Validity and Reliability in a nutshell

Article *in* Journal of Clinical Nursing · December 2009

DOI: 10.1111/j.1365-2702.2009.02939.x · Source: PubMed

Katrina Bannigan
University of Plymouth
66 PUBLICATIONS **421** CITATIONS

SEE PROFILE

Roger Watson
University of Hull
570 PUBLICATIONS **6,259** CITATIONS

SEE PROFILE

METHODOLOGICAL PAPER

Journal of
Clinical Nursing

Reliability and validity in a nutshell

Katrina Bannigan and Roger Watson

Aims. To explore and explain the different concepts of reliability and validity as they are related to measurement instruments in social science and health care.

Background. There are different concepts contained in the terms reliability and validity and these are often explained poorly and there is often confusion between them.

Design. To develop some clarity about reliability and validity a conceptual framework was built based on the existing literature.

Results. The concepts of reliability, validity and utility are explored and explained.

Conclusions. Reliability contains the concepts of internal consistency and stability and equivalence. Validity contains the concepts of content, face, criterion, concurrent, predictive, construct, convergent (and divergent), factorial and discriminant. In addition, for clinical practice and research, it is essential to establish the utility of a measurement instrument.

Relevance to clinical practice. To use measurement instruments appropriately in clinical practice, the extent to which they are reliable, valid and usable must be established.

Key words: nursing, questionnaires, reliability, stability, utility, validity

Accepted for publication: 22 June 2009

Introduction

In the field of questionnaire development and psychometrics the concepts of reliability and validity are both central and crucial to the development of measurement instruments that can be useful in research, clinical practice or education where they are applied. In the physical world we generally take the reliability and validity of measurements for granted: when we measure weight, length, height and temperature, we usually have a familiar standard and even personal experience against which to compare measurements. For example, if we are told that someone is 2 m tall we know that this is both credible and verifiable and we also know what it means relative to our own height and the height of other familiar objects. Also, if we meet the person – unless we have been widely misinformed or someone has made a terrible mistake is measuring them – then we can see for ourselves that the person in question is, indeed, 2 m tall. Within limits, the same applies for other anthropomorphic measures and for these measures applied to familiar objects such as buildings and vehicles, for example. Moreover, in the physical world, where we depend so much on measurements to be accurate for purchasing goods, cooking food and travelling, there are indeed verifiable physical standards against which many measures may be compared and government departments concerned with the honest and consistent application of, for example, weights and measures, to ensure that trade and commerce ensue fairly.

However, in the world of psychometrics, where measures are of less tangible entities such as educational attainment, psychological state and clinical conditions, we are usually in unfamiliar territory and dependent on measures which, at

Authors: *Katrina Bannigan*, PgCLTHE, PhD, FHEA, Reader in Occupational Therapy/Director of the Research Centre for Occupation and Mental Health, Faculty of Health and Life Sciences, York St John University, York, UK; *Roger Watson*, FIBiol, FRCN, FAAN, Professor of Nursing, School of Nursing and Midwifery, The University of Sheffield, Sheffield, UK

Correspondence: Katrina Bannigan, Reader in Occupational Therapy/ Director of the Research Centre for Occupation and Mental Health, Faculty of Health and Life Sciences, York St John University, York YO31 7EX, UK. Telephone: +44 (0) 1904 876793. **E-mail:** k.bannigan@yorksj.ac.uk

doi: 10.1111/j.1365-2702.2009.02939.x

best, are mere proxies, or worst, are inaccurate and misleading. Thus, while reliability and validity are no less crucial in the physical world, they are relatively easily attained; in psychometrics they are also crucial but, it could be said, are rarely attained and, when they are, are very 'hard won'.

The details of the different concepts of reliability and validity (and utility) will be described in detail below. Suffice to say here that an understanding of the relationship between reliability and validity is necessary to understand both why these concepts are worth exploring but also why they are both necessary. A cursory glance at any research methods textbook will explain that reliability is the extent to which a measure is the same each time it is performed and by whoever performs it. In the physical world this would, for example, mean that a pole purporting to measure 1 m in length will always be that length each time it is used. Thus the 1 m pole is completely reliable; however, this is no guarantee that it is a valid measure of 1 m in length: it could be reliably wrong. The only way to verify (validate) that it is 1 m long would be to use a standard measure of 1 m with which to compare it. Grasping the above fundamentals leads to a complete understanding of the concepts of reliability and validity and their relationship: validity is totally predicated upon reliability and reliability in itself is insufficient. This is all that needs to be known about psychometrics.

Nevertheless, the frequency with which the above is reiterated in research textbooks seems to do little, in our experience, to dispel confusion about the concepts of reliability and validity; especially among research students. This confusion is compounded by the frightful array of terminology that underpins the field of psychometrics and – without wishing to exemplify – it is our experience that different textbooks and different fields present slightly different definitions and explanations of the different concepts thereby confounding the difficulties that students – and some academics – find in this field.

Thus the present paper: the text is based on a significant section of a chapter from KB's PhD thesis which RW supervised. In an effort to clarify the field of psychometrics for herself (and RW) KB wrote this section – which we reproduce largely unchanged from the original – and used it to preface the relevant chapter of her thesis. This comes to be published in this form at this time due to RW having used the section repeatedly with subsequent PhD students as an alternative to a lengthy explanation or an even longer reading list. Invariably the response from students is positive ranging from it being helpful to the hagiographic and urging its publication. Neither KB nor RW believe that this is the last word on reliability and validity but we share it and welcome comment and debate. Notwithstanding a relatively recent and excellent paper in this field (De Von *et al.* 2007) we believe this paper – whatever its veracity – to be one of the most complete and circumscribed accounts in the field… reliability and validity in a nutshell.

Methods

In any research study the theoretical basis of the study can serve as a sound foundation on which to build data collection and data analysis methods (Polit & Hungler 1995). In the case of developing a measurement tool the literature militates against this because the concepts of reliability and validity are often explained poorly and there is confusion between them. Therefore, before it was possible to develop a proposal for measuring the reliability and validity of the Bannigan Utilisation of Research Profile (Bannigan 2004) – the study from which this paper emanated – a conceptual framework had to be developed to explore and explain the different concepts of reliability and validity as they are related to measurement instruments in health and social care. This involved using the existing literature to formulate a generalised scheme of relevant concepts, i.e. conducting a conceptual analysis of the literature, and then assessing the accuracy of the resultant framework by verifying the conclusions drawn with the existing literature (see Miles & Huberman 1994, Polit & Hungler 1995). The conceptual framework is reported here.

Results: the concepts of reliability, validity and utility

The field of psychometrics 'provides a way to quantify the precision of measurement of qualitative concepts such as satisfaction' (Utwin 1995 p. 1). The product of psychometrics is measurement scales. Reliability and validity are research techniques used to assess the accuracy of measurement scales. Reliability (or consistency) refers to the stability of a measurement scale, i.e. how far it will give the same results on separate occasions, and it can be assessed in different ways; stability, internal consistency and equivalence. Validity is the degree to which a scale measures what it is intended to measure. The different terms in common usage (of the 35 possible available) related to validity include:

- content validity (which includes face validity);
- criterion validity, (which includes concurrent and predictive validity);
- construct validity (which includes convergent, divergent, factorial and discriminant validity) (Lynn 1986).

Utility, i.e. how practical the measurement scale is for use in the field, is another key issue to be considered alongside reliability and validity when developing, or assessing the quality of, a measurement scale. This is because if the scale is not actually used the time spent developing it and testing for reliability and validity will have been wasted.

Reliability

Reliability is essentially concerned with 'error in measurement' (McDowell & Newell 1996, p. 37) i.e. how consistently or dependably does a measurement scale measure what it is supposed to be measuring (Polit & Hungler 1995). The premise for conducting reliability tests is that there will always be a degree of random error in the administration of measurement scales. An example of a random error is a mistake in measurement due to the respondent or rater being distracted. Reliability assesses 'the extent to which a score is free of random error…[and]…is defined as the proportion of observed variation in scores' (McDowell & Newell 1996, p. 37). Essentially, 'the less variation an instrument produces in repeated measurements of an attribute, the higher its reliability' (Polit & Hungler 1995, p. 347). Therefore, 'reliability is a statistical measure of how reproducible the instrument's data are' (Utwin 1995, p. 6) and can be equated with stability, consistency and dependability (Polit & Hungler 1995). Reliability can be assessed in different ways; test-retest reliability for stability, inter-item reliability for internal consistency and interrater reliability or parallel scale for equivalence.

Stability

A measurement scale's stability is '…the extent to which the same results are obtained on repeated administrations of the instrument. The estimation of reliability here focuses on the instrument's susceptibility to extraneous factors from one administration to the next' (Polit & Hungler 1995, p. 347). This is assessed through 'test-retest reliability', a commonly used indicator of the reliability of a measurement scale (Utwin 1995). The measurement scale under development is administered on two separate occasions to the same sample and the scores are compared. Statistical procedures are used to elucidate a reliability coefficient; 'a numerical index of the magnitude of the test's reliability…[where]…the higher the coefficient, the more stable the measure' (Polit & Hungler 1995, pp. 348–9). No test will yield exactly the same results from test to test; therefore, it is necessary to determine what is an acceptable level of error. The issues that have to be considered in designing test-retest studies are:

- That the construct being measured may change over time regardless of the stability of the measure and so may confound the calculation of a reliability coefficient.
- Memory of the first administration of the test may influence the second (Eysenck 1994).
- Subjects may actually change as a result of the first test administration.
- Subjects may not be as careful when using an scale a second time.

Polit and Hungler (1995) advise: 'Stability indexes are most appropriate for relatively enduring characteristics such as personality, abilities, or certain physical attributes such as height' (p. 349).

Internal consistency

Internal consistency '…is applied not to single items but to groups of items that are thought to measure different aspects of the same concept' (Utwin 1995, p. 21). It is used to assess how well the different items measure the same characteristic (Utwin 1995, p. 25). 'An instrument may be said to be internally consistent or homogeneous to the extent that all of its subparts are measuring the same characteristic' (Polit & Hungler 1995, pp. 349–350). Internal consistency is a widely used method of testing for reliability because it is economical and it identifies errors in the sampling of items (Polit & Hungler 1995). A variety of procedures exist for measuring internal consistency including the 'split-half technique', 'Cronbach's alpha' (or 'coefficient alpha') and the 'Kuder–Richardson formula 20' (KR-20). Nunally (1967) advises coefficient alpha is the best estimate of reliability because most major sources of error are due to the sampling of instrument contents. The theory behind this procedure is 'the higher the internal consistency, the higher the test-retest reliability will be' (McDowell & Newell 1996, p. 40). However, the procedures for internal consistency do not consider fluctuations over time (Polit & Hungler 1995).

Equivalence

Equivalence can be addressed in two ways (i) the use of the scale by the same administrators at the same time (i.e. interrater reliability) or (ii) administering two parallel forms of the same scales to the same sample successively (i.e. alternative form reliability). In relation to assessing interrater reliability it is suggested that 'The use of Pearson correlations can seriously exaggerate the impression of reliability' (McDowell & Newell 1996, p. 38). Kendall's tau is used as it takes tied scores into account (Bryman & Cramer 1994). Other techniques suggested are intraclass correlations, analysis of variance, Rasch's item response model and rank-order correlations (Polit & Hungler 1995, McDowell & Newell 1996).

Scalability

Scalability, as determined using item response theory (IRT), is included here due its growing popularity and application to a range of well known psychological inventories such as the NEO Five Factor Index (Watson *et al.* 2007). Scalability is the extent to which individual items in a scale measure the latent trait that is being measured and do so distinctly from other items in the scale. Establishing scalability provides an ordering of items along the latent trait (e.g. depression) and also helps to exclude redundant items. Thereby, a cumulative scale is produced, one where the score on an individual item provides a measure of the latent trait. A popular method of establishing scalability is Mokken scaling which provides parameters that describe the extent to which items are order along the latent trait and the extent to which they do so reliably and also the extent to which the ordering if items has occurred by chance (Watson *et al.* 2008a,b).

Validity

Once a measurement scale has been shown to be reliable over time it should be assessed to establish whether or not it is reliably measuring what you want it to measure (Utwin 1995). Validity is concerned with the meaning and interpretation of a scale. There are many ways of testing validity and it has been suggested that 'A variety of approaches should be used in testing any index, rather than relying on a single validation procedure' (McDowell & Newell 1996, p. 37). This is because validity is not absolute. It is a matter of degree rather than an 'all or nothing' concept' (Carmines & Zellar 1979). 'In reality…it is not possible to take one form of measurement validity in isolation, as several forms may be applicable' (Gould 1994, p. 102).

Content validity

Face validity and content validity are two closely related forms of validity and they are the minimum requirement of acceptance of a scale. However, Streiner and Norman (1995) recommend that '…this judgement should comprise only one of several used in arriving at an overall judgement of usefulness and should be balanced against the time and cost of developing a replacement.' (p6).

Face validity

Dempsey and Dempsey (1992) describe face validity as the quickest method of determining validity. It is an assessment of whether a measurement scale looks reasonable, i.e. are the items included in the scale relevant? Face validity is directly related to the 'subjects acceptance of the text' (Payton 1988). 'The measurement tool must be understandable and perceived as relevant by the subjects to ensure their co-operation and motivation' (Gould 1994, p. 99). Face validity is not tested using statistical procedures. Subjects, experts and/or the researcher may be involved in a consideration of whether a scale appears to be relevant. Obviously, the more people and different groups related to the subject who are involved in the process the more acceptable it is likely to be. An assessment of face validity is important because acceptability of a scale is important to its utility.

Content validity

Content validity considers whether a scale has included all the relevant and excluded irrelevant issues in terms of its content. From a psychometric perspective this means the extent to which the measure adequately samples all possible questions that exist. For example, in behaviour scales it would be the extent to which the measure samples behaviours representative of the entirety of behaviours (Carr 2001). It is usually assessed by either:

1 a critical review by an expert panel for clarity and completeness or,
2 comparing with the literature or,
3 both.

This is done to achieve authenticity, i.e. to ensure all concepts relevant to the construct of interest are included in the instrument (Messick 1994), and assure directness, i.e. not including items which are not relevant. Content validity is closely related to construct validity. This is because the domain of content must first be defined (Rothstein 1985, Arnell & Sim 1993) and then it must be investigated to see if the measurement scale adequately reflects the domain (Gould 1994). The difficulty for researchers with content validity is there is no definitive list of 'correct content' (Gould 1994). It is, therefore, impossible to sample the content of a concept and establish total content validity (Arnell & Sim 1993). It can also be difficult to ensure that the measuring scale includes all the components of a concept (Brink 1991). Content validity should be carried out in the planning stages to try to ensure content validity from the outset, rather than making a judgement on it at a later stage (Gould 1994).

Whilst content validity is not usually assessed using formal statistical procedures, Lynn (1986) and Cohen *et al.* (1996) have suggested ways of quantifying content validity using the Index of Content Validity (CVI) and Content Validity Ratio (CVR) respectively. Unlike content validity, criterion validity and construct validity are tested using more formal statistical procedures.

Criterion validity

Criterion validity involves comparing the scale being developed with a criterion measure that has been established as valid. Criterion related validity is relatively straightforward if a valid criterion is already in existence (Gould 1994). There are two subdivisions of criterion validity (i) concurrent validity when the information about the criterion is available at the time the test is administered (Eysenck 1994) and (ii) predictive validity where the criterion measure is obtained after the test has been administered (Eysenck 1994).

Concurrent validity

Concurrent validity assesses the extent to which a measurement scale under development correlates with the 'gold standard' (McDowell & Newell 1996), i.e. is similar to the currently accepted scale for measuring the construct of interest (Polit & Hungler 1995). Concurrent validity may test the accuracy of a complete measure or each question ('item analysis'). The procedure is to apply the scale under development against the established test to an appropriate sample of people and compare the results to test level of agreement. 'The correlation of each question with the criterion score is used to select the best questions and thereby refine draft versions of the questionnaire' (McDowell & Newell 1996, p. 31). However, it is important to be sure that the gold standard is a true gold standard in terms of its psychometric properties and not just a scale that is in common usage but has no reliability or validity. In most instances there will not be another reliable and valid measure available. However, if another reliable and valid test exists, 'it begs the question of whether a new test is needed in the first place. It must offer something different to be of any use' (Carr 2001, p. 15).

Predictive validity

As with concurrent validity, predictive validity involves correlating the results of one scale with the results of a second scale that is administered much later (Utwin 1995). It is used to measure the accuracy of a measurement scale because it 'measures how well the item or scale predicts expected future observations' (Utwin 1995, p. 45).

Construct validity

If a gold standard or other measure does not exist, and there is no way of directly testing the relationship between the measurement scale and the underlying concept (Arnell & Sim 1993), validity can be tested by assessing to what extent the measurement scale under development correlates with the construct under investigation (Polit & Hungler 1995). Construct validity, is the main form of validation for a test, it is an indirect approach and multiple measures can be used to determine validity (Seaman 1987). Construct validity is relevant when a scale has been developed on the assumption of a particular theory. It is demonstrated by investigating the convergence or divergence of similar tests and by logical theoretical argument (Domholdt 1993). The procedure for testing construct validity begins with defining the topic or construct to be measured (McDowell & Newell 1996). 'These may be expressed as hypotheses indicating, for example, what correlations should be obtained with other instruments, which respondents should score high or low, or what other findings would be predicted from the scores' (McDowell & Newell 1996, p. 33). Construct validity is part art and part science that cannot be proven definitively 'it is a continuing process in which testing often contributes to our understanding of the construct, following which new predictions are made and tested' (McDowell & Newell 1996, p. 36). Features of good studies of construct validity (McDowell & Newell 1996) will:

- State clear hypotheses with justification of why they are the most relevant.
- Test the hypotheses stated.
- Try to disprove the hypothesis that the method measures something other than its stated purpose.

Construct validity can be assessed through convergent validity (that uses correlational evidence), factorial validity and discriminant validity (that uses group differences or discriminant evidence).

Convergent (and divergent) validity

Correlational evidence evolves by testing *a priori* hypotheses developed about how the measurement under development will correlate with another measurement scale. The testing of hypotheses formulated about the measurement scales the measure will correlate with is known as 'convergent validity'. Conversely, 'divergent validity' tests hypotheses formulated about the measurement scales the measure will not correlate with. This may involve several other indices. Convergent validity assesses the sensitivity and divergent validity tests the specificity of a measurement scale. McDowell and Newell (1996) recommend 'Construct validation should begin with a reasoned statement of the types of variable with which a measure should logically be related…The expected strength of correlation coefficients (or of the variance to be explained) should be stated prior to the empirical test of validity.' (p. 34).

Factorial validity

Factorial validity involves factor analysis, which is 'a statistical procedure for reducing a large set of variables into a smaller set of variables with common characteristics or underlying dimensions' (Polit & Hungler 1995, p. 642). It is used 'to describe the underlying conceptual structure of an instrument; it examines how far the items accord in measuring one or more common themes' (McDowell & Newell 1996, p. 35). In relation to construct validity it is used to establish whether the items in the scales group together in a consistent and coherent way (Bowling 1995, p. 293). There are two main approaches to factor analysis exploratory factor analysis (EFA) and Confirmatory factor analysis (CFA). EFA, which is considered adequate by some authors (Watson & Thompson 2006) is used to identify a set of factors, which are not easily observed in a large set of variables (Watson & Deary 1997). CFA 'allows hypothetical models to be set up before the data are analysed and subsequently tested for their fit, by a number of criteria, to the data' (Watson & Deary 1997, p. 407). McDowell and Newell (1996), provide the following guidance for those carrying out factor analysis:

- Items should be measured at the interval-scale level.
- The response distributions should be approximately normal.
- There should be at least five (some authors say 20) times more respondents in the sample than there are variables to be analysed.

Discriminant validity

Measurement scales should be able to discriminate between different people being measured by it in a way that would be expected. As such, discriminant validity is 'The extent to which scores on a measurement distinguish between individuals or populations that would be expected to differ (e.g. people with or without a disease)' (McDowell & Newell 1996, p. 500). It is assessed using a multivariate statistical procedure (discriminant analysis) that 'selects the set of questions that shows the most marked contrast in the pattern of replies between the groups' (McDowell & Newell 1996, p. 500).

Utility

When developing a measurement scale a researcher also needs to consider its utility by assessing how practical the scale is to use in the field. Aspects to consider are the:

- time it takes to administer;
- ease of administration;
- language used to ensure the phrasing is clear (McDowell & Newell 1996, p. 31).

McDowell and Newell (1996) advise that new measurements should be re-tested in a variety of settings to assess how far different people are able to use a measure.

Discussion

At one level, the concepts of reliability and validity are relatively easy to understand. However, when it comes to translating this into the reality of psychometrics, where concepts are not tangible and standards are scarce then it becomes less easy to understand and, in fact, quite complex. The lack of 'gold standard' measures for most psychological and clinical concepts means that the issues of reliability and validity have to be approached in a multifaceted way: no single estimation of either is sufficient. If an observational instrument proves to be highly reliable when one person uses it on two occasions then the question remains as to whether or not it is reliable when used by two people on the same occasion; if an instrument is reliable when correlated against another similar measure the question remains as to whether or not it is valid in discriminating between two levels if the trait or state it purports to measure. Furthermore, meticulous psychometric investigation may reveal adequate reliability and validity but the important question remains about the utility of the instrument in the 'real world' of research or practice. It is our observation that, for all but a few well financed and commercially viable psychometric instruments which – by their very nature have utility – the whole range of estimations of reliability, validity and utility have very rarely been addressed for many instruments some of which are commonly used. It is easy to see why this is the case due to the detailed and lengthy work that is required, often over many years.

Conclusion

We have experienced personal confusion over the concepts of reliability and validity and also amongst those we teach, supervise and mentor. Slightly different definitions of reliability and validity exist in the literature but we consider that the main concepts – described in this paper – can be described and differentiated in a framework that permits their application both to the understanding of the extent to which existing instruments have been adequately developed and in the development of new instruments.

Relevance to clinical practice

The issues considered in the present paper have utility in clinical practice as it is essential that instruments used to measure psychological states such as anxiety, patient experiences such as pain and clinical symptoms such as fatigue measure what they say they measure and do so each time they are used. In an increasingly evidence based practice environment it is necessary that practitioners can evaluate the utility of the instruments that they are expected to use.

Conflict of interest

None.

Contributions

Study design: KB; data collection and analysis: KB; manuscript preparation: KB, RW.

References

Arnell P & Sim J (1993) Measurement validity in physical therapy research. *Physical Therapy* **73**, 102–110.

Bannigan K (2004) *Increasing the Use of Research Findings in Four Allied Health Professions* (Unpublished PhD thesis). School of Nursing, University of Hull, Hull.

Bowling A (1995) *Measuring Disease: A Review of Disease Specific Quality of Life Measurement Scales*. Open University Press, Buckingham.

Brink PJ (1991) Issues of reliability and validity. In *Qualitative Nursing Research: A Contemporary Dialogue* (Morse JM ed.). Sage publications, London, pp. 151–156.

Bryman A & Cramer D (1994) *Quantitative Data Analysis for Social Scientists*, 2nd edn. Routledge, London.

Carmines EG & Zellar RA (1979) *Reliability and Validity Assessment*. Sage Publications, Newbury Park.

Carr J (2001) *Assessing Change in Exercise Based Therapy Programmes for Back Pain Management: Is the Chronic Pain Coping Inventory a Reliable and Valid Measure?* (Unpublished MSc dissertation). City University, London.

Cohen RJ, Swerdlik ME & Phillips SM (1996) *Psychological Testing and Assessment: An Introduction to Tests and Measurement*, 3rd edn. Mayfiled Publishing Company, Mountain View, California.

De Von HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, Savoy SM & Kostas-Polston E (2007) A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship* **39**, 155–164.

Dempsey PA & Dempsey AD (1992) *Nursing Research and Basic Statistical Applications*, 3rd edn. Jones and Bartlett, Boston.

Domholdt E (1993) *Physical Therapy Research, Principles and Applications*. WB Saunders Company, Philadelphia.

Eysenck MW (1994) *Individual Differences: Normal or Abnormal*. Lawrence Erlbaum Associates, Hove.

Gould A (1994) The issue of measurement validity in health-care research. *British Journal of Therapy and Rehabilitation* **1**, 99–103.

Lynn MR (1986) Determination and quantification of content validity. *Nursing Research* **35**, 382–385.

McDowell I & Newell C (1996) *Measuring Health: A Guide to Rating Scales and Questionnaires*, 2nd edn. Oxford University Press Inc, New York.

Messick S (1994) The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* **3**, 13–23.

Miles MB & Huberman AM (1994) *Qualitative Data Analysis*, 2nd edn. Sage, Thousand Oaks, CA.

Nunally JO (1967) *Psychometric Theory*. McGraw Hill, New York.

Payton OD (1988) *Research: The Validation of Clinical Practice*, 2nd edn. FA Davis Company, Philadelphia.

Polit DF & Hungler BP (1995) *Nursing Research Principles and Methods*, 5th edn. JB Lippincott Company, Philadelphia.

Rothstein JM (1985) Measurement and clinical practice: theory and applications. In *Measurement in Physical Therapy: Clinics in Physical Therapy* (Rothstein JM ed.), Vol. 7. Churchill Livingstone, New York, pp. 1–46.

Seaman CHC (1987) *Research Methods, Principles, Practice and Theory for Nursing*, 3rd edn. Appleton and Lange, California.

Streiner DL & Norman GR (1995) *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford University Press, Oxford.

Utwin MS (1995) *How to Measure Survey Reliability and Validity*. Sage Publications, Thousand Oaks.

Watson R & Deary I (1997) Feeding difficulty in elderly patients with dementia: confirmatory factor analysis. *International Journal of Nursing Studies* **34**, 405–414.

Watson R & Thompson DR (2006) Use of factor analysis in *Journal of Advanced Nursing*: literature review. *Journal of Advanced Nursing* **55**, 330–341.

Watson R, Deary I & Austin E (2007) Are personality trait items reliably more or less 'difficult'? Mokken scaling of the NEO-FFI *Personality and Individual Differences* **43**, 1460–1469.

Watson R, Deary I & Shipley B (2008a) A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine* **28**, 575–579.

Watson R, Roberts (nee Shipley) B, Gow A & Deary IJ (2008b) A hierarchy of items within Eysenck's EPI. *Personality and Individual Differences* **45**, 333–335.