more complex. There is not the scope to address it fully here; however, for a more comprehensive account see, for example, Embretson & Reise (2000).

In IRT, reliability is described in the form of "information" (the more information there is, the higher the accuracy). Because IRT focuses on the psychometric properties at the item level (rather than the overall test level as with classical test theory), the information is calculated first at the item level (in the form of the information function, or IF), and then it can be summed to give an overall information level across a test (the test information function or TIF).

Another difference with IRT is that the reliability is not constant across the item or test, but varies according to a test taker's ability or trait level (called theta or the Greek symbol θ). In other words, the amount of information an item provides differs according to the proficiency of the person completing it. A very difficult item, for example, may tell us a lot about high ability performers, but very little about those with low ability (as they will all get it wrong).

Of course, all this complexity makes evaluations more difficult for practitioners unfamiliar with the concepts. *Marginal reliability* (Green, Bock, Humphreys, Linn & Reckase, 1984) presents an example of a measure developed to provide a single reliability index for a test, which can help with interpretation and comparisons with tests developed using classical test theory. Essentially it is a calculation of the average reliability across the test and different levels of theta, and interpretation is similar to the correlation coefficient used for classical test theory reliability (Thissen & Wainer, 2001).

There are New Zealand examples of studies using IRT approaches to testing, and consequently to estimating reliability (Krynen et al., 2013; Sibley & Pirie, 2013).

Validity

Validity is crucial to both the usability of the test with candidates, establishing whether it is measuring what we want to measure, and giving us a picture of how

certain we can be about the results. All forms of validity are important, although criterion validity is the ultimate judge of whether tests should or should not be used, as it is the measure of whether we can predict future behaviour – which is ultimately why we use tests!

Face validity. Face validity is a qualitative assessment, gauging whether the test looks relevant to the context in which it is being used. For example, in an occupational setting are the types of questions being asked likely to be things the candidate has to deal with on the job, or are they personal in nature? The item "I hate opera singing" has been questioned in the media (Duff, 2013). Questions of a personal nature may be relevant in a clinical setting but would lack face validity in an occupational setting (see the section above on instrument content). Questions the candidate could see as discriminatory should not be asked, even if they are not being used in that way. This is particularly important when selection or promotion decisions are based on assessment results.

Face Validity

- Look relevant
- Comply with legislation
- Not unnecessarily intrusive

Construct Validity. Four distinct questions relate to construct validity: Is there a good theoretical reason why the construct should predict future behaviour; is the test delivering similar results to other tests also measuring that construct; is the test really measuring everything it says it is measuring; and is the description of the construct reflective of what it is measuring.

Construct Validity

- Based on a sound theory
- Correlate 0.6 with tests measuring the same thing.
- Is measuring what it says it is measuring

- There should be a good theoretical reason as to why the construct is going to be related to job performance. Without a good understanding of the relationship, we cannot fully manage any adverse impact or other contamination factors that may occur. This factor has been emphasised by the International Test Commission to ensure tests are cross-culturally appropriate. For example, a question on whether a person preferred baths to

showers being positively correlated with other behaviour is not enough to justify its use, unless there is a good theoretical reason why it relates to that other behaviour. We would also need to be assured that it related to the behaviour in different cultural contexts, where baths or showers may be less or more prevalent.

- The test, and the subscales within it, should correlate well with other tests or scales measuring similar constructs. Too high a correlation (0.8: Morrow, 1983) asks what additional value this measure has over existing tools. The benchmark tests must themselves have proven validity and reliability, otherwise the comparison is meaningless. The European Federation of Psychological Associations has set a median correlation coefficient of 0.55 as acceptable and 0.6 as good. For such analyses sample sizes should be greater than 100. Furthermore, the test or scales should not show a strong relationship with other unrelated constructs, as that would suggest measurement error is occurring. New Zealand has good examples of construct validity being tested with clinical, education, and occupational samples (Barker-Collo et al., 2008; Christianson & Leathem, 2004; Palmer, 2004)

- Does the test actually measure the number of constructs it claims to? Factor analysis can be used to verify whether the test is independently measuring the number of constructs that it says it is measuring. For example, if the test claims to measure 12 distinct scales, but factor analysis reveals only four factors, the claim is not supported. There are some nice examples of the work that should be done to validate the structure of an international questionnaire in New Zealand (Brown, Jose, Ng & Guo, 2002; Guenole & Chernyshenko, 2005; Knight, et al., 2008; Roberts & Wilson, 2008; Sibley & Pirie, 2013; Wicks, Siegert, & Walkey, 2004; Wright, Burt, & Strongman, 2006).

- Qualitatively, the definition of the construct should match what it is measuring. For example, high reliability can be achieved by having a very narrow range of questions, or even by repeating the same question a number of times. This is okay if the definition of the construct is similarly narrow. However, if the construct that the test claims to measure is much broader than the content of the questions, this is misleading. The type of scoring should also be addressed; for example, if the test gives marks for approximate answers it needs to be called a test of estimation, not of reasoning.

Criterion Validity. For assessments that are intended to be used in employment contexts, criterion-related validity is essential. Assessments that have no proven relationship to job performance or behaviour add little or no value to decision making processes. The Human Rights Act reinforces this importance, requiring that employment decisions must be based on job-relevant criteria. Criterion-related validity helps establish this relevance. Without proven criterion validity, a test is unlikely to be legally defensible and should not be used in employment settings.

Criterion Validity

- Total sample size more than 100
- Two or more studies
- Statistically significant ($p<.05$).
- Relationship of 0.3

Criterion-related validity coefficients should be statistically significant and be carried out on an appropriately sized sample based on the validity of the assessment. A total sample of at least 100 people, preferably in the form of two groups of 50 people each (in order to verify findings), is a good standard to use. This obviously may be difficult to achieve in a New Zealand setting given the average size of organisations, although some good examples are coming through now where people have managed to achieve this (Black, 2000; Hambleton, Kalliath, & Taylor, 2000). While the size of the coefficient is important (the higher the better), 0.20 has been set by the European Federation of Psychological Associations as a minimum acceptable level. Detecting values below that require large samples, but can also be useful in recruitment where, for instance, a very small selection ratio is being used.

As a rule of thumb, however, coefficients of 0.3 or greater are desirable as this equates to an 80% success rate in the prediction (based on a selection ratio of 1:20). That is, 80% of those who score above average on the test will be rated as above average on what you are trying to predict (Taylor & Russell, 1939).

Further considerations include the relevance of the criterion measure to the intended test use (direct measures of outcomes or behaviour are ideal), and the relevance of the sample (e.g., job incumbents for occupational tests, students for educational tests).

Adverse Impact.

The issue of adverse impact takes on special significance given the Treaty of Waitangi, and the Anglo-Saxon origin of testing, as discussed earlier. A problem arises when a meaningful difference is found between the average performance of different ethnic or cultural groups, or men and women on the assessment. A test or measure should not produce large (greater than one standard error of measurement) systematic differences between different groups of people (on the basis of age, gender, or ethnicity). This is especially common where socio-economic conditions impact on the educational opportunities available to particular groups, or where a candidate is not a native speaker of the language in which the assessment tool is presented, although the cause may be as simple as the perceived threat around the testing process itself (Brown & Day, 2006). 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 (Hunter, Schmidt, & Hunter, 1979). That is, if an unjustifiable entry standard is set and demanded of all applicants, the lower scoring group would find it harder to comply with the requirement and, hence, would be indirectly discriminated against.

Adverse Impact

- Information is available.
- Cutoff scores are set.
- Differences should be less than 1 standard error of measurement.
- More than four fifths of a group included.
- Practice materials are available.
- Tests measuring elements of crystallised intelligence, e.g., word knowledge are treated with extra caution.

Validation evidence showing that those who perform poorly on the assessment also perform poorly on the job confirms that rejecting low scoring candidates is reasonable, as they are not being unfairly discriminated against. However, where the validation evidence shows that individuals are not performing worse on the job but are scoring lower on the test or questionnaire, disparate impact is occurring (differential prediction). The greater the degree of disparate impact resulting from the use of a psychometric instrument, the higher the validity should be to justify its use. However, this would need to be weighed against diversity goals that an organisation may have and whether a better instrument can be found.

The possibility remains that overall validity is masking cases where an instrument has poorer or no predictive validity for some groups, or that group differences in assessment scores are not reflected in job performance (differential validity). For example, albeit with a wholly inadequate sample, Barker-Collo et al. (2008) found two measures that correlated strongly for Pakeha had no interrelationship for Māori. If one of these measures was significantly related to clinical outcomes for one group it would have no relationship for the other. Extensive research into these issues 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, when best practice has been followed (Hartigan & Widgor, 1989; Hunter & Hunter, 1984). However, care needs to be taken that performance measures are not also contaminated (Roth, Huffcutt, & Bobko, 2003).

Some experts argue that where cultural or gender group differences on assessment scores exceed group differences in job performance, separate norm tables for each group should be used for evaluating scores (Department of Labour & The Office of Personnel Management, 2011). Use of separate norms in these circumstances has not been tested in New Zealand courts or tribunals, but it would not be justifiable based on differences in test performance alone. The availability of direct or relevant instrument validation data means that discriminatory practices can be avoided.

If group differences have not been researched for a test, questionnaire, or selection process, the following steps should be taken:

- Put pressure on test publishes to sponsor such research or collect user data that will lead to a better understanding of any group differences in test scores.
- Test in the candidate's first language where possible.
- Thoroughly analyse the instrument's content to ensure that questions are free from sexist language and are equally meaningful to Māori and other ethnic groups to whom they may be administered. For example, normative questionnaires containing double negatives, which are not part of Pacific languages. may impact on the accuracy of those measures for Pacific peoples. Questionnaires designed to be non-transparent to avoid distortion often contain significant cultural context that would need to be validated before they could be used outside of the countries they were developed in.
- Set a test cut-off that relates to performance requirements rather than taking the best scoring candidates. Taking the best scoring candidates will impact the most on lower scoring groups.
- Monitor the proportion of candidates who are meeting criteria. Where the proportion of an ethnic or gender group passing is less than four fifths of the proportion of other groups passing, adverse impact is likely to be occurring (Singer, 1993).
- Ensure that the standard error of measurement is taken into consideration in score interpretation and in making decisions between candidates, as well as in the use of cut-offs.
- Use practice materials or practice sessions with candidates to reduce the perceived "threat" or unfamiliarity of the actual testing sessions. Likewise, avoid terminology that could increase anxiety about the session such as "test". "Assessment" and other softer terminology can reduce or eliminate the impact.

It should be remembered that group differences relate to average performance. Even where substantial group differences exist there will be members of the lower scoring group who have better results than many people from the higher scoring group and vice versa. Furthermore, job success does not generally depend on a single ability or preferred style, and assessment tools do not have perfect predictive power. Therefore, on occasion, those with poorer results on an assessment will do better in a job than an assessment result may suggest. For this reason it is preferable to interpret assessment results along with other available information.

If differential scoring patterns are used with different gender or ethnic groups, it is particularly important that appropriate guidelines are followed to avoid improper use of psychometric tools. Considerations of fairness are important in themselves, particularly when the legal implications under the Human Rights Act 1993 for engaging in discriminatory practices in the selection and promotion of employees are taken into account.

While access to internet is continually improving, in New Zealand one in five still do not have internet available in their home. For Māori (32%) and Pacific Islanders (35%) this number is even higher (Statistics New Zealand, 2013). Therefore, to reduce the risk of adverse impact where assessments are delivered to candidates via the internet, efforts should be made to ensure that candidates have appropriate access to internet facilities, or that alternative provisions can be made for candidates without access. In addition, practitioners should ensure candidates are comfortable using the technology, and that there is appropriate user support available.

Identifying whether people have disabilities that may impact on the assessment result is also an important consideration. This can be achieved by notifying candidates of selection processes well in advance, supplying practice materials to enable them to see what the process will entail, and asking if there is anything that may impact on the accuracy of the assessment. If candidates do have disabilities, consult with your test publisher on the best way to manage the assessment process.

Representative Norm Groups

The last factor to consider is the availability of representative and relevant comparison groups. This will be covered further below.

Norms

As we compare a person's aptitudes and attributes in relation to other people, scores on tests and questionnaires also need to be compared to relevant comparison groups. We do this through seeing how a person's score sits in relation to others' scores on a normal distribution or bell curve. Norms are sets of data derived from groups of individuals who have already completed a test or questionnaire. These norm groups enable us to establish where an individual's score lies on a standard scale, by comparing that score with that of other people.

Within the field of occupational testing, a wide variety of individuals are assessed for a broad range of different jobs. Clearly, people vary markedly in their abilities and qualities, and therefore the norm group against which an individual is compared crucially important. Given our original purpose of using tests to gather valid base rate information on behaviours, we need to make sure the comparison is similar to the group we are trying to make predictions about. It is very likely that the conclusions reached will vary considerably when an individual is compared against two different groups – for instance, school leavers and managers in industry. Therefore, it is important to ensure that the norm groups used are relevant to the given group or situation that the data are being used for.

Norming Systems

A number of different norming systems are available for use, each of which have strengths and weaknesses in different situations. These can be grouped into two main categories: Rank order and standard score systems.

Rank Order (Ordinal scale) Systems. When a group of people are given a test or questionnaire we expect to observe a range of different scores as people differ in

their abilities or personal qualities. This spread of results allows us to arrange people in a rank order scale according to their performance. As percentiles are a description of a rank order they have the disadvantage that they are not equal units of measurement. Accordingly, percentiles cannot be averaged or treated in any other mathematical fashion.

However, they have the advantage of being easily understood and can be very useful when giving feedback of results to candidates or discussing results with managers. They also represent the true distribution of scores for a test or questionnaire. If the distribution of scores for a particular test is not "normal" then the percentile interpretation will be more reflective of real differences in performance.

Standard score (Interval scale) Systems. To overcome the problems of interpretation implicit with rank order systems, various types of standard scores have been developed. All systems describe how far away the candidate's result is from the average in terms of standard deviations. Systems include T-Scores, Sten scores, standard nine scores (Stanine) and intelligence quotients (IQ).

Standard score systems have the advantage of being able to be mathematically manipulated (added together and weighted) and assume a normal distribution.

Choosing Norm Groups

Using New Zealand-based norms is important for face validity and construct validity. When providing feedback to a respondent, giving the information in the context of a group they feel they should be compared with is very important for the person's acceptance of the results. Secondly, the results may be less meaningful when compared to an inappropriate group, as the reference group may have different means or smaller distributions than the New Zealand group, thereby masking or accentuating the differences being explored.

International reference groups may be applicable for multinational companies where selection decisions involve applicants from a number of countries and the position involves working internationally.

Norm group size should be determined by the standard deviation of the sample and the reliability of the instrument. A rough rule of thumb is that for instruments that meet the gold standard of reliability (0.75), norm groups should be made up of at least 150 people, preferably 300 or more, and should be directly relevant to the purpose of the test (Evers et al., 2013).

As discussed previously, the purpose of norms is to provide base rate information about the likelihood of a skill or behaviour being displayed by the individual. For base rate information to be meaningful the characteristics of the comparison group must be as similar as possible to the reference group of our audience or the situation in which we are trying to predict the behaviour. A person applying for a role in senior management needs to have their assessment results compared to a senior management norm group to enable a better understanding what their behaviour is likely to be on the job compared with the base rate of that group.

In-house norms provide the most directly relevant base rate information, as the recipient of the information will be very familiar with the prevalence of the behaviour being measured. The disadvantage of norms produced in-house is that the comparability with the whole population is lost. For example, are the people in your own institution more or less able than the whole population; are your applicants more or less capable at what you are measuring than those applying or shortlisted by other companies?

A number of test suppliers (for example, CEB, NZCER, and Opra Consulting Group) in New Zealand supply relevant norm groups for use with their assessments. You should stipulate that you want relevant New Zealand norm groups when ordering products. There are also a number of published studies providing New Zealand normative sets for clinical, educational, and occupational groups (Fernando et al.,

2003; Knight, McMahon, Green, & Skeaff 2004; Knight et al, 2008; Wicks, Siegert, & Walkey, 2004; Sibley & Pirie, 2013).

Preparing Candidates and Administrating Assessment Tools

Standardisation is the fundamental concept in the utility of tests (Mischel, 1986). It is the standardised conditions, instructions, time, content, scoring, and interpretation that make psychometric tools (or any assessment tool or methodology) effective.

Use of Practice Materials

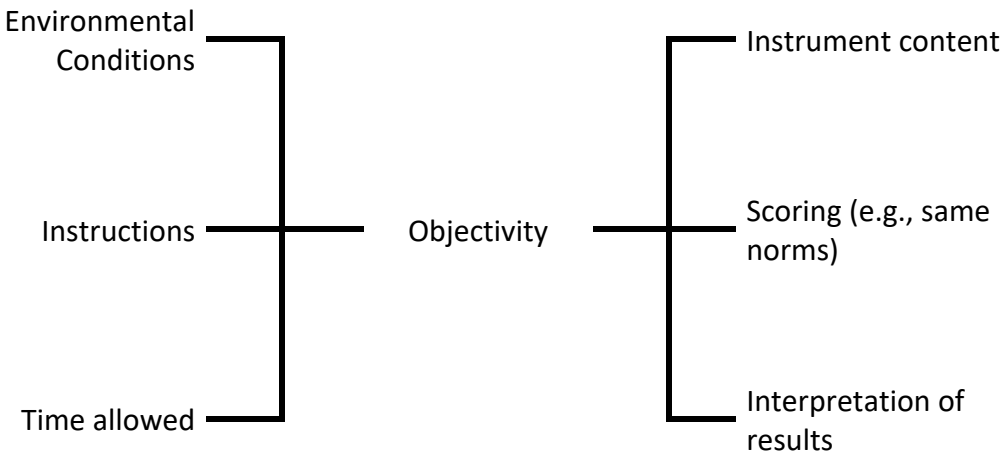
Some candidates may be unfamiliar with assessment processes so it may be difficult for them to perform at their best. Others may find the assessment situation very stressful. Ethnic minority candidates in particular may under-perform because of the effects of educational disadvantage or race discrimination. Older candidates and those with less educational experience are also likely to suffer from these sorts of problems. Practice items at the beginning of an assessment procedure can reduce the bias that may arise from differential assessment sophistication. They can also reduce nervousness by allowing a candidate to gain confidence in their ability to perform well in the assessment. People with disabilities can flag in advance potential issues that may arise during the testing process, allowing time for appropriate adjustments to be made. If possible, candidates should be notified a week in advance that they will be assessed. Examples or descriptions of what assessment instruments will be like (practice leaflets or website links) should be provided, so that candidates can familiarise themselves with the type of tasks involved (Kellett, Fletcher, Callen, & Geary, 1994). Such practice increases the effectiveness of the assessment proper by giving an accurate measure of a candidate's style or ability.

Administration of Psychometric Instruments:

The administration instructions are extremely important and must always be strictly adhered to. Only qualified persons should administer assessment tools. Abuse of procedures described in the instrument manual can lead to bias and possible

unlawful discrimination. Special care should be taken with people whose first language is not English to ensure they have understood the administration instructions properly. Some assessments that are fair for native English speakers will present problems for people with a lesser command of the English language. Instruments requiring reading skills that are not an integral part of the job are particularly likely to be unfair. Where appropriate, such candidates should be assessed in their native language. A number of aspects to instrument administration should be standardised. This is illustrated in Figure 1:

Figure 1: Standardisation



An encouraging attitude on the part of the administrator is always desirable, but it is particularly important to establish rapport with individuals who might lack confidence or who feel anxious about the assessment. The introduction to the assessment session is an important part of the administration procedure and instructions should be clear and not rushed. It allows the establishment of this rapport and should be conducted in a serious yet friendly manner. Information on the following should be provided during instrument administration:

- why instruments are being used and how they fit into the assessment procedure
- the intended recipients of the results or outcomes
- the name and address of the organisation responsible for conducting the assessments
- the name and address of the organisation who will hold the assessment results
- that completing the assessment is not compulsory
- the implications, if any, should the candidate choose not to complete the assessment
- that the candidate will receive meaningful feedback relating to their performance on all assessments conducted.

Candidates should have an opportunity to ask general questions before the formal assessment procedure starts.

Remote Administration

Although educational and clinical use of tests is still predominantly via paper and pencil tests, increasingly occupational assessment is delivered via the internet or even smartphones. When choosing which method of administration is appropriate, you should always follow the advice of the test publisher, and look at the way in which the instrument was designed to be used.

One key difference with the introduction of internet based testing is that it has opened up the possibility of remote and unsupervised (unproctored) assessment. Research to date has found that this mode of administration does not affect the psychometric properties of self-report personality instruments (e.g., Bartram, 2005; Bartram & Brown, 2004), and remote administration of such questionnaires is now a fairly well-established practice.

Typically with cognitive ability tests, where standardisation is critical to the reliability and validity of the results, a supervised mode of administration with a trained test administrator would always be recommended. Recent advancements, however, have seen the introduction of online ability tests designed specifically to be completed remotely. The integrity of the test content is preserved by a unique test being created for each test taker (items, calibrated for their difficulty level, are drawn randomly, or using adaptive testing methods[1], from large item banks). Such tools have the most efficacy in a screening situation – the tests can be taken by a large number of geographically dispersed applicants without the logistical difficulties of conducting supervised assessments. They also allow cognitive ability testing to be brought forward in a selection process, which is beneficial both practically and psychometrically (using more valid techniques early with high volume groups will bring through a higher yield of candidates to subsequent stages). Remote proctoring is another solution proposed by some providers, where candidates are supervised via webcams, for example.

Although the advantages of remote testing are substantial, its use does raise some interesting issues, especially in moderate and high-stakes situations (e.g., for recruitment and selection) where the temptation, and opportunity, to cheat is high. As the identity of the test taker cannot be fully verified, applicants who pass through the screening phase should ideally be retested under supervised conditions at a later stage. Some providers offer specific verification tests, designed to confirm whether scores are consistent with those achieved in the unsupervised conditions. In addition, having candidates agree to an honesty contract can act as a deterrent to discourage cheating in the first place. As with personality questionnaires, feedback interviews can further help verify results, and identify any issues the test taker may have experienced at the time of completion.

[1] Typically referred to as Computer Adaptive Tests (CATs), the difficulty of the questions is adapted to match the ability of the test taker. If questions are answered correctly the difficulty level will gradually increase and vice-versa.

Confidentiality and Storage

Materials

The security of materials is paramount. Free circulation leads to over-familiarity and devalues psychometric instruments. Responsible instrument publishers only supply materials to trained users who, in turn, must ensure untrained users do not gain access to them. Within an organisation, decisions should be taken about who should hold assessment materials and who should have access. It may not be desirable for all users to have access to all materials. Central storage can help prevent unnecessary duplication of materials but may not be practical in decentralised organisations. An organisation must supply instrument users with appropriate storage space where the materials can be kept under lock and key. It is highly desirable that all materials are logged in and out of storage when used. This helps ensure materials are not carelessly left lying around or misplaced. Failure to keep track of materials can be expensive where replacements have to be purchased or annual lease fees paid on missing booklets.

In July 2014 The International Test Commission released guidelines on the security of test materials with particular focus on the issues of security of materials and technology (*The ITC Guidelines on the Security of Tests, Examinations and Other Assessments Version 1*). It is suggested practitioners refer to these guidelines if needed.

The Role of Feedback

The Privacy Act 1993 requires that whenever assessment results are used assessors should be honest and open with candidates about why the instruments are being used and what will happen to the results. Members of the New Zealand Psychological Society are also bound by the Code to obtain the informed consent of the individuals to be assessed when undertaking a psychological assessment. Individuals must be informed of their right to know the content of psychological assessment reports, and in reporting findings psychologists must endeavour to

ensure that appropriate explanations of the findings and their interpretations are given. All individuals assessed should be offered meaningful feedback about their results as soon after the assessment process as possible. Personality and motivation questionnaire feedback is critical and will often enhance the interpreter's own understanding of assessment results.

Feedback does not need to be lengthy; indeed, with a large number of applicants this might be very time consuming. A face-to-face interview is preferred, but telephone feedback may be the only option in some circumstances. Feedback should be given by people trained and qualified in the assessment tooland should be an open two-way process. The NZPsS guidelines do not sanction the release of uninterpreted data from assessments to individuals untrained in their use and interpretation; *profile charts* may be shown to respondents but they should not be given copies to take away.

Computer generated or narrative reports can support, but should not replace, the feedback interview. Some may be suitable to give to respondents, but many are intended as aids to interpretation to the trained instrument user and could easily be misinterpreted by others. Users should follow guidelines provided by the author or publisher of such systems.

Under Information Privacy Principle 8 of the Privacy Act 1993, personal information gathered and held about an individual should not be used without ensuring that the information is accurate, up to date, complete, relevant, and not misleading. Feedback interviews are an important part of this validation process.

Assessment Results

Assessment results, like all personal information, should be stored with the strictest regard to confidentiality. Access should be restricted to those with a need to know and in accordance with what has been agreed with the respondent during administration and feedback. Persons who are untrained should not be allowed access to raw data from assessments, but only to clear interpretations of those results.

Individuals change and develop and so psychometric data can become less accurate over time. Therefore, Instrument scores should not be kept on file indefinitely. The time period for which scores are valid will differ depending on the measures and the particular use made of them. Care should be taken in using results that are more than 6 to 12 months old for selection purposes. Little reliance should be put on results over two years old for any purpose.

The Privacy Act requires that all personal information, including assessment results, should be protected by such security safeguards as it is reasonable to take to ensure against (a) loss and (b) unauthorised access, use, modification, or disclosure and should only be kept for as long as is legally useful. Given that decisions can be challenged for up to 90 days under the Employment Relations Act or up to one year under the Human Rights Act (unless there are exceptional circumstances), this would suggest keeping results for longer than two years is also unlikely to be defensible.

CONCLUSIONS

We generally understand people's behaviour in relation to models we build about the world. These normative theories encompass the concept of relativity, but even as professionals we underutilise important information in forming these attributions. Gathering and using base rate information is the easiest and arguably the most effective way to correct for these errors of judgement. Gathering standardised information and interpreting it by comparing it with a relevant population provides us with important base rate information, if the standardised information has a proven relationship with what we are trying to measure (validity). Norms need to be relevant to what we want to use the results for, for example work-related samples in work contexts and educational samples in educational contexts. New Zealand norms are an important aspect in the effective interpretation and understanding of test scores. Taking a contextual view, we also need to take account of different individual, social and cultural influences on test behaviour and outcomes.

Research and legislative precedent is firmly entrenching the importance of rigorous analysis to justify the use and choice of tests. Our ethical and professional obligations as psychologists are clearly laid out and oblige us to only use tools we are trained to use. Validity and reliability and appropriateness of the items should underlie our evaluation of tests.

Standardisation is the key principle behind the application of psychometric tools. As practitioners this does not always come easily to us, but being rigorous must be foremost in our minds.

Separate norms can be used for different gender or ethnic groups if it is proven that differential validity exists; that is, the relationship between test scores and behaviour is different for the groups concerned.

The internet provides innovative and efficient methods of testing, but also risks compromising some of the key benefits of assessment. The core principals outlined above apply equally to assessment over the internet.

References

Barker-Collo, S., Bartle, H., Clarke, A., van Toledo, A., Vykopal, H., & Willetts, A. (2008). Accuracy of the National Adult Reading Test and Spot the Word Estimates of Premorbid Intelligence in a non-clinical New Zealand sample. *New Zealand Journal of Psychology, 37*(3)*,* 53-61.

Baron, J. (1994). *Thinking and deciding* (2nd ed.). Cambridge: Cambridge University Press.

Bartram, D. (2005). The changing face of testing. *The Psychologist, 18*, 666–668.

Bartram, D., & Brown, A. (2004). Online testing. Mode of administration and the stability of OPQ32i scores. *International Journal of Selection and Assessment, 12*, 278–284.

Bartrum, D. (2013). ISO10667: An international standard for psychological assessment. *OP Matters*, *19*, 37–40.

Black, J. (2000). Personality testing and police selection: Utility of the 'big five'. *New Zealand Journal of Psychology*, *29*(1), 2–9.

Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: stereotype threat and the race gap on Raven's advanced progressive matrices. *Journal of Applied Psychology, 91*(4), 979–985.

Brown, J., Jose, P., Ng, S. H., & Guo, J. (2002). Psychometric properties of three scales of depression and well being in a mature New Zealand sample. *New Zealand Journal of Psychology, 31*, 39–46.

Christianson, M., & Leathem, J. M. (2004). Development and standardisation of the computerised finger tapping test: Comparison with other finger tapping instruments. *New Zealand Journal of Psychology, 33,* 44-49

Code of Ethics for Psychologists Working in Aotearoa/New Zealand (2002). Retrieved from www.psychology.org.nz.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.

Department of Labour & The Office of Personnel Management. (2011). *Uniform Guidelines on Employee Selection Procedures. 4, 29.* Retrived from www.gpo.gov/fdsys/pkg/CFR-2011-title29-vol4-part-1607.xml

Duff, M. (2013, September). Job personality tests may be illegal. *Stuff.* Fairfax Media.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mawah, New Jersey: Erlbaum.

Evers, A., Hagameister, C., Hostmaelingen, A., Lindley, P., Muniz, J., and Sjoberg, A. (2013). *EFPA review model for the description and evaluation of psychological and educational tests.* European Federation Psychologist Association. Retrieved from www.efpa.eu

Fernando, K., Chard, L., Butcher, M., & McKay, C. (2003). Standardisation of the Rey Complex Figure Test in New Zealand children and adolescents. *New Zealand Journal of Psychology, 32*(1), 33-38.

Gilbert vs Transfield, NZ Employment Court 71, CRC 46/10.

Ginosar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgements about another person. *Journal of Experimental Social Psychology, 16*, 228–242.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347–360.

Guenole, N., & Chernyshenko, O. S. (2005). The suitability of Goldberg's big five IPIP personality markers in New Zealand: A dimensionality, bias, and criterion validity evaluation. *New Zealand Journal of Psychology, 34*(2), 86-96

Guenole, N., Englert, P., & Taylor, P. J. (2003). Ethnic group differences in cognitive ability test scores within a New Zealand applicant sample. *New Zealand Journal of Psychology, 32*(1), 49–54.

Haitana, T., Pitama, S., & Rucklidge, J. J. (2010). Cultural biases in the Peabody Picture Vocabulary Test III: Testing tamariki in a New Zealand sample. *New Zealand Journal of Psychology, 39*(3), 24-34.

Hambleton, A., Kalliath, T., & Taylor, P. (2000). Criterion-related validity of a measure of person-job and person-organisation fit. *New Zealand Journal of Psychology, 29*(2), 80–85.

Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases*. Journal of Personality and Social Psychology, 39,* 578–589.

Hartigan, J. A., & Widgor, A. (1989) *Fairness in employment testing: Validity generalisation, minority issues, and the general aptitude tests battery***.** Washington DC: National Academy Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: Wiley.

Human Rights Commission. (2007). *Getting a job. An A–Z for employers and employees. Pre-employment guideline.* Wellington, New Zealand: Government Print.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98**.**

Hunter, J. E., Schmidt, F. L., & Hunter, K. (1979). Different validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, *86*, 721–735.

International Test Commission. (2013). *International guidelines on computer-based and internet delivered testing*. Retrieved from www.intestcom.org/guidelines

Jeng, M. (2006). A selected history of expectation bias in physics. *American Journal of Physics, 74 (7),* 578-583.

Kahneman, D. (2000). Evaluation by moments, past and future. In D. Kahneman & A. Tversky (Eds.), *Choices, values and frames* (pp. 693-708). New York: Cambridge University Press.

Kellett, D., Fletcher, S., Callen, A., & Geary, B. (1994). Fair testing: The case of British Rail. *The Psychologist*, *36 (10),* 26-29.

Kelly, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation (vol. 15).* Lincoln, NE: University of Nebraska Press.

Knight, R. G., McMahon, J., Green, T. J., & Skeaff, C. M. (2004). Some normative and psychometric data for the Geriatric Depression Scale and the Cognitive Failures Questionnaire from a sample of healthy older persons. *New Zealand Journal of Psychology, 33*(3), 163-170.

Knight, R. G., McMahon, J., Skeaff, C. M., & Green, T. J. (2008) Normative data for persons over 65 on the Penn State Worry Questionnaire. *New Zealand Journal of Psychology, 37*(3), 4-9.

Krynen, A. M., Osborne, D., Duck, I. M., Houkamau, C. A., & Sibley, C. G. (2013). Measuring psychological distress in New Zealand: Item response properties and demographic differences in the Kessler-6 screening measure. *New Zealand Journal of Psychology, 42*(2), 95- 109.

Lichtwark, I. T., Starkey, N. J., & Barker-Collo, S. (2013). Further validation of the New Zealand Test of Adult Reading (NZART) as a measure of premorbid IQ in a New Zealand Sample. *New Zealand Journal of Psychology, 42*(3), 60-68.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the literature*. Minneappolis, MN: University of Minnesota Press.

Mischel, W. (1986). *An introduction to personality*. New York: CBS College Publishing.

Morrow, P. C. (1983). Concept redundancy in organizational research: The case of work commitment. *The Academy of Management Review, 8*(3), 486–500.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement*. Englewood Cliffs, NJ: Prentice-Hall.

New Zealand Psychologists Board (2015). *The Use of Psychometric Tests.* Retrieved from www.psychologistsboard.org.nz

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. In R.F. Pohl (Ed.). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (79-96). Hove, United Kingdom: Psychology Press.

Palmer, S. (2004). Homai te waiora ki ahau: A tool for the measurement of wellbeing among Maori - the evidence of construct validity. *New Zealand Journal of Psychology, 34*(1), 50-58

Palmer, S. (2005). Psychometrics: An ancient construct for Maori. *New Zealand Journal of Psychology, 34*(1), 44-51.

Roberts, M. E., & Wilson, M. S. (2008). Factor structure and response bias of the Obsessive-Compulsive Inventory – Revised (OCI-R) in a female undergraduate sample from New Zealand. *New Zealand Journal of Psychology, 37*(2), 2-7.

Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology, 66*, 225–241.

Roth, P. L., Huffcutt, A., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88*(4), 694–706.

Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, B. (1978). From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology, 14*, 237–255.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178–200.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*(2), 195-202.

Sibley, C. G., & Pirie, D. J. (2013). Personality in New Zealand: Scale norms and demographic differences in the Mini-IPIP6. *New Zealand Journal of Psychology, 42*(1), 41-50.

Singer, M. (1993). *Diversity-based hiring: an introduction from legal, ethical and psychological perspectives*. Aldershot, United Kingdom: Avebury.

Starkey, N.J. & Halliday, T. (2011). Development of the New Zealand adult reading test (NZART): Preliminary findings. *New Zealand Journal of Psychology, 40*(3), 129-141.

Statistics New Zealand. (2013). *Household use of information and communication technology.* Retrieved from www.stats.govt.nz/

Strauss, H., Leathem, J., Humpries, S., & Podd, J. (2012). The use of brief screening instruments for age-related cognitive impairment in New Zealand. *New Zealand Journal of Psychology, 41*(2), 13-22.

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

Taylor, P., Keelty, Y., & McDonnell, B. (2002). Evolving personnel selection practices in New Zealand organisations and recruitment firms. *New Zealand Journal of Psychology, 31*, 8–18.

Taylor, S. E., & Thompson, S. C. (1982). Stalking the elusive "vividness" effects. *Psychological Review, 89*, 155–181.

Wicks, L., Siegert, R. J., & Walkey, F. H. (2004). A confirmation of the eight factor structure of the Eating Disorder Inventory in a non-clinical sample, with New Zealand norms. *New Zealand Journal of Psychology, 33*(1), 3-7.

Wright, S. L., Burt, C. D. B., & Strongman, K. T. (2006). Loneliness in the workplace: Construct definition and scale development. *New Zealand Journal of Psychology, 35*(2), 59-88