

Psychometrics: An Ancient Construct for Maori

Stephanie Palmer

Massey University, Palmerston North

Capacity to measure the mind and monitor changes in psychological attributes is an ancient and inherent component of classical Maori culture. Within a contemporary context, however, Maori have yet to fully realise the power and potential of psychometric paradigms. As a particular discipline, psychometrics provides methodologies for constructing measurement tools and frameworks for testing whether such tools achieve expected objectives. Psychometric theory provides the rationale for critical analysis and evaluation of assessment tools commonly used on Maori. The advancement of psychometric skills and expertise among present-day Maori will enable the establishment of world class tools that meet the needs and aspirations of Te Ao Maori. The following discussion aims to raise awareness, generate debate and facilitate understanding of psychometric techniques, principles and issues that hold relevance for Maori engaged in the development and use of measurement tools.

Capacity for conceptualisation and measurement of conscious and sub-conscious psychological qualities is evident in classical Maori tradition and culture. The universe itself, for example, is seen to be pure energy eternally engaged in a process of logical progression: i te kore, ki te po, ki te ao marama (Shirres, 1997). Maori pantheon and cosmology is premised upon concepts of ebb and flow between material, psychic and spiritual realms interwoven and influenced by knowledge of Te Kete Aronui, Te Kete Tuauri and Te Kete Tuatea (Marsden, 1975). In retelling of Te Wehenga, the ancient creation story painstakingly describes subtle differences in character, form, disposition and quality. For example, the concepts of te korekore, te korekore-te-rawea, te-korekore-te-whiwhia, te-korekore-te-tamaua and te po-i-tuturi, te-po-i-pepeke, te-po-uriuri, and te-po-tangotango, represent particular

states with discernible purpose, intent and implication (Fitzgerald, 2002; Marsden, 1975). Within the unfolding of the universe there is establishment of hierarchy, relativity and conceptual frameworks for the measurement of difference over aeons.

Evidence of the importance our tupuna rangatira placed on meticulous definition and classification is demonstrated in karakia and other genre. In *Tenei Au*, Ruawhoro from Takitimu refers to Rangi-tu-haha and Tihi-o-manono, both archetypal constructs for distinguishing between multiple levels of vitality, existence, consciousness and enlightenment (Shirres, 1998). In his oriori for Tu-Tere-Moana, Te Matorohanga (1865) of Wairarapa describes the sequential, incremental growth and development of cognition and consciousness during human gestation. Similarly, Enoka Te Pakaru uses the waka wairua of kumara to

describe the methodical implantation of essential human qualities in her oriori *Po Po* (Te Reo Rangatira Trust, 1998).

Around each atua, there are psychological benchmarks for the conceptualisation and manifestation of human potential. For example, Io-matua-kore is the all-encompassing source. Tane is associated with forty-one qualities, each with its own set of attributes and implications, such as Tane-nui-a-rangi, Tane-te-wananga, Tane-matua and Tane-te-waiora. Likewise, the female archetype, Hine is associated with multiple domains, such as, Hine-angiangi, Hine-i-te-korikori, Hine-rauwhangi, Hine-i-te-iwaiwa, Hine-te-hihiri and Hine-i-te-whita.

Whare runanga represent another example of Maori determination to establish systems for classification. Each whare contains spiritual dividers, spatial divisions, symbolic pointers and numerous meaningful artefacts, for example, kopaiti, ihonui, kauwhanga, tahuu, kaho, paepae, rehutai and rukatai (Fitzgerald, 2002). Each division serves its own purpose in terms of discerning difference and grouping people on the basis of physical, intrinsic, seemingly esoteric thresholds or attributes for example, mate/ora, tangata whenua/manuhiri, rangatira/ringawera, tane/wahine, puhi/pakeke, kauwae runga/kauwae raro and tapu/noa. The whare runanga provides a pragmatic system for classifying and identifying relative difference between people. Adherence to the system promises collective benefit. It helps to determine personal

role, function and responsibility in relation to the group. It helps to shape personal attitude, ethics and values as well as consolidate human capacity and capability, most importantly, the capacity for creation, integration and transmission of knowledge.

The establishment of systems for measurement of intrinsic, interpersonal, psychological qualities is exemplified also in *te reo Maori*. That is, the structure, semantics, composition and constructs of Maori language itself. Taina and Hariata Pohatu (2002) use the word *ata* to demonstrate the notion of conscious movement. In terms of interpersonal dynamics and relationships this small kupu contains socialisation, behavioural and transformative possibilities. These authors suggest that the information obtained from *ata* provides guidance on relationship boundaries and appropriate behaviour in terms of space, amount of energy to invest, need for respectfulness and level of reciprocity. As a measurement tool, *ata* can be used to monitor the quality of relationships, safeguard against disadvantage or injury and inform strategic planning.

The primordial concept of *mauri* has a similar, but distinct, function (Pohatu & Pohatu, 2003). *Mauri* is the unique life-force, the vitality, source and essential energy that drives existence, aliveness and being. *Mauri* is an intangible quality, but powerful yardstick for monitoring the integrity of engagement. *Mauri* is the sentinel of capacity and opportunity to realise potential. It is calibrated on a bipolar continuum that ranges from *mauri ora* to *mauri mate* with various stages in between: *mauri tu*, *mauri oho* and *mauri noho*, for example, each aiming to provide information about the integrity of interaction, the quality of processes, the need for resolution, improvement and change.

Maori culture revolved around the socialisation, transmission and refinement of mechanisms for measurement of intrinsic psychological qualities. Within a contemporary context Maori have been quick to realise the potential of psychometric pathways and paradigms (Dewes, 1995; Jones, 1995; Mead, 1995; Murchie, 1984; Parata, 1995; Puketapu, 1995; Ramsden, 1995; Walker, 1995). There are at least four reasons why the development

of psychometric expertise would be particularly useful for Maori.

Firstly, the discipline contains methodologies that will help Maori to evaluate, monitor and demonstrate the advantages and benefits of participation in *te ao Maori*. Indeed, the collection of such data is consistent with political demand for evidence-based practice and use of assessment tools, outcome measures and performance indicators, particularly in health and education. Within mental health services, for example, Puahou proposes the need for strategies to collect data about cultural enhancement, active participation in Maori centred values and beliefs (cited in Durie, 1998b). Others have identified similar needs (Galvin, 1998; Health Research Council of NZ, 2003; Ministry of Education, 2003; Ministry of Health, 2002; Te Puni Kokiri, 2001; Walker, 1998).

Secondly, the use of psychometric techniques can help Maori explore, define, describe and identify culturally relevant concepts and constructs integral to kaupapa Maori and/or Maori centred research designs (Bevan-Brown, 1998; Cram et al, 2002; Crengle, 1998; Durie, 1998a; Henare, 2000; Jackson, 1998; Jahnke, 1997; Keefe et al, 1999; Lawson-Te Aho, 1998; Maynard, 1999; Mead, 1998; Ministry of Education, 1996; Pitama et al, 2002; Potiki, 2000; Smith, 1999; Walsh-Tapiata, 1998). Sharples (2001), for example, is among those who have called for research methodologies that validate the authenticity of Maori pedagogy and *kawa*.

Thirdly, psychometric methodologies provide empirical pathways for the development of theory and accumulation of knowledge about theoretical mechanisms which may underpin the burgeoning number of models and paradigms put forward to explain or conceptualise Maori cognition, behaviour and social phenomena (Jahnke, 2001; Kingi & Durie, 2000; Pere, 1991; Reedy, 2000; Royal, 1998; Takino, 1998). Using knowledge of Maori culture as a reference point, Tukukino, for example, has presented a model for the classification of Maori identity on a continuum ranging from marginal through to emergent, adaptive or traditionalist (1989). Similarly, Putangitangi and Te Hoe Nuku Roa have

been developed for the measurement of Maori identity (explained in Davies et al, 2002; Durie, 1995; Durie et al, 2002).

Fourthly, an understanding of psychometric theory and principles can provide the foundation and rationale for critical analysis of measurement tools and the skills to accept, reject and/or build upon the information and outcomes of techniques for assessment of Maori (Alpass, Neville & Flett, 2000; Bennett & Flett, 2001; Brough & Kelling, 2002; Brown et al, 2002; Goulton, 1998; Hirini & Flett, 1999; Oliver & Brough, 2002).

The motivation for writing this paper is derived from involvement in the construction of a tool to measure *waiora* among Maori (Palmer, 2002, 2004). The main objectives are to improve awareness and understanding of psychometric principles, generate interest in the use of psychometric methodologies and provide a platform for critical analysis of tools for the assessment and measurement of Maori. The remaining discussion aims to outline the key components of psychometric theory, namely classical test theory, reliability theory, generalizability theory, validity theory and item analysis theory. It seeks to enhance understanding of the principles and rationale that underpin each theory and provide some examples of commonly used techniques.

Te Hangaitanga

Andrews, Peters & Teesson (1994) have said that tools for the measurement of health outcomes should be applicable, acceptable, practical, reliable, valid and sensitive to change. Sansoni (1996) added that a good outcome measure should demonstrate reliability, validity, discriminatory power, responsivity, practical utility, freedom from confounding factors, relevance of application and appropriate mode of administration.

However, the meaning of reliability and validity has proven difficult to define. Substantial research has shown that reliability varies with the purpose of the test, the particular attributes being measured and the circumstances in which it is applied (Aitken, 1997; Cohen & Swerdlik, 1999; Moss, 1994; Murphy & Davidshofer, 2001). Thus, a test may

be reliable for some purposes but not others. Furthermore, reliability is said to be a necessary but not sufficient condition for validity, suggesting that an unreliable test cannot possibly be valid but a valid test is not necessarily reliable. In other words, validity has pre-eminence over reliability (Anastasi & Urbina, 1997; Johnston & Pennypacker, 1980).

The meaning of validity has been debated for decades and is also known to vary with the context and purpose of tests (Bracht, Hopkins & Stanley, 1972; Johnson & Pennypacker, 1980). Popular definitions of validity are said to be confusing and unsatisfactory (Embretson, 1983; Messick, 1995). For example, Andrews, et al (1994) and Sansoni (1996) would not be able to distinguish the meaning of validity from qualities of acceptability, applicability, user-friendliness, practical utility, mode of administration, relevance of application and even freedom from confounding factors. Lack of a clear definition has led several to propose the notion of construct validity as an overarching concept to capture the many characteristics of validity (Embretson, 1983; Messick, 1995; Murphy & Davidshofer, 2001).

Embretson (1983) suggests the demonstration of construct validity requires a paradigm shift towards modern psychometric theory and methodologies that gradually accumulate information from a variety of sources. Such methodologies are seen to provide opportunities for task decomposition and theory construction which facilitate understanding of individual differences, the processes that underlie behaviour and the way in which test scores may interact with a universe of related variables (Anastasi & Urbina, 1997; Embretson, 1983; Murphy & Davidshofer, 2001).

Though never widely accepted, the terms *vaganotic* and *idemnotic* help to describe key differences between classical and modern psychometric methodologies (Johnston & Pennypacker, 1980). The *vaganotic* tradition is based on assumptions of bell-shaped distribution and the notion that error, or variability from true ideals, will reduce with more measurements. This school of thought

uses variability to measure phenomena. It aims to reduce variability and holds that large sample sizes ensure representativeness or the generality of data to even larger populations. Such inferential statistics are well suited to the study of population behaviour but not the behaviour of an individual.

In contrast, *idemnotic* techniques use variability to study behaviour at an individual level. This approach measures objective differences in absolute quantities, for example, reaction time, the number of trials or the number of correct items. Various techniques, like standardization and or the anchoring of data, provide mechanisms for comparing different data sets. The overall objective is to establish the generality, or generalisability, of data beyond test conditions. *Idemnotic* theory assumes that a small number of subjects will provide sufficient data to demonstrate functional relationships which are a pre-requisite for generality to a larger group.

Vaganotic theory has underpinned the development and use of classical psychometric methods whereas modern psychometric theory, notably, item response theory, is firmly based on *idemnotic* principles.

Classical Test Theory

Classical test theory is based on two principles that have provided a powerful foundation for the measurement of psychological phenomena (Aitken, 1997; Anastasi & Urbina, 1997; Cohen & Swerdlik, 1999; Murphy & Davidshofer, 2001; Trochim, 2002).

The first principle assumes a distribution of scores will always resemble the so-called normal, bell-shaped curve. The larger the group, or sample size, the more closely a distribution will approximate this bilaterally symmetrical curve. This assumption underlies the rationale for many techniques to describe and interpret test data. In particular, it allows individual and aggregate test scores to be compared in terms of 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 is 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 assumption of true score theory or classical test theory, is that every measurement, or test score (X), is a composite of two components: true score (T) and random error (e_x). In terms of variance (var) across a set of scores, it is assumed $\text{var}(X) = \text{var}(T) + \text{var}(e_x)$. Hence, the variability of a measure 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. The notion of error is central to the establishment of reliability.

Reliability Theory

The basic tenet of reliability theory is that a test score reflects two factors: those that contribute 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 error.

In accordance with classical test theory, it is assumed 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 any individual, the reliability of scores for this measure is represented by the ratio of true scores to true score plus error scores. However, reliability is a characteristic of responses taken across individuals and this ratio is, therefore, expressed in terms of variance, that is, $\text{var}(T) : \text{var}(X)$.

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 $\text{var}(T)$ 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 standard deviations. The resultant coefficient reflects the degree of consistency between two independently derived sets of scores. Reliability is always expressed as a correlation co-efficient. Three assumptions underpin this approach: (1) the mean error of measurement will always sum to zero¹; (2) true scores and errors are not correlated, and (3) there is no correlation between errors on different measures. In general, the reliability coefficient is defined as the ratio of true score variance to total variance of test scores.

The third goal of reliability theory is to improve test scores by suggesting ways to minimize error. Classical theory has five main methods to calculate the reliability of test scores and each provides different information about the source of error (Anastasi & Urbina, 1997; Murphy & Davidshofer, 2001; Trochim, 2002). The rationale underlying each approach is that two equivalent, or parallel, forms of a test will produce a 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. An alternate forms approach looks at the consistency of scores when different tests, constructed from the same content, 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 average inter-correlation among test items. For example, Chronbach's co-efficient α estimates the mean reliability coefficient obtained when each item is

compared with every other item. This method looks at consistency between item scores or the extent to which each item represents the things being measured by other test items.

Reliability co-efficients are used to calculate the standard error of measurement, which determines the preciseness of test scores, confidence levels in a test score and the extent to which scores are likely to vary from one administration to another. In general, the more items in a test, and the higher the correlation between test items, the higher the reliability. For this reason, composite scores are usually most reliable.

Generalisability Theory

There are two main criticisms of reliability theory: (1) the assumption that error is always random; and, (2) both measurement error and the reliability co-efficient can vary, thus neither are stable characteristics of a test.

Generalisability theory has extended reliability theory in a number of ways (Embretson 1996; Embretson & Reise, 2000; Murphy & Davidshofer, 2001; Shavelson, et al 1989). In particular, it recognises that measurement error may not always be random. Generalisability theory has divided the error component of 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. For this reason, random error is seen to be a nuisance variable and is often called noise. In contrast, systematic error, also called bias, will affect the average performance of a group because it has a systematically positive, or negative, effect on responses.

The main concern with generalisability theory is the extent to which the results obtained from one measure can be generalised 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 variability can be attributed to systematic or unsystematic (random) sources.

Murphy & Davidshofer (2001) suggest the main advantage of

generalisability theory is conceptual because each test score is seen to be a single sample from a universe of possible scores. Generalisability theory encourages the identification of situations in which a test might be reliable, or unreliable, it provides a context for thinking about test results. In this light, reliability becomes a characteristic of the use of test scores rather than a characteristic of the scores themselves. Whereas, classical techniques can be useful under highly standardised conditions, generalisability theory is more appropriate when conditions are likely to effect test scores or test scores are used for different purposes. As a method for estimating reliability, generalisability theory is expected to replace older methods.

Validity Theory

Validity is usually associated with the meaning of test scores and notions of correctness and freedom from bias. Validity is used to determine how well a test measures the concept, or construct, it sets out to measure. In general, this is estimated through analysis of test content, correlation between test scores and criteria or investigation of underlying properties.

Trochim (2002) suggests the need to consider the validity of operationalisations, rather than the validity of measures. In his view, it is the way in which a concept is translated into a functioning, operating reality that is really of interest. 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). They claim the various techniques for measurement of validity provide information about different aspects of construct validity not different types of validity. Collectively, it is argued, the various types of validity measures are not mutually exclusive and all have a bearing on construct validity.

There are three slightly differing perspectives on the underlying approach to collection of data about construct validity. Firstly, Trochim (2002) claims techniques for the measurement of construct validity

provide information on translation validity or criterion-related validity. The term translation validity was coined to describe techniques that aim to show the construct is well defined. Methodologies for the assessment of face and content validity are seen to serve this purpose well. That is:

- face validity looks at whether the physical appearance and content of a test has relevance for intended respondents and administrators
- content validity considers whether the content of a test is representative of the conceptual domain it is intended to measure.

Criterion-related validity looks at whether the operationalisation behaves as it should. A standard, or criterion, is used to test whether the item, that is the operationalisation of a construct, functions in a predictable way. The rationale is relational and acknowledges that change from one construct to another may be gradual. That is, a construct is not necessarily one thing or another but may be a mixture of inter-related concepts. Criterion-related validity aims to provide information about the correlation between constructs, in the form of a correlation, or validity, coefficient. The four main techniques for assessment of criterion-related validity provide information on:

- predictive validity - ability to predict something the construct should theoretically predict, such as an increase or decrease in scores with age or experience
- concurrent validity - ability to distinguish between variables and concepts that should theoretically be different
- convergent validity - that theoretically similar constructs should correspond or converge
- discriminant/divergent validity - degree of difference from theoretically distinct variables or constructs.

Secondly, Murphy & Davidshofer (2001) suggest that pathways towards construct validity provide information on the validity of measurement or the validity of decisions. In their view, content-orientated methodologies provide information on the validity

of measurement, that is, whether the test measures what it is supposed to measure. Such methodologies aim to describe the content domain, determine which part of the content domain is measured by each item and compare the structure of a test to the structure of content domain. Internal consistency, for example, provides information about content validity.

In contrast, construct-orientated methodologies provide information on the validity of decisions, that is, whether the test helps to make correct or accurate decisions. Such methodologies will aim to identify behaviours that relate to either the construct being measured or similar constructs. The main techniques for gathering information about this type of construct validity are correlations such as those which test for predictive or concurrent validity, factor analysis and/or experimental manipulation of the construct being measured.

Embretson (1983) presents a third approach to the measurement and conceptualisation of construct validity. In her view, methodologies for the collection of data about construct validity should provide information about construct representation and nomothetic span. Construct representation aims to identify theoretical mechanisms that underlie task performance. Nomothetic span looks at the extent to which a construct may be embedded in a network of relationships with other constructs. In general, nomothetic span provides information on the utility, or usefulness, of a construct as a measure of individual difference. For example, an individual's test scores may be correlated against ethnicity, gender, cultural identity, age, socio-economic status or any other socially important criteria to check theoretical assumptions about the construct being measured.

Embretson (1983) believes construct representation and nomothetic span are interactive phases of the construct validation process. That is, construct representation informs the development of nomothetic span. Furthermore, this approach to construct validity incorporates both classical and modern psychometric techniques. More specifically, the modern methodologies,

particularly item response theory, provide the most effective techniques for gathering information about construct representation whereas nomothetic span can be identified through conventional methods for structural equation modeling.

Item Analysis Theory

No matter which way it is approached, the reliability, validity and construct validity of any test depends upon the quality of items. For example, the reliability or validity of a test may be limited because items are poorly worded or poorly designed. Item analysis aims to identify ineffective items and improve the psychometric properties of a test. Classical approaches to item analysis use two main techniques to look at individual differences in the pattern of responses to each item: item difficulty and item discrimination (Aitken, 1997; Anastasi & Urbina, 1997; Murphy & Davidshofer, 2001).

Item difficulty (p) looks at the relative frequency of correct responses among those taking the test. When all of the items in a test are difficult or easy, there will be little variability. In terms of central tendency and the normal curve, item difficulty is therefore 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 of scores at the upper end suggests a low ceiling because the test lacks difficult items. Such skewness makes it impossible to measure individual differences.

Three techniques are mainly used to measure the discriminating power of an item. Firstly, the D statistic is based upon the rationale that a good item will discriminate between those with high and low scores. Secondly, item-total correlations consider the strength of correlations between item score and total test scores. Thirdly, inter-item correlations consider the correlations between all test items and sort out the items that are internally consistent. Inter-item correlations also provide information on predictor validity or item ability to predict an external criterion.

Items selected on the basis of a D statistic will yield a different kind of test than one composed of items selected because of correlations with an external criterion. The method for determining discriminatory power therefore depends on the purpose of the test. When external validity is important, inter-item correlations are most useful. When internal consistency is important, the D statistic is relevant although, in some instances, a combination of the two approaches may be appropriate.

Item Response Theory

For a number of psychometricians, conventional item analysis has too many limitations, notably:

- responses are sample dependent or influenced by the types of people who take the test
- large sample sizes are needed to test each item
- longer tests are more reliable than short forms
- tests are static and items cannot be changed which means the results obtained from different versions of a test cannot be compared
- acceptable p and D values do not guarantee that an item will function effectively in all situations
- reliability and validity is based on overall test scores rather than response to individual items.

For these reasons, item response theory is providing an increasingly popular alternative to conventional item analysis (Anastasi & Urbina, 1997; Embretson, 1983, 1996; Embretson & Reise, 2000; King, 2002; Hambleton, Swaminathan & Rogers, 1991; Murphy & Davidshofer, 2001; van der Linden & Hambleton, 1996). The rationale for item response theory is that each item measures a specific attribute and the more of the attribute a person has the more likely it will be that they answer the item correctly. Item difficulty and discrimination is therefore described in terms of the relationship between attribute and response. Item response theory is based on four main assumptions:

- performance is explained by a set of factors called latent traits, or abilities

- only one ability is measured by each set of items in a test, each test is unidimensional
- when ability is held constant there is no relationship between item and response, ability is the only factor that influences response
- the relationship between item performance and underlying ability is described by a monotonically increasing function called an item characteristic curve.

An item characteristic curve 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 provides a single graph summary of item difficulty, discriminatory power and the probability of answering correctly by guessing. In addition, the item characteristic curve can provide information on how the item is functioning across a range of criteria.

For psychometricians, the emergence of item response theory has introduced new rules for measurement that have a number of advantages over older methods. Furthermore, Murphy and Davidshofer (2001) believe item response theory has important conceptual advantages because it encourages researchers and test administrators to think about why people respond the way they do. The chief limitation of item response theory is simply that the discipline is still developing. Support services and mentorship, for example, may be hard to find and the software is stand alone, which means it is not yet compatible with popular statistical packages like SAS or SPSS. However, the accessibility of item response theory is rapidly improving (Baker, 2001).

Hei Mutunga

The above discussion has outlined key principles in psychometric theory and some areas of debate. In summary, classical test theory, along with assumptions of normal distribution and random error, underpin the use of most psychometric techniques including methodologies for estimating reliability, validity, confidence intervals and variance or standard deviation.

Concepts of reliability and validity are neither gold standards nor constant.

Both have proven difficult to define and estimates are known to vary with test context, purpose or participants. Several techniques provide information about reliability, or consistency of scores, but generalisability is the more relevant construct and validity is the more sensible place to start.

The notion of construct validity is set to replace more conventional approaches to assessment of validity and involves the gradual accumulation of knowledge from a range of sources including different validity co-efficients. Both classical and modern psychometric methodologies are used to gather information about construct validity

Finally, item analysis is the most important component of psychometric design. Modern techniques are quickly replacing older methods and item response theory offers a particularly attractive alternative. The item characteristic curve captures information about item difficulty, item discrimination, response guessing and criterion validity on a single graph.

Development of Maori psychometric expertise will serve to increase opportunities for engagement and participation in Te Ao Maori. It is argued that the principle of psychometric measurement has been an inherent component of Maori tradition and culture. Within the context of Maori health assessment, an understanding of psychometric theory and principles, particularly construct validity and item analysis, will generate frameworks for critical analysis of mainstream tools and help identify pathways for integration of matauranga Maori content. The development of matauranga Maori based health assessment tools will foster understanding of Maori concepts and constructs, assist the socialisation of Maori world-views and help to shape the values of Te Ao Hou.

Although psychometric methodologies may have benefits for Maori, there are limits to its use. Marsden (1975) helped define that boundary when he said that the main distinction between conventional science and a matauranga Maori approach to science is know-how and know-why. Conventional science is seen to have a know-how approach to knowledge. As long as there is know-how the knowledge will be pursued,

no matter what the consequences. In contrast, matauranga Maori science is driven by know-why. That is, the reasons for seeking knowledge must be irrefutable. In other words, some forms of knowledge are sacred, dangerous and best protected until know-why is understood.

He mana te matauranga.

References

- Aitken, L.R. (1997). *Psychological Testing and Measurement*, 9th edition. Needham Heights, MA: Allyn & Bacon.
- Alpass, F., Neville, S., & Flett, R. (2000). Contribution of retirement-related variables to wellbeing in an older male sample. *New Zealand Journal of Psychology*, 29(2), 74-79.
- Anastasi, A. & Urbina, S. (1997). *Psychological testing*, 7th edition. New Jersey: Prentice Hall.
- Andrews, G.A., Peters, L. & 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.
- Baker, F.B. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment & Evaluation, University of Maryland, College Park, M.D. Available at <http://ericae.net/irt/baker/> on 3 June 2001.
- Bennett, S. & Flett, R., (2001). Te Hua o te Ao Maori. *He Pukenga Korero*, 6(2), 29-34.
- Bevan-Brown, J. (1998). By Maori, For Maori, About Maori – is that enough? *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Bracht, G.H., Hopkins, K.D. & Stanley, J.C. (1972). *Perspectives in Educational and Psychological Measurement*. Englewood Cliffs, New Jersey: Prentice-Hall Inc.
- Brough, P. & Kelling, A. (2002). Women, work and wellbeing: the influence of work, family and family work conflict. *New Zealand Journal of Psychology*, 31(1), 29-38.
- Brown, J., Jose, P., Ng, S.H. & Guo, J. (2002). Psychometric properties of three scales of depression and wellbeing in a mature New Zealand sample. *New Zealand Journal of Psychology*, 31(1), 39-47.
- Cohen, R.J. & Swerdlik, M.E. (1999). *Psychological Testing and Assessment: an introduction to tests and measurement*, 4th edition. California: Mayfield Publishing Company.
- Cram, F., Pihama, L., Jenkins K. & Karehana, M. (2002). *Evaluation of Programmes for Maori Adult Protected Persons under the Domestic Violence Act 1997*. Ministry of Justice, Wellington.
- Crengle, S. (1998). Ma Papatuanuku, ka Tipu nga Rakau. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Davies, S., Elkington, A. & Winslade, J. (1994). Putangitangi: a model of cultural identity. *Perspectives on Counselling 1031.330 Reading Guide*. Hamilton, Department of Education, Waikato University.
- Dewes, K. (1995). Principles for Resource Distribution. *Proceedings of the Hui Whakapumau Maori Development Conference*. Department of Maori Studies, Massey University.
- Durie, A. (1998a). Me Tipu Ake te Pono: Maori Research, Ethicality and Development. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Durie, M.H. (1995). Te Hoe Nuku Roa Framework. *Journal of the Polynesian Society*, 104(4), 462-470.
- Durie, M.H. (1998b). Puahou: a five part plan for improving Maori mental health. *He Pukenga Korero*, 3(2), 61-70.
- Durie, M., Fitzgerald, E., Kingi, T.K., McKinley, S. & Stevenson, B. (2002). *Te Hoe Huku Roa: Maori Specific Outcomes and Indicators*. A report prepared for Te Puni Kokiri, The Ministry of Maori Development by Te Putahi-a-Toi, Massey University, Palmerston North.
- Embretson, S. (1983). Construct Validity: Construct Representation versus Nomothetic Span. *Psychological Bulletin*, 93(1), 179-197.
- Embretson, S. (1996). The New Rules of Measurement. *Psychological Assessment*, 8(4), 341-349.
- Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates Inc.
- Fitzgerald, S. (2002). *The Whare Runanga: The House of Learning*. Maori Studies. Palmerston North, Massey University.
- Galvin, P. (1998). Te Puni Kokiri Agency Review Methodology: an examination of Government departments's provision of services to or for Maori. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Goulton, F., (1998). He Huarahi Ako: Pathways to learning the academic and cultural self-efficacy of Maori student teachers. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Hambleton, R.K., Swaminathan, H., & Jane Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, California: Sage Publications Ltd.
- Health Research Council of NZ (2003). Alcohol and Drug Outcomes Project (ADOPT). *HRC Newsletter*, 43, 7.
- Henare, D. (2000). Millennial Thoughts on Maori Development. *He Pukenga Korero*, 5(2), 21-32.
- Hirini, P. & Flett, R. (1999). Aspects of the Maori All Black Experience: the value of cultural capital in the new professional era. *He Pukenga Korero*, 5(1), 18-24.
- Jahnke, H. (1997). Towards a theory of mana wahine. *He Pukenga Korero*, 3(1), 27-36.
- Jahnke, H. (2001). Navigating the Education Workplace: a Maori Centred Approach to Researching Maori Women in Educational Organisations. *He Pukenga Korero*, 6(2), 9-17.
- Jackson, M. (1998). Research and the Colonisation of Maori Knowledge. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Johnston, J.M. & Pennypacker, H.S. (1980). *Strategies and tactics of human behavioural research*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Jones, S. (1995). Te Ara Whakatuputupu. *Proceedings of the Hui Whakapumau, Maori Development Conference*. Department of Maori Studies, Massey University.
- Keefe, V., Ormsby, C., Robson, B., Reid, P., Cram, F., Purdie, G. & Ngati Kahungunu Iwi Authority Inc. (1999). Kaupapa Maori meets Retrospective Cohort. *He Pukenga Korero*, 5(1), 12-17.
- King, M. (2002). Item Response Theory: Applications to health outcomes measurement. *Health Outcomes 2002 – Current challenges and future frontiers*. Conference Proceedings, 8th National Health Outcomes Conference, 17-18 July, Australian Health Outcomes Collatoration, Canberra.
- Kingi, Te K.R. & Durie, M.H. (2000). Hua Oranga: A Maori Measure of Mental Health Outcome. *Research Report TPH 00/01*. Palmerston North: Massey University, Te Pumanawa Hauora, School of Māori Studies.
- Lawson-Te Aho, K. (1998). Maori Youth Suicide – Colonisation, Identity and Maori Development. *Proceedings of Te*

- Oru Rangahau, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Marsden, M. (1975). God, man and universe: a Maori view. In King, Michael (ed). *Te Ao Hurihuri*, pp 191-219, Hicks, Smith & Sons, Wellington.
- Maynard, K. (1999). Kimihia: Maori culture-related needs – seeking more effective ways to assess and address Maori offending. *He Pukenga Korero*, 5(1), 25-33.
- Mead, A. (1995). Maori Leadership: the waka tradition – the crews were the real heroes. *Proceedings of the Hui Whakapumau, Maori Development Conference*. Department of Maori Studies, Massey University.
- Mead, A. (1998). Sacred Balance. *He Pukenga Korero*, 3(2), 22-27.
- 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.
- Ministry of Education. (1996). *Te Whariki: Early Childhood Curriculum*. Ministry of Education, Wellington.
- Ministry of Education. (2003). *Education Priorities for New Zealand*. Ministry of Education, Wellington.
- Ministry of Health. (2002). *He Korowai Oranga: Maori Health Strategy*. Ministry of Health, Wellington.
- Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher*, 23, 5-12.
- Murchie, E. (1994). *Rapuora Health and Maori Women*. Wellington, Lincoln Print.
- Murphy, K.R. & Davidshofer, C.O. (2001). *Psychological testing: principles and applications, 5th edition*. Upper Saddle River, New Jersey: Prentice-Hall.
- Oliver, J. & Brough, P. (2002). Cognitive appraisal, negative affectivity and psychological wellbeing. *New Zealand Journal of Psychology*, 31(1), 2-6.
- Palmer, S.K. (2002). *Hei Oranga mo nga Wahine Hapu i roto i te Whare Ora*. PhD Thesis, Psychology Department, Waikato University, Hamilton.
- Palmer, S. K. (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* 33(2): 50-59.
- Parata, R. (1995). Maori Investments for the Future. *Proceedings of the Hui Whakapumau, Maori Development Conference*. Department of Maori Studies, Massey University.
- Pere, R. (1991). *Te Wheke – a celebration of infinite wisdom*. Gisborne: Ako Global Learning.
- Pitama, D., Ririnui, G. & Mikaere, A. (2002). *Guardianship, Custody and Access: Maori Perspectives and Experiences*. Ministry of Justice, Wellington.
- Pohatu, T.W. & Pohatu, H.R. (2002). *Ata: Growing Respectful Relationships*. Maori Studies. Auckland Institute of Technology.
- Pohatu, T.W. & Pohatu, H.R. (2003). *Mauri: Re-thinking Human Wellbeing*. Maori Studies. Auckland Institute of Technology.
- Potiki, T. (2000). A Traditionalist Approach to Iwi Government. *He Pukenga Korero*, 5(2), 51-58.
- Puketapu, B. (1995). Hokia ki te Kopae a nga Pahake: a classical Maori journey. *Proceedings of the Hui Whakapumau, Maori Development Conference*. Department of Maori Studies, Massey University.
- Ramsden, I. (1995). Maori Policy, Maori and Government Objectives. *Proceedings of the Hui Whakapumau, Maori Development Conference*. Department of Maori Studies, Massey University.
- Reedy, T.M. (2000). Ko te Huringa i te Mahara. *He Pukenga Korero*, 6(1), 7-14.
- Royal, Te A.C. (1998). Te Ao Marama – a research paradigm. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Sanson, J. (1996). The centrality of health outcome measurement. *Integrating Health Outcomes Measurement in Routine Health Care*. Conference Proceedings, Canberra: Australian Health Outcomes Clearing House.
- Sharples, P. (2001). *Kaupapa Maori Health Research*. Presentation at Hui Whakatipu, Tame-te-Kapua Marae, Rotorua. Maori Health Research Council, Health Research Council of New Zealand, Auckland.
- Shavelson, R.J., Webb, N.M. & Rowley, G.L. (1989). Generalizability Theory. *American Psychologist*, 44(6), 922-932.
- Shirres, M.P. (1997). *Te Tangata: the human person*. Auckland: Accent Publications.
- Smith, L.T. (1999). *Decolonizing Methodologies: Research and Indigenous Peoples*. Dunedin: University of Otago Press.
- Takino, N.M. (1998). Academics and the Politics of Reclamation. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Te Reo Rangatira Trust (1998). *He Waiata Onamata - Songs from the Past*. Wellington: Huia Publishers
- Te Matarohanga, M. (1865). *Te Kauwae Runga*. Transcription written by H.T. Whatahoro. Hamilton: Te Ataranga o Waikato Polytechnic.
- Te Puni Kokiri. (2001). *Strategic Plan 2001/02 – 2003/04*. Te Puni Kokiri, Wellington.
- 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.
- Tukukino, H. (1989). Presentation at Totara Toa Hui, *In Search of Hauora*, Taupo, 3-5 March.
- van der Linden W.J. & Hambleton, R. K. (1996). *Handbook of Modern Item Response Theory*. New York, Springer-Verlag.
- Walker, R. (1995). Maori Resistance to State Domination. *Proceedings of the Hui Whakapumau, Maori Development Conference*. Department of Maori Studies, Massey University.
- Walker, V. (1998). Quality Evaluation for Maori: overcoming difficulties. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.
- Walsh-Tapiata, W. (1998). Research within you own iwi – what are some of the issues. *Proceedings of Te Oru Rangahau*, Te Putahi-a-Toi, School of Maori Studies, Massey University, Palmerston North.

Note

1. 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 errors will equal zero.

Author Note:

Stephanie Palmer
Te Aitanga-a-Mate, Te Whanau-a-Iritekura, Te Whanau-a-Rakairoa
Erihapeti Rehu-Murchie Post-Doctoral Research Fellow
Te Pumanawa Hauora

Address for correspondence:

Stephanie Palmer
School of Psychology
Massey University
PO Box 11-222
Palmerston North.
waea: (04) 801 5799 extn 6028
Email: S.K.Palmer@massey.ac.nz