

## **Options and issues in the development of validation methodologies for Hua Oranga and Hōmai te Waiora ki Ahau**

The following paper sets out to describe options and issues which need to be considered in the development and implementation of validation methodologies for Hua Oranga and Hōmai te Waiora ki Ahau.

### **Hua Oranga**

Hua Oranga is a Māori measure of mental health outcome. Kingi and Durie (2000a) say it is “a cultural tool designed specifically to consider aspects of mental health outcome relevant to Māori mental health consumers” (p.57). It aims to provide:

- an appropriate theoretical perspective for the measurement of Māori mental health outcomes
- information which is accessible and acceptable for clinicians, providers and consumers of Māori mental health services, and
- information which contributes to health gains for Māori

The measure comprises three schedules of sixteen items grouped into four dimensions, ie four items for each dimension. It is administered to tangata whaiora (consumers), clinicians and a member of the consumer's whānau. For each of the three respondents, 5-point rating scales are used to obtain the sixteen scores. These are summed, combined and averaged to produce a single outcome score.

In contrast, Hōmai te Waiora ki Ahau is a twelve item tool for the measurement of psychological wellbeing among Māori (Palmer, 2002). It was not specifically designed as an outcome measure but it certainly has the capacity to be used for this purpose. Like Hua Oranga, this tool seeks to provide a culturally relevant paradigm for the conceptualisation and measurement of psychological wellbeing among Māori. Hōmai te Waiora ki Ahau is presented as a series of pictures which aim to describe the meaning of each concept and ensure that the measure can be administered even when respondents have little, or no, understanding of the Māori language. Thirteen-point

rating scales are used to obtain data on each of the twelve component items, these are then summed to give an overall waiora score. A poster which shows the pictures for each of the twelve component items and the scale used in the collection of data is available at: <http://www.publichealth.massey.ac.nz/homai/homai.htm>.

Work on the validation of both measures has already commenced and the available data is described below. The following discussion mainly aims to provide a framework for decisions regarding the next stages in development and implementation of validation methodologies for Hua Oranga and Hōmai te Waiora ki Ahau. This material is presented in four sections. The first section outlines the characteristics and qualities of a good health outcome measure. Section two presents an overview of the principles which underpin classical and modern psychometric approaches to the development of health outcome measures. Against this background, the third section looks at the data and implications of procedures which have, thus far, been used to validate Hua Oranga and Hōmai te Waiora ki Ahau. The final part of this paper aims to identify methodologies which would seem pertinent to pursue if work on the validation of these measures is to continue.

## What is a good health outcome measure?

Health outcome research is not well developed in Aotearoa (Kingi & Durie, 2000). In comparison with other countries, however, the New Zealand Government has a unique commitment to the implementation of health services which meet the needs and expectations of the indigenous population, Māori (Durie, 1997). Within the health arena, particularly mental health, Māori have a long history of involvement in the identification of criteria for evaluating the effectiveness of services (Durie, 2000). Over the last decade, national impetus for the development of mental health outcome measures has been generated with establishment of the Mental Health Commission and Mental Health Research and Development Project (Kingi & Durie, 2000). Among other goals, these initiatives aimed to assist the implementation of systems for measuring mental health service outcomes in order to improve the planning, purchasing and delivery of mental

health services in New Zealand. For Māori, it was recommended such outcome measures needed to be consistent with Māori concepts of health and wellbeing.

In 1997, Durie & Kingi developed a framework for the measurement of Māori mental health outcomes (MMHO). Table 1 displays the five principles which underpinned the development of this framework: ie wellness, cultural integrity, specificity, relevance and applicability.

**Table 1: The Five MMHO principles (Durie & Kingi, 1997)**

principles	characteristics
wellness	promotes wellness not just the absence of symptoms
cultural integrity	meets Māori expectations and aspirations for health development
specificity	has clear goals and objectives
relevance	is both useful and appropriate
applicability	is practical and manageable as an outcome tool

More specifically, the MMHO framework identified the need for measures which:

- ♠ consider the views of key stakeholders, ie clinicians, consumers and their families
- ♠ are consistent with Māori concepts of health and wellbeing
- ♠ can be administered at any, or all, of the main clinical endpoints, ie after assessment, in-patient treatment, outpatient treatment, community care and/or discharge from treatment or care
- ♠ provide information which have relevance for all stakeholders but complements the data obtained from clinical tools

With regard to a measure which is consistent with Māori concepts of health and wellbeing, the MMHO framework identified the need for a tool which has the capacity to measure the four domains identified in an accepted model of Māori health, Te Whare Tapa Wha. Māori are aware that a good health outcome in one culture may not be regarded positively in another (Kingi & Durie, 2000). However, it is suggested the likelihood of a favourable approach to Māori health outcome measures may be facilitated with development of tools which can demonstrate transparency, reliable and robust systems, validity, acceptability, adherence to consultation processes and cost-efficiency.

**Table 2: The qualities and characteristics of a good health outcome measure (Andrews et al, 1994)**

qualities	expected characteristics of data obtained
applicability	meaningful, useful, able to assist treatment
acceptability	brief, user-friendly, easily understood
practicality	minimal cost and training, simple instructions
reliability	acceptable psychometric qualities
validity	measures what it purports to measure
sensitivity to change	able to pick up changes

Internationally, it is the Australians who have shed considerable light on the qualities and characteristics of a good health outcome measure (Peters, 1994). Roughly a decade ago, for example, a group of authors have said that health outcome measures should be applicable, acceptable, practical, reliable, valid and sensitive to change (Andrews, Peters & Teesson, 1994). Table 2 presents some of the characteristics associated with data displaying these qualities. It seems a health outcome measure should provide accurate and meaningful information, be user-friendly and cost-efficient to administer and display appropriate psychometric properties as well as the ability to pick up change.

**Table 3: Ten criteria for the selection of a good health outcome measure (Sanson, 1996)**

criteria	definition of each criteria
reliability	appropriate psychometric properties
validity	does it measure what it claims to measure
discriminatory power	ability to discriminate between groups
responsivity	ability to detect change over time
type of instrument	generic, condition specific or both
style of instrument	rating scale, self-report or both
practical utility	user friendly, easy to administer
freedom from confounding factors	social desirability of responses, bias
relevance of application	is it the most appropriate measure
mode of administration	interview, telephone, mail-out

In 1996, Jan Sansoni put forward a similar but more elaborate framework for the selection of a good health outcome measure. Table 3 briefly describes the ten criteria of reliability, validity, discriminatory power, responsivity, type and style of instrument, practical utility, freedom from confounding factors, relevance of application and mode of

administration. Although both authors have contributed much to discussion about this issue, there is some doubt about the usefulness of these frameworks, in terms of their ability to guide the development and identification of a good health outcome measure.

Both authors, for example, have acknowledged the importance of reliability and validity but they fail to mention that the meaning of these terms, as well as the distinction between these conceptual domains, can be hazy (Aitken, 1997; Cohen & Swerdlik, 1999; Moss, 1994; Murphy & Davidshofer, 2001). Reliability is often said to be a necessary but not sufficient condition for validity. This seems to mean that an unreliable test cannot possibly be valid but a valid test is not necessarily reliable. The concept of validity, therefore, may have a degree of pre-eminence over that of reliability (Anastasi & Urbina, 1997; Johnston & Pennypacker, 1980). Even so, the definition of reliability, as well as the way in which it is estimated, clearly depends upon the purpose of the test, the particular attributes being measured and the circumstances in which it is measured. In other words, an indicator of reliability does not necessarily mean the test is reliable as it is may be reliable for some purposes but not others.

The concept of validity is similarly confusing. Indeed, the meaning of validity has been debated for a great many decades and clearly depends upon the context and/or purpose of tests (Bracht, Hopkins & Stanley, 1972; Johnson & Pennypacker, 1980). Popular definitions are often fragmented and incomplete (Embretson, 1983; Messick, 1995). In terms of the above models, however, validity is a concept that has the capacity to encompass most items. For example, a valid health outcome measure could also be shown to have acceptability and applicability. Although the importance of user-friendliness and cost-efficiency is not likely to be disputed, it could easily be argued that these are also components of acceptability, applicability or validity. Similarly, the type and style of instrument as well as its practical utility and mode of administration will surely have an impact on acceptability, applicability and validity. Discriminatory power, freedom from confounding factors and relevance of application are also functions of validity. For this reason, several support the idea that construct validity may be an overarching concept which has the capacity to encapsulate all that is meant by validity and its many qualities or criteria (Embretson, 1983; Johnston & Pennypacker, 1980;

Messick, 1995; Murphy & Davidshofer, 2001). No wonder, those involved with the development of social research methodologies have long been criticized for wrapping old concepts in new packages and failure to build on fundamental theory (Ebel, 1972; Embretson, 1983; Johnston & Pennypacker, 1980; Trochim, 2002).

At Australia's recently held Annual Health Outcomes Conference, therefore, it was not surprising to find a shift in the thinking behind a good health outcome measure (Eagar, 2002; Glasziou, 2002; King, 2002; Sansoni, 2002). Within the key challenges, it is interesting to note, there was no mention of reliability or validity but, instead, a clear message about the need for:

- ♠ greater expertise in the use and development of tools which have empirical, theoretical and conceptual relevance
- ♠ a broader perspective on research designs, not just RCTs, focus on effectiveness, as well as efficacy, greater co-ordination and collaboration across the research effort
- ♠ efficient and accessible information transfer, computer assisted technologies and user-friendly tools which do not duplicate the collection of data nor serve to further disadvantage the disadvantaged
- ♠ strategies which link health outcomes to policy, practice and resource allocation, re-definition of context, focus on generalizability versus sustainability of measures, more efficient models, short terms costs versus long term gains, cost-efficiency
- ♠ consumer-oriented measures and models which meet clinical expectations and reflect community values

Such change has been spurred by developments in applied psychology and a mounting call to revise the tools of psychological measurement (Anastasi & Urbina, 1997; Embretson, 1996; Embretson & Reise, 2000; Murphy & Davidshofer, 2001).

The problem has been a mindset towards classical rather than modern, or model-based, psychometrics (Embretson, 1996; Embretson & Reise, 2000; Johnston & Pennypacker, 1980; van der Linden & Hambleton, 1996). Many, it seems, have been reluctant to explore how the development of tools for psychological measurement, as well as the collection and interpretation of data, may be influenced by psychometric theory. Within psychology and the health outcomes movement, enthusiasm for the classical methodologies has left a legacy of literature on reliability and validity but little which

addresses the broader theoretical concepts integral to an understanding of construct validity. Embretson (1983) has suggested the ability to demonstrate construct validity requires a paradigm shift towards a structural, rather than functional, approach and the gradual accumulation of information from a variety of sources. For many, therefore, the pathways towards construct validity provide important opportunities for task decomposition and theory construction which allow the researcher to understand not only individual differences but also the processes that underlie behaviour and the way in which test scores may interact with or be influenced by a universe of other variables (Anastasi & Urbina, 1997; Embretson, 1983; Johnston & Pennypacker, 1980; Murphy & Davidshofer, 2001).

The next section seeks to understand the manner in which classical and modern psychometric techniques may contribute to the development of construct validity and establishment of health outcome measures.

## Psychometric theory – key components and principles

Johnston and Pennypacker (1980) coined the terms *vaganotic* and *idemnotic* to illustrate a fundamental difference between two empirical approaches in the social scientists. The *vaganotic* tradition, traced to the 17<sup>th</sup> century, is based on an assumption of natural norms and the notion that error, or variability from true ideals, will display a bell-shaped distribution and reduce with more measurements. This approach uses variability, or error, to define and measure phenomena. By late 1800, Francis Galton, Alfred Binet and others had synthesized these ideas to launch various movements for the measurement of mental states<sup>1</sup>. The aim of *vaganotic* measurement is to reduce error and variability. This school of thought holds that large sample sizes are necessary to ensure representativeness and, thus, the generality of data to even larger populations. This is the principle of inferential statistics which is well suited to predict the behaviour of a population but not an individual (Johnson & Pennypacker, 1980).

---

<sup>1</sup>With help from mathematical statisticians, like Karl Pearson and Ronald Fisher, who provided techniques for calculating the correlation co-efficient, variance and the standard error of measurement.

In contrast, the term idemnotic implies the use of standard and absolute measurement units. This approach, evident by the mid-19<sup>th</sup> century, is based upon the enumeration of objective differences in terms of exact and absolute quantities, or unit standards like, for example, reaction time and the number of trials or correct items. Such paradigms were followed by the forefathers of experimental psychology - Wilhelm Wundt, Thorndike, Pavlov and Skinner. Within the idemnotic tradition, variability is considered to be a window through which individual differences may be observed. Various techniques, such as the standardization and/or anchoring of data, provide mechanisms for comparing different sets of data and working towards the overall objective which is to establish the generality, or generalisability, of data beyond test conditions. Idemnotic methodologies aim to understand behaviour at the individual level. The theory assumes that a small number of subjects will provide sufficient data to demonstrate sound functional relations which are a prerequisite for generality to a larger group (Johnson & Pennypacker, 1980).

Among psychometricians, the words vaganotic and idemnotic did not seem to catch on but the need for terms which distinguish between these two approaches, or methodological paradigms, has, nevertheless, been relevant. In recent years, psychometricians have witnessed a dramatic rise in the popularity of methodologies which clearly stem from more idemnotic than vaganotic origins. Within the literature, such idemnotic methodologies have been aligned with modern psychometric theory whereas those with vaganotic origins have been grouped under the realm of traditional, or classical, psychometric theory. The following discussion looks at the way in which classical and modern psychometric theory may contribute to the establishment of construct validity.

### Classical test theory

Classical test theory is based on two main principles which have provided a powerful foundation for the evaluation of psychological tests and health outcome measures



(Aitken, 1997; Anastasi & Urbina, 1997; Cohen & Swerdlik, 1999; Murphy & Davidshofer, 2001; Trochim, 2002)<sup>2</sup>.

The first principle assumes that a distribution of scores will always resemble the theoretically normal, bell-shaped curve. The larger the group, or sample size, the more closely a distribution will approximate this bilaterally symmetrical curve. This assumption provides the rationale for many techniques to describe and interpret the meaning of test data. In particular, it allows individual and aggregate test scores to be described and compared in terms of central tendency, variability around a central point or divergence from a mid-point in the frequency distribution. The mean, mode, median, variance, deviation and concepts of skewness, kurtosis and ceiling or floor effects have commonly served this purpose. These indicators, and others, provide powerful tools for describing relative performance in terms of central tendency or deviation from a normal distribution. The normal curve has been extensively used as a reference point for the transformation of raw data into standardized scores and the establishment of relative norms (Anastasi & Urbina, 1997; Murphy & Davidshofer, 2001).

The second main principle of true score theory assumes that every measurement, or test score (**X**), is a composite of two components: true score (**T**) and random error (**e**). Mathematically, this relationship is represented by the equation  $\mathbf{X} = \mathbf{T} + \mathbf{e}_x$ . In terms of variance (**var**) across a set of scores, the so-called variability of a measure, it is assumed that  $\mathbf{var}(\mathbf{X}) = \mathbf{var}(\mathbf{T}) + \mathbf{var}(\mathbf{e}_x)$ . The variability of a measure, therefore, is the sum of variability due to true score and variability due to random error. When sample size is large enough, the distribution of true scores and error will approximate the normal curve. More importantly, however, the notion of error is central to the establishment of reliability.

## Reliability Theory

In psychometric terms, the concept of reliability means repeatability or consistency. The basic tenet of reliability theory is that test scores reflect two factors: those that contribute

---

<sup>2</sup> Classical test theory is invariably called reliability theory, classical reliability theory and/or true score theory. The principles which underlie these theories are largely interchangeable.

to consistency, or true score, and those that contribute to inconsistency, or error (Murphy & Davidshofer, 2001). The theory posits that a measure with no random error (which means it is all true score) is perfectly reliable whereas a measure with no true score (which means it is all random error) will have zero reliability (Trochim, 2002). The first goal of reliability theory is to estimate the size of measurement error.

In accordance with classical test theory principles, it is assumed that two measurements of a person's response to the same test will be related to the degree that they share true score. The error component will vary randomly around true score. For this individual, the reliability of this measure is represented by the ratio of true scores to true score plus error scores (Trochim, 2002). Reliability, however, is a characteristic of measures taken across individuals and this ratio is, therefore, expressed in terms of variance,  $\text{var}(\mathbf{T})/\text{var}(\mathbf{X})$ . The second goal of reliability theory is to estimate how much variability in test scores is due to errors in measurement and how much is due to variability in true scores.

Because true scores are never known, an estimate of  $\text{var}(\mathbf{T})$  is calculated from the correlation between two measures or observations. The correlation coefficient is largely derived from the cross-product of variance for each measure divided by the product of their standard deviations. The resultant coefficient reflects the degree of consistency between two independently derived sets of scores. The concept of reliability, therefore, is always expressed as a correlation co-efficient (Anastasi & Urbina, 1997). Three assumptions underpin this approach: the mean error of measurement will always sum to zero<sup>3</sup>; true scores and errors are not correlated and there is no correlation between errors on different measures. In general, the reliability coefficient ( $r_{xx}$ ) is defined as the ratio of true score variance to total variance of test scores (Murphy & Davidshofer, 2001).

The third goal of reliability theory is to improve tests by suggesting ways to minimize error. Classical theory uses five main methods to calculate the reliability of test scores and each provides information on the source of error (Anastasi & Urbina, 1997; Murphy

& Davidshofer, 2001; Trochim, 2002). The rationale underlying each approach is simply that two equivalent, or parallel, forms of a test will produce a consistent, similar, result and any differences between the test scores will be due to errors in measurement. *Inter-scorer* techniques assess the degree to which different raters, or observers, give consistent estimates of the same phenomenon. The *test-retest* method examines the consistency of scores when the same measure is administered to the same group of people on different occasions. The *alternate forms* approach looks at consistency of scores when different tests, constructed from the same content domain, are administered to the same group of individuals. The *split-half* method administers the test just once but the results are split in two and the consistency of one half-test is compared against the other. In the *internal-consistency* approach, reliability is a function of the number of test items and the average inter-correlation among test items (Murphy & Davidshofer, 2001). For example, Chronbach's coefficient alpha provides an estimate of the mean reliability coefficient obtained when each item is compared with every other item. This method looks at the extent to which each item, or item content, represents an observation of the same thing observed by other test items.

Reliability theory provides information about the accuracy of a test and whether it is worthwhile investing more resources in the development of strategies to improve a test (Aiken, 1997; Cohen & Swerdlik, 1999; Murphy & Davidshofer, 2001). For example, the reliability co-efficient is used in the calculation of standard error of measurement which is, in turn, used to determine the preciseness of test scores, the confidence levels in a test score or the extent to which scores are likely to vary from one administration to another. The reliability co-efficient is also part of the Spearman-Brown formula which estimates how much time and effort is needed to increase the reliability of a test and whether this will lead to other improvements, like increased validity.

In general, however, the reliability of a test is influenced by characteristics of the people taking the test, characteristics of the test itself, the intended use of test scores and/or the method used to estimate reliability. Although the proper method for estimating and

---

<sup>3</sup> Random error pushes observed scores up or down randomly which means there will be an equal distribution of positive and negative errors. With enough observations, therefore, the sum of adding positive and negative errors together is zero.

interpreting reliability depends on what the test measures and how it is measured, the best way to ensure reliability is to obtain as many observations as possible of the attribute to be measured. Composite scores, therefore, are typically most reliable. The more tests, or items, that are combined and the higher the correlation between them the higher the reliability of the composite score (Murphy & Davidshofer, 2001).

## Generalizability Theory

Although reliability theory has provided a useful foundation for analysis of test data for many decades, the approach has limitations (Embretson 1996; Embretson & Reise, 2000; Murphy & Davidshofer, 2001; Shavelson, Webb & Rowley, 1989). The main criticisms of classical reliability theory appear to be its' acceptance that error is always random and the fact that neither measurement error nor the reliability co-efficient are stable as both vary in accordance with the method for collecting data and/or the method of calculation.

Generalizability theory has extended classical reliability theory in a number of important ways (Murphy & Davidson, 2001; Shavelson, et al, 1989)<sup>4</sup>. In particular, it has recognized that measurement error may not always be random. Generalizability theory has divided the error component in true score theory into two sub-components: random error ( $e_r$ ) and systematic error ( $e_s$ ). Random error adds variability to the data but does not affect the average performance of a group<sup>5</sup>. In contrast, systematic error will affect the average performance of a group because it has a systematically positive, or negative, effect on the responses of all participants<sup>6</sup>.

The central issue in generalizability theory is the extent to which the results obtained from one measure can be generalized to another. Instead of looking at the consistency of test scores, this theory examines the causes of variability in test scores and the extent to which such variability can be attributed to systematic and unsystematic (random) sources.

---

<sup>4</sup> Classical reliability theory is seen to be a variant of generalizability theory.

<sup>5</sup> For this reason, it is described as a *nuisance* variable and is often called *noise*.

<sup>6</sup> Systematic error is, therefore, called *bias*.

Murphy & Davidshofer (2001) say the main advantage of generalizability theory is conceptual. A test score is seen to be a single sample from a universe of possible scores. The theory encourages the identification of situations in which a test might be reliable, or unreliable, and provides a context for thinking about test results. The concept of reliability is a characteristic of the use of test scores rather than a characteristic of the scores themselves. Although classical techniques may be useful when tests are given under highly standardized conditions, generalizability theory is more appropriate when conditions are likely to affect test scores or test scores are used for different purposes. As a method for defining and estimating reliability, many expect generalizability theory to replace the more traditional methods.

### Validity Theory

In a statistical context, the concept of validity has always aimed to understand the meaning and implications of test scores (Murphy & Davidshofer, 2001). It is usually associated with notions of correctness, accuracy and freedom from bias. And the question of validity is often phrased in terms of how well a test measures the concept, or construct, it sets out to measure. In general, the main approaches to validity measurement have been through analysis of test content, correlations between test scores and criteria or investigations of the construct's underlying properties. Validity estimates, however, often depend upon the specific purposes for which the test was designed, the target population and the methodological approach to measurement (Aitken, 1997; Cohen & Swerdlik, 1999; Murphy & Davidshofer, 2001).

Trochim (2002) has provided a contemporary framework for thinking about the meaning and measurement of validity. He points out that those involved in the development of psychological measures need to talk about the validity of operationalizations, rather than the validity of measures, because it is the way in which a concept has been translated into a functioning, operating reality that is really of interest. Furthermore, Trochim (2002) is among a growing number of researchers who have called for a unitary view of validity (Anastasi & Urbina, 1997; Embretson, 1983; Messick, 1995; Murphy & Davidshofer, 2001). These authors maintain the various techniques used to measure validity provide information on different aspects of construct validity not different types of validity.

Collectively, it is argued, the different approaches to validity measurement are not mutually exclusive and all have a bearing on construct validity. Most would say that the ability to demonstrate construct validity is reliant on the gradual accumulation of knowledge from a range of sources. However, debate about the way in which such sources of knowledge should be conceptualised is clearly evident.

Trochim (2002), for example, has said that techniques for the measurement of construct validity fall into two main categories: translation validity and criterion-related validity. He coined the term translation validity to describe procedures which aim to show that the construct is well defined and the operationalization is a good reflection of the construct. Techniques for the assessment of face validity and content validity serve this purpose:

- ♠ face validity looks at whether the physical appearance and content of a test have relevance for the people being tested and/or those who will be administering the test, it is usually based on subjective judgement and looks at issues of appropriateness and acceptability
- ♠ content validity looks at whether the test content is adequately representative of the conceptual domain it was designed to measure, this may involve the development of a blueprint, or universe of possible content, from which test items can be chosen; asking experts, referees or members of the target population to rate the relevance of test content - the content validity ratio provides a way to express this information

In contrast, the procedures for describing criterion-related validity provide an opportunity to examine whether the operationalization behaves the way it should given the theory behind the construct. Criterion-related validity refers to the use of a standard, or criterion, to test whether the operationalization of a construct functions in predictable ways. The rationale is relational, rather than definitive. A relational approach acknowledges that change from one construct to another may be gradual, constructs are not necessarily one thing or another but may be a mixture of several constructs. The objective of criterion-related validity techniques is to provide information on the extent to which constructs may correlate with one another. This information is always in the form of a correlation, or validity, coefficient. The assessment of criterion-related validity has traditionally involved techniques which provide information on:

- ♠ predictive validity, or the construct's ability to predict something the measure should theoretically predict, such as an increase or decrease with age and/or experience
- ♠ concurrent validity, or the ability to distinguish between groups that should theoretically be different
- ♠ convergent validity which is based on the idea that theoretically similar constructs should be related to each other and should correspond or be able to demonstrate convergence
- ♠ discriminant/divergent validity which looks at the degree of difference between theoretically different concepts, it reflects the idea that theoretically different constructs should not be related to each other

Murphy & Davidshofer (2001) concur with the idea that validation strategies can be grouped under the broad heading of construct validity but they suggest the outcomes of these processes have two distinct uses. In their view, such strategies aim to provide information on either the validity of measurement or the validity of decisions. To this end, content and construct validation strategies are seen to provide information on the validity of measurement, ie - the test is valid if it measures what it is supposed to measure whereas predictive and concurrent validation strategies provide information on the validity of decisions, ie - the test is valid if it can be used to make correct or accurate decisions. Briefly, it is suggested that content-oriented validation procedures will:

- ♠ describe the content domain
- ♠ determine which areas of content domain are measured by each test item, and
- ♠ compare the structure of the test with the structure of the content domain.

Internal consistency, therefore, is an important indicator of content validity. In contrast, construct-oriented validation procedures will aim to identify:

- ♠ behaviours that relate to the construct to be measured
- ♠ similar constructs and their relationship to the construct being measured
- ♠ behaviours associated with the similar constructs and, on the basis of the relations among constructs, determine whether these behaviours are related to the construct being measured

Murphy & Davidshofer (2001) have shown that the usual technique for assessing construct validity is through correlations between test scores and other tests, factor

analysis and/or experimental manipulation of the construct being measured<sup>7</sup>. Factor analysis is a method for estimating the correlations between a specific variable, like the test score, and scores on the factor, or items in a test. This provides information on the relationships between variables and whether expected relationships exist. By general consensus, however, the multitrait-multimethod matrix (MTMM) appears to be one of the most useful techniques for assessment of construct validity (Anastasi & Urbina, 1997; Murphy & Davidshofer, 2001; Trochim, 2002). The MTMM requires several concepts, or traits, to be measured by a number of different methods. This technique provides a method for interpreting multiple sources of construct validity data simultaneously. In particular, it provides information on convergent validity, discriminant validity and test bias.

Nevertheless, Embretson (1983) has presented an alternative approach to the issue of construct validity. In her view, the process of construct validation should involve research methods which aim to provide information on construct representation and nomothetic span. More specifically, construct representation is concerned with identifying the theoretical mechanisms that underlie individual task performance whereas nomothetic span looks at whether the test is embedded in a network of relationships to other measures. The latter term clearly stems from Chronbach and Meehl's notion of a nomological network (Trochim, 2002). It seems the concept of a nomological network provided the rationale for linking theory to observations, as a means to investigate construct validity, but nomothetic span has provided the method (Embretson, 1983; Trochim, 2002).

Nomothetic span looks at the utility of a construct as a measurement of individual differences. Such differences are correlated with other test scores, group membership and/or various socially important criteria to check theoretical assumptions about the construct, or constructs, being measured. Construct representation and nomothetic span are interactive phases of the construct validation research process. However, procedures for the achievement of nomothetic span are most effective when the

---

<sup>7</sup> They use the term 'construct explication' to describe the procedure for establishing construct validity, the end result of construct explication is a detailed description of the relationships among a set of constructs and behaviours.



constructs likely to underlie item solving have been identified in a construct representation phase. Embretson (1983) has shown that item response theory, in particular, the multicomponent latent model (MLTM), provides the best approach to construct representation whereas an understanding of nomothetic span can be achieved through the more conventional process of structural equation modeling<sup>8</sup>. This discussion will return to the methodological objectives of construct representation and their significance for construct validity. At this point, however, it is pertinent to address the final topic in this discussion on psychometric methodologies. The following section turns attention to the principles which underlie item analysis and item response theory.

### Traditional Item analysis Techniques

No matter which way it is approached, the reliability and validity of any test depends upon the characteristics of its items. For example, the reliability and validity of a test may be limited because items are poorly worded and/or do not match the content domain the test was designed to measure. Item analysis aims to identify ineffective items and provide diagnostic information that may help to improve reliability and validity (Anastasi & Urbina, 1997; Murphy & Davidshofer, 2001). An item analysis looks at individual differences in the pattern of responses to each item in three main ways, ie: are responses likely to be the result of guessing; was the item difficult and did the item discriminate between groups within the sample. These strands of information are conceptually distinct but empirically related. This section will briefly describe the approach to item difficulty and item discrimination.

Item difficulty is defined in terms of the relative frequency with which those taking the test choose the correct response.<sup>9</sup> It is measured as a percentage, or p value, and is, thus, a characteristic of both the item and the population taking the test. Item difficulty affects the variability of test scores and the precision with which test scores can discriminate between groups. When all items are difficult or easy, there will be little

---

<sup>8</sup> Structural equation modeling is similar to path analysis and/or causal modeling – all techniques aim to identify whether a causal relationships exists between constructs.

<sup>9</sup> The concept of “correct” may be in terms of either a dichotomous (yes/no) response or a keyed response, such as a high, rather than low, score on a rating scale. If everyone chooses the correct answer the item is easy whereas if most get it wrong the item is defined as difficult.

variability. In terms of central tendency and the normal curve, this will be reflected in the distribution of scores. For example, a piling of scores at the lower end of the scale suggests the test floor is too high, or lacks a sufficient number of easy items to discriminate properly at this end of the range. Alternatively, a piling at the upper end suggests a low ceiling because the test is lacking in difficult items. Such skewness makes it impossible to measure individual differences.

Three strategies are used to measure the discriminating power of an item:

- ♠ The  $D$  statistic is based on the rationale that a good item will discriminate between those who score high and those who score low, that is, more people in the top-scoring group will have answered the item correctly and vice-versa. Selecting items with high  $D$  values will result in an internally consistent test as the correlations among items will be highly positive
- ♠ The item-total correlation, also called the point biserial correlation, represents a correlation between the score on an item and the total test score. A positive correlation suggests two things: that the item successfully discriminated between those who do well and those who do poorly and the item is internally inconsistent, or measures the same thing that is being measured by the rest of the test
- ♠ The interitem correlation matrix looks at the correlations between all test items, this will sort the items of a test into those which are internally consistent and those which are not and may help to explain why some items fail to discriminate

The point biserial correlation is also used to examine the validity of an item in terms of its ability to predict an external criterion. In this case, item scores are correlated with scores on the criterion measure. It should be noted, however, that items selected on the basis of a  $D$  statistic will yield a different kind of test than one composed of items selected because of high correlations with an external criterion. The method for determining discriminatory power, therefore, depends on the purpose of the test. When external validity is important, items having high correlations with the criterion but low correlations with other items are best because they make a more independent contribution to the prediction of criterion scores. When internal consistency is important, items which correlate with each other are chosen. A combination of the two approaches may sometimes be appropriate. Relatively homogenous items can be sorted into separate tests, or subtests, each of which covers a different aspect of the external criterion. A composite test may then be constructed from subtests which have low correlations with each other and substantial correlations with an external criterion, but

the items within each subtest are internally consistent (Aitken, 1997; Anastasi & Urbina, 1997; Murphy & Davidshofer, 2001).

### Item response theory – a modern approach item analysis

Traditional item analysis techniques have a number of limitations. Examples of common criticisms would be:

- ♠ responses are sample dependent or influenced by the types of people who take the test
- ♠ large sample sizes and complex research paradigms are needed
- ♠ longer tests are more reliable than shorter forms
- ♠ tests are static and cannot be changed, different versions of the test are not comparable
- ♠ even acceptable p and D values do not guarantee that an item is functioning effectively across all levels of overall test performance
- ♠ reliability and validity is based on test scores rather than responses to individual items

Item response theory (IRT), also called latent trait theory, item characteristic curve theory and/or model-based test theory, is providing an increasingly popular alternative to traditional item analysis techniques (Anastasi & Urbina, 1997; Embretson, 1983, 1996; Embretson & Reise, 2000; King, 2002; Hambleton, Swaminathan & Rogers, 1991; Murphy & Davidshofer, 2001; van der Linden & Hambleton, 1996). Item difficulty and discrimination is defined in terms of the relationship between the attribute being measured and the individual's response<sup>10</sup>. The underlying rationale is simply that: if a test measures a specific attribute, then the more of the attribute the person has the more likely it will be that they answer the item correctly. IRT has evolved around four main assumptions:

- ♠ individual performance on a test item can be predicted or explained by a set of factors called latent traits, or abilities
- ♠ only one ability is measured by a set of items in a test, each test is unidimensional
- ♠ items have local independence, when ability is held constant there is no relationship between responses to test items, abilities are the only factors which influence responses

---

<sup>10</sup> In the traditional approach, difficulty and discrimination is defined in terms of the people taking the test (the p statistic) and other item on the test (the D statistic).

- ♠ the relationship between examinees item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (ICC).

The ICC plots the probability of choosing a correct answer to an item as a function of the level of attribute being measured by the test. It can provide a graphic summary of item difficulty, discriminatory power and the probability of answering correctly by guessing. In addition, the ICC can provide information on how the item is functioning across a range of criterion scores.

All IRT models contain one or more parameters to describe the item and the respondent's ability or trait. An item response curve is constructed by plotting the proportion of people who answer an item correctly against an estimate of their true standing on a unidimensional latent trait or characteristic. These estimates are derived by mathematical functions which vary with the assumptions and estimation procedures prescribed by the particular approach. Different IRT models utilize different mathematical functions based on different sets of assumptions. Some, for example, use normal ogive functions while others use logistic functions. Estimates of the item and ability parameters are computed by successive approximation. This process of iterative approximation is repeated until the values stabilize and there is a goodness of fit, or the mathematical model can accurately predict and estimate the data.

With the rapid escalation of computer assisted technologies, the popularity and sophistication of item response theory has increased dramatically in recent years (King, 2002; Sansoni, 2002). Some of the newer models, for example, have the capacity to handle multidimensional tests as well as graded and/or multiple choice data (Embretson & Reise, 2000; Hambleton, 1983; van der Linden & Hambleton, 1999). For psychometricians, the emergence of IRT has introduced new rules for measurement which have a number of advantages over the methods provided by classical test theory (Embretson 1983; Embretson & Reise, 2000). In particular, item response theory offers:

- ♠ sample invariance because ability estimates are not test dependent and item indices are not group dependent, IRT provides a uniform scale of measurement, different set of items can be administered to different groups and their scores will be directly comparable

- ♠ adaptive testing and/or the ability to tailor tests to specific needs, common items - called anchor, linkage or calibration items – can be used to bridge the gaps across large sample groups and link a number of different constructs together, once items in the pool, or bank, are calibrated any subset of items can be administered to any group or individual and the resulting scores will be comparable
- ♠ shorter tests can be more reliable than longer tests
- ♠ better methods for detecting and understanding the consequences of test bias, the ICCs allow differences between individuals and groups to be easily displayed
- ♠ standard error is estimated for each individual, rather than the entire group, the standard error of measurement differs across scores, or response patterns, but generalizes across populations
- ♠ falsifiable models in that a given model may, or may not, be appropriate for a particular set of test data. In classical theory, item responses are linked to total test scores rather than the construct, if the test is a poor measure of the construct the relationship between item scores and total test scores will say little about the item
- ♠ statistical robustness, various assumptions can be violated without distortion of the results eg unbiased estimates can be obtained from unrepresentative samples, data can accumulate or aggregate over time, randomized controlled trials are not necessary

Murphy & Davidshofer (2001) suggest the main advantage of IRT is conceptual because it encourages researchers and test administrators to think about why people respond the way they do. At this point, however, the chief limitation of IRT would appear to be that the discipline is still being developed (Embretson & Reise, 2000). For this reason, the available software and supporting material is neither user-friendly nor appropriate for use among those who have little understanding of IRT. Some IRT programs, for example, have user options with unexplored empirical consequences and all IRT software is stand alone which means it is not compatible with the more popular statistical packages, like SAS or SPSS. Unfortunately, assistance and expertise in IRT principles and the use of IRT techniques can be hard to find (Embretson & Reise, 2000).

## Summary of Psychometric methodologies

The above discussion has briefly outlined principles which underlie the classical approach to measurement of reliability and validity. One of the main points to be taken from this discussion is that the actual methodology will depend upon the objectives and purposes of each test. It is clear that generalizability, rather than reliability per se, is the

more relevant goal because it seeks to understand the extent to which test scores may not only be influenced by systematic error but also generalizable to other populations. Furthermore, a comprehensive approach to item analysis can provide otherwise inaccessible opportunities to gather information on factors which contribute to reliability and the validity of tests.

Within psychology and the health outcomes movement, current thought would suggest that validity issues have an initial pre-eminence over those of reliability, or generalizability, for two important reasons. Firstly, it is clear that reliability is a necessary but not sufficient condition for validity. This means that an unreliable test can never be valid whereas a valid test may, or may not, be reliable. The issue of validity, therefore, would seem to be a sensible place to state. More importantly, however, there is an international call for the tests and measures which have a strong theoretical orientation and the capacity to verify links to psychological theory through empirical and experimental hypothesis.

Despite consensus on the value of construct validity as a suitable mantle for strategies associated with the assessment of validity, opinion on the methodological approach to this issue is divided. On one hand, there is support for processes which foster the gradual accumulation of knowledge about networks and links which may exist between content domain and psychological constructs. On the other hand, there is disgruntlement with traditional approaches to construct validity and demand for innovative strategies which can identify the theoretical mechanisms behind test performance and can, therefore, assist the establishment of nomological knowledge.

As a first step in the construct validation process, Embretson (1983), has shown that multicomponent latent trait modeling has all the advantages of a conventional multitrait multimethod matrix and more. It seems the IRT model is able to not only decompose a test into basic theoretical constructs but also identify person differences and item indices which permit the test to be compared by item difficulty parameters that represent the theoretical properties embedded in each item.

For many researchers, Embretson & Reise (2000) suggest the old rules of measurement still have relevance and will be adequate for most purposes. Nonetheless, it is important to understand the implications of IRT as an innovative psychometric technique and increasingly popular basis for the development of psychological tests. Among those less experienced in the development of psychological measures, IRT offers a very attractive alternative to classical test theory for a variety of reasons especially the notions of sample invariance, statistical robustness and adaptive testing as well as the availability of information on item characteristics, individual abilities and test bias. As a psychometric method, the mathematical complexity of IRT is only accessible when there is compatibility with computer assisted technologies. However, the level of analytical sophistication offered by the combination of these techniques means that IRT has the capacity to test the falsifiability of its' models in terms of their ability to identify theoretical mechanisms. In the establishment of construct validity, therefore, IRT would seem to have a major advantage over traditional techniques.

In practical terms, however, IRT is in an emergent discipline which means the expertise, technology and support processes necessary for mainstream use are still being developed. Although appropriate mentorship may be hard to find, those involved in the development of psychological tests and health outcome measures need to be open to the possibilities of IRT. Indeed, it would seem important to work towards the establishment of methodologies which reflect and explore both classical and modern psychometric theories.

**hua oranga & hōmai te Waioira ki Ahau: current status and available data on psychometric properties**

### **Hua Oranga**

- ♠ a draft tool was developed on the basis of MMHO framework recommendations, ie
  - the outcome will reflect the views of key stakeholders - clinicians, tangata whaioira and their whanau
    - separate outcome schedules/questionnaires were developed for each of the three stakeholders

- the interview schedule was framed around the four dimensions of te whare tapa wha - taha wairua, taha hinengaro, taha tinana and taha whānau
  - four baseline questions were developed, one for each dimension of te whare tapa wha
  - an appropriate version of the baseline question was developed for each stakeholder, the total schedule comprised 12 items
- the information would be accessible, acceptable to clinicians, providers and consumers
  - a simple scoring system was developed
  - the scoring system reflected both positive and negative consequences of treatment, ie scores for each question ranged from -2 to +2
  - the scoring schedule is triangulated or combines the three stakeholder scores to give a combined score which is then averaged to produce an outcome score
  - the questionnaires can be administered at any or all of the five clinical endpoints: assessment, inpatient treatment, outpatient treatment, community care and discharge
  - the measure provides a unidimensional outcome score and/or a profile of scores for each stakeholder by content domain and/or the clinical end-point(s)
- ♠ evidence of validity
  - key individuals were asked to comment on the MMHO framework and draft measure
    - this resulted in some modifications to the draft tool eg the fifth clinical end-point was changed from discharge to community support
  - Phase One validity test
    - six different settings within four geographic regions were chosen as test sites: urban/rural mix, varying degrees of acculturation and de-culturation
    - consultation hui were held with representatives from each test site
    - feedback from the consultation hui participants was used to modify the tool
    - 170 participants took part in an initial test (59 clinicians, 59 tangata whaiora and 52 whānau members) - the questionnaires were administered and respondents were asked to evaluate their experience in terms of whether they understood the questions; thought the questions could be improved; thought the questions were relevant as a way of measuring the outcome of treatment or care and/or had suggestions for improvement
    - 86-100% of the respondents understood the question and 79-91% thought the questions had relevance



- respondents identified the need to refine and clarify the content of questions and the inappropriateness of the tool for certain groups, eg those with significant mental health impairments and/or little cultural knowledge
- some whānau members did not complete the questionnaire
- feedback from respondents was used to modify the interview schedule, namely four dimensions for each of the four te whare tapa wha domains were identified and twelve items were added to the schedule so that four items, instead of one, described each domain of te whare tapa wha
- in the new schedule each stakeholder would be asked sixteen questions, one question for each dimension of the four domains, when multiplied by three the total number of items was 48
- Phase Two validity test: the tool was re-tested in the same settings with a different, smaller group of stakeholders
  - consultation hui were held with representatives from each of the test sites
  - 75 participants took part in the re-test, 25 in each stakeholder group
  - respondents were asked to complete the measure then evaluate their experience in a manner similar to that identified in Phase One
  - responses suggested between 85-100% felt the measure was able to be understood and relevant
  - feedback suggested some groups would find the questions difficult, ie those who were significantly impaired, not comfortable in cultural settings and children
  - some respondents said some of the items in some of the domains were difficult to understand
- No further changes to Hua Oranga have been made

## Hömai te Waiora ki Ahau (HtWkA)

Item content was mostly developed from a review of the literature but the appropriateness of this material was informally discussed with a range of Māori experts.

The intended applications are:

- ♠ a tool for the measurement of waiora among Māori
- ♠ able to be administered when respondents have little, or no, understanding of te reo Māori
- ♠ provides an opportunity for Māori to think about the waiora they obtain from each dimension
- ♠ provides an aggregate, unidimensional, score which may help to measure the effectiveness of health interventions for Māori

- ♠ provides multidimensional information, or a profile of scores, which may help to identify domains of wellness and/or possible pathways for personal development;
- ♠ may assist a transformation of consciousness towards psychosocial constructs which have relevance for Māori
- ♠ may have use as a health outcome measure

A pre-pilot test (n=10, Maori women aged 16-65 years) displayed evidence of face and content validity, ie the measure was able to be administered, participants were able to differentiate between each component and responses varied. As an indicator of criterion-related validity, a linear relationship was evident between aggregate waiora scores and participants' self-rated feelings of overall waiora ( $r = 0.91$ ,  $p > 0.01$ ). No relationship was found between the waiora scores and a non-Māori measure of psychological wellbeing. At .69 Chronbach's alpha coefficient for the twelve component items suggested a borderline level of internal consistency but it was acceptable to proceed to the pilot study.

Results of the pilot-study showed (based on interviews with n =31 self-identified Māori women, aged 16-34 years, during the last trimester of pregnancy):

- ♠ evidence of irregular score distribution: three of the twelve item means did not fall within the middle zone of the rating scale; the difference between item mean and median was  $\geq 1$  for six items; six items displayed significant skewness or kurtosis; five items yielded low standard deviations ( $\sigma \leq 3$ ).
- ♠ low variability on some items: the full range of score alternatives were not always utilized, respondents tended to score towards the upper end of the scale and responses were sometimes clustered too closely together
- ♠ evidence of reliability:
  - the distribution of scores for the measure was normal
    - the overall mean ( $\chi = 95.64$ ) fell within one standard deviation of the mid-score for the total measure (48-96);
    - the distribution of scores for the measure did not show significant skewness or kurtosis and there was no disparity between the mean and median.
  - the robustness of this measure could be improved
    - scores tended to fall towards the upper end of the rating scale
    - the standard deviation was relatively small ( $\sigma = 16.98$ )
    - respondents did not utilize the full range of rating options available

- the measure was sensitive to individual differences – both the one-way ANOVA ( $F_{(30,11)} 9.4559$ ,  $p > .001$ ) and Hotelling's T Squared ( $F_{(11,20)} 4.3069$ ,  $p > .001$ ) were significant.
  - the scale was largely composed of internally consistent items:
    - none of the items had an  $r_{\text{itot}} > .7$  but six items failed to reach significance
    - Chronbach's co-efficient  $\alpha$  for the scale was .6486 and this would only have improved slightly with the removal of items
  - Such findings suggest a borderline level of internal consistency which may improve with development of the measure's robustness and sensitivity.
- ♠ evidence of validity:
- content validity – concepts used in the development of this measure were derived from the literature, expert opinion and peer group discussions.
  - face validity – the measure was able to be administered, respondents understood the questions and were willing to participate in the interview.
  - convergent validity: a significant correlation coefficient was found between aggregate waiora scores and self-rated feelings of overall waiora ( $r = .49$ ,  $p < .01$ ). This relationship also provides evidence of criterion-related validity, that is, the aggregate score was a predictor of self-rated waiora.
  - discriminant validity/freedom from confounding factors:
    - no evidence of a linear relationship was found between waiora scores and the scores obtained from administration of a non-Māori measure of psychological wellbeing, Affectometer 2.
    - no evidence of linear relationships were found between the aggregate waiora scores and indicators of ethnic identity, ie - whether respondents considered Māoritanga to be important, were able to identify their iwi, identified with an active stage of Māori identity development and/or thought of themselves as Māori (rather than part Māori, part Pākehā or mostly Pākehā).

### Possible issues for consideration in the development of Hua Oranga & Hömai te Waiora ki Ahau

At the beginning of this paper, the objectives which underpin a Māori mental health outcome measure were described along with a summary of the qualities and characteristics that have been associated with a good health outcome measure. This provided a broad framework for conceptualising the long-term objectives of Hua Oranga and Hömai te Waiora ki Ahau as health outcome measures for Māori. The second part of this paper looked at the methodological issues which underlie the establishment of a good health outcome measure. Against this background, available data on the current

status and psychometric qualities of Hua Oranga and HtWkA have been summarized. In conclusion, the following section aims to identify issues which would seem to have immediate relevance for the development of each measure.

## Hua Oranga

- ♠ the existing measure has yet to be tested – no data is available as yet
  - what is an appropriate research paradigm
  - can the data obtained from implementation of the measure during the Phase One and Two validity tests be analyzed?
- ♠ What single construct is the tool measuring, ie Hua Oranga? Te whare tapa wha? Or is it four separate constructs .... wairua, whānau, hinengaro, tinana?
- ♠ Is the measure unidimensional or a multidimensional composite score
  - should it be a composite/ aggregate score made up of 16 internally consistent items, or
  - 4 independent subtests each of which contains 4 internally consistent items
- ♠ Content domain
  - How was the content of each item chosen , who designed the questions? Why was this process valid
  - Are the items representative of the content domain they are meant to represent? What is the content domain ? Strategies to define the content domain? eg ... how do we know that wairua is measured by the question - Does it make you feel stronger as a Maori? Why is this question any better than the numerous other ways in which wairua could be tested, ie – effect of specific processes (use of water, karakia, te reo, lifting of tapu, what value is placed on the involvement of whānau/tohunga, is there a choice, how do consumers feel about the clinical processes - taking of bloods, diagnoses, labels), comparison between clinical/pathological/whānau oriented models of treatment, involvement in decision-making (what promotes a feeling of empowerment, are there processes for resolution of grievances, closure to diagnoses/prognosis), confidence/satisfaction (was it psychoses or mate Maori)
  - What does the item analysis for existing data say (discrimination, difficulty, ICC?)
  - would an IRT model offer the opportunity to develop the content domain for each item but still maintain the objectives of this measure
  - Is there a need to further examine construct validity – can both classical (transitional, convergent, divergent, criterion, predictor) and modern paradigms (item banking, identification of anchor items) be implemented
  - Is there a need to develop a nomological network – are responses influenced by, for example, ethnic identity, level of acculturation, satisfaction with services, cultural audit – would a combination of methods

be appropriate (structural equation modeling/ multitrait multimethod matrix/ multicomponent latent trait model?)

♠ Method

- Is the triangulated approach reliable, ie: consistency in scores obtained from tangata whaiora, clinician, consumer (test-retest, alternate forms, inter-rater reliability, generalizability)?
- What are the sources of error?
  - Is there evidence of test bias, halo/leniency effects?
  - Is there a ceiling effect in the Phase One and Two validity data?
- Is there a better method for presenting the items?
  - Eg open-ended vs multi-choice vs structured interview vs hui
  - Computer vs face-to-face vs self-rated

## Hömai te Waiora ki Ahau

♠ Construct

- is it multi or unidimensional?
- Is it acceptable, perceived value, relevance
  - survey referee/expert opinion
  - strategy to check relevance of concepts and construct as a tool for the measurement of waiora in a range of contexts eg community, professional, clinical, whanau, iwi, hapu,
  - capacity as wellbeing tool vs health outcome measure

♠ Content domain

- is it adequately defined?
- do the items represent the concepts
  - strategies to check for content validity eg focus group discussions with a range of groups, multiple choice survey (check which items represent the concept), survey perceived relevance of items
  - item banking – aim for internal consistency in item?
  - item analysis and ICCs of existing data/new data
  - check for evidence of test bias
- establishment of nomological network/theoretical base
  - establish strategies for construct explication/representation?
  - structural equation modelling/MTMM or multicomponent latent trait models to identify relationships between concepts and constructs, assist with the development of theory
  - Relationship between waiora and te whare tapa wha items?
- moderator variables

- Influence of ethnic identity development, ie active/passive stages?
- Does te reo Maori make a difference?

♠ Method

- Is there a better approach than pictures, eg pen and paper, multi-choice, CAT
- adaptive tests/shorter tests – is this a possibility? Strategies for item banking, anchor items to bridge across populations
- consequences of the method – does it have an effect – design/administer post-test questionnaire to gain qualitative information about perceived value eg did you understand questions, do they have relevance, has it influenced your thinking
- feasibility in different contexts? eg the health arena, within whānau, schools

References

1. Aitken, L.R. (1997). *Psychological Testing and Measurement*, 9<sup>th</sup> edition. Needham Heights, MA: Allyn & Bacon.
2. Anastasi, A. and Urbina, S. (1997). *Psychological testing*, 7<sup>th</sup> edition. New Jersey: Prentice Hall.
3. Andrews, G.A., Peters, L. and Teesson, M. (1994). *Measurement of Consumer Outcome in Mental Health*. A Report to the National Mental Health Information Strategy Committee. Sydney: Clinical Research Unit for Anxiety Disorders.
4. Bracht, G.H., Hopkins, K.D. and Stanley, J.C. (1972). *Perspectives in Educational and Psychological Measurement*. Englewood Cliffs, New Jersey: Prentice-Hall Inc.
5. Cohen, R.J. and Swerdlik, M.E. (1999). *Psychological Testing and Assessment: an introduction to tests and measurement*, 4<sup>th</sup> edition. California: Mayfield Publishing Company.
6. Durie, M.H. (1997). *Whaiora: Māori Health Development*. 2<sup>nd</sup> Edition. Auckland: Oxford University Press.
7. Durie, M.H. (2000). *Mauri Ora: The Dynamics of Māori Health*. Auckland: Oxford University Press.
8. Durie, M.H. and Kingi, Te K.R. (1997). *A Framework for Measuring Māori Mental Health Outcomes*. A report prepared for the Ministry of Health by Te Pūtahi-a-Toi, Massey University, Palmerston North.
9. Eagar, K. (2002). Keynote address. *Health Outcomes 2002 - Current Challenges and Future Frontiers*, Conference Proceedings, 8<sup>th</sup> National Health Outcomes Conference, 17-18 July, Australian Health Outcomes Collatoration, Canberra.

10. Ebel, R.L. (1972). Some Limitations of Criterion-Referenced Measurement in *Perspectives in Educational and Psychological Measurement*. Bracht, G.H., Hopkins, K.D., and Standley, J.C. (Eds). Englewood Cliffs, N.J.: Prentice-Hall Inc.
11. Embretson, S. (1983). Construct Validity: Construct Representation versus Nomothetic Span. *Psychological Bulletin*, 93(1), 179-197.
12. Embretson, S. (1996). The New Rules of Measurement. *Psychological Assessment*, 8(4), 341-349.
13. Embretson, S.E. and Reise, S.P. (2000). *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates Inc.
14. Glasziou, P. (2002). Navigating Best Practice in the Information Deluge. *Health Outcomes 2002 – Current Challenges and Future Frontiers*, Conference Proceedings, 8<sup>th</sup> National Health Outcomes Conference, 17-18 July, Australian Health Outcomes Collatoration, Canberra.
15. Hambleton, R.K. (1983). *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.
16. Hambleton, R.K.; Swaminathan, H.; and Jane Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, California: Sage Publications Ltd.
17. Johnston, J.M. and Pennypacker, H.S. (1980). *Strategies and tactics of human behavioural research*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
18. King, M. (2002). Item Response Theory: Applications to health outcomes measurement. *Health Outcomes 2002 – Current challenges and future frontiers*. Conference Proceedings, 8<sup>th</sup> National Health Outcomes Conference, 17-18 July, Australian Health Outcomes Collatoration, Canberra.
19. Kingi, Te K.R. and Durie, M.H. (2000a). *Hua Oranga: A Māori Measure of Mental Health Outcome*. Research Report TPH 00/01. Palmerston North: Massey University, Te Pūmanawa Hauora, School of Māori Studies.
20. Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performance as scientific inquiry into scoring meaning. *American Psychologist*, 9, 741-749.
21. Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher*, 23, 5-12.
22. Murphy, K.R. and Davidshofer, C.O. (2001). *Psychological testing: principles and applications*, 5<sup>th</sup> edition. Upper Saddle River, New Jersey: Prentice-Hall.
23. Palmer, S.K. (2002). *Hei oranga mo ngā wāhine hapū i roto i te whare ora*. Unpublished PhD Thesis, Psychology Department, Waikato University, Hamilton.

24. Peters, J. (1994). *Performance and outcome indicators in the Mental Health Service: a review of literature*. A report prepared for the Ministry of Health, Wellington.
25. Sansoni, J. (1996). The centrality of health outcome measurement. *Integrating Health Outcomes Measurement in Routine Health Care Conference Proceedings*, Canberra: Australian Health Outcomes Clearing House.
26. Sansoni, J. (2002). Australian Health Outcomes collaboration. Challenges and frontiers: a pot pourri. *Health Outcomes 2002 – Current challenges and future frontiers*. Conference Proceedings, 8<sup>th</sup> National Health Outcomes Conference, 17-18 July, Australian Health Outcomes Collatoration, Canberra.
27. Shavelson, R.J., Webb, N.M. and Rowley, G.L. (1989). Generalizability Theory. *American Psychologist*, 44(6), 922-932.
28. Trochim, W. (2002). *Threats to construct validity and Pattern Matching for Construct Validity*. Available at <http://www.prr.msu.edu/trochim/relialt.htm> on 23 July 2002.
29. Van der Linden, W.J. and Hambleton, R.K. (1996). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.