The application of Classical Test Theory (CTT) to the development of Patient-Reported Outcome Measures (PROMs) in Health Services Research

Matthew Hankins

Submission for PhD by Publication

University of Brighton
Contents

Abstract.................................................................................................................................................. 3
1. Introduction and overview................................................................................................................ 5
  1.1 Patient-reported Outcome Measures in health services research................................. 5
  1.2 Classical Test Theory (CTT) and Patient-reported Outcomes......................................... 7
  1.3 The problem of measurement error in health services research................................. 10
2. Classical Test Theory (CTT)............................................................................................................. 14
  2.1 The measure model and its assumptions.............................................................................. 14
  2.2 The observed score, the true score and reliability.............................................................. 15
  2.3 Estimates of reliability $r_{xx}$............................................................................................... 16
  2.4 The Standard Error of Measurement..................................................................................... 17
  2.5 Reliability and validity.............................................................................................................. 18
  2.6 Unidimensionality and factor analysis................................................................................... 18
  2.7 Reliability and the assumptions of Classical Test Theory................................................. 20
  2.8 Classical Test Theory and the Theory of Test Discrimination........................................ 20
3. Discussion of Publications Included in the Thesis...................................................................... 25
  3.1 The Satisfaction with Information about Medicines Scale (SIMS): A new measurement tool
      for audit and research.............................................................................................................. 25
  3.2: Measuring patient satisfaction for the Quality Outcomes Framework ......................... 27
  3.3 The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result
      of negative phrasing?............................................................................................................. 30
  3.4 The reliability of the twelve item General Health Questionnaire (GHQ-12) under realistic
      assumptions............................................................................................................................. 30
  3.5 Time orientation and health-related behaviour: measurement in general population
      samples....................................................................................................................................... 32
  3.6 Statistical guidelines for studies of the theory of reasoned action and the theory of
      planned behaviour.................................................................................................................... 33
  3.7 The expectancy-value muddle in the Theory of Planned Behaviour - and some proposed
      solutions................................................................................................................................... 33
  3.8 Questionnaire discrimination: (re)-introducing coefficient $\delta$........................................ 36
  3.9 How discriminating are discriminative instruments?........................................................... 38
  3.10 Is the Satisfaction with Cancer Information Profile (SCIP) valid for tailoring information
      for patients with head and neck cancer? ................................................................................. 40
4.0 Conclusion ..................................................................................................................................... 41
5.0 References..................................................................................................................................... 43
Appendix 1: Published works and correspondence........................................................................ 47
Abstract

Patient-Reported Outcome Measures (PROMs) are increasingly used in health services research and clinical practice for the quantification of patient experiences, including quality of life, mood (e.g. depression), and satisfaction with services. Such PROMs usually take the form of questionnaires. The underlying measurement model is derived from psychometric theory, specifically Classical Test Theory (CTT). This model requires statistical analysis of questionnaire data to establish the quality of data so collected, with emphasis on the reliability (reproducibility) and validity (domain-specific measurement) of the data.

The widespread adoption of CTT by health service researchers and clinicians is potentially problematic because very few health service researchers receive training in psychometric methods. Researchers and clinicians are therefore unaware of the limitations of CTT and the assumptions of the statistical methods used. This has led to a number of theoretical and empirical problems, illustrated in this thesis, the common theme being that inappropriate methods applied uncritically yield data of questionable value.

A critical review (Hankins et al. 2007) and subsequent published correspondence reveals that unvalidated questionnaire measures were adopted by the NHS Confederation in the New General Medical Services Contract (2003) for UK general practitioners. These questionnaires were used in annual national surveys of patient satisfaction as part of the Quality Outcomes Framework. In contrast, a measure of patient satisfaction developed by the author (Horne, Hankins, & Jenkins 2001) demonstrates that reliability and validity can be established if appropriate piloting is undertaken.

A second strand of this thesis examines and expands upon a little-explored facet of questionnaire validity, the index of discrimination. Building on work by Ferguson (1949) the index is expanded in generality to take into account modern multi-item scales (Hankins 2007) and applied to a practical problem of designing a PROM to determine information needs of patients with head and neck cancer (Hankins & Llewellyn 2008). It is further argued that the index of discrimination is a core concept in discriminative health-related quality of life measures (Hankins 2008a and subsequent published discussion).

A third stand examines the applicability of CTT to psychological models of behaviour widely used in health services research and problems deriving from inappropriate
conceptualisation of the measurement model (Hankins, French & Horne 2000; French & Hankins 2003).

Finally, the fourth strand examines the technical detail of some of the assumptions and methods used to establish the dimensionality of PROMs. The specific example chosen is the General Health Questionnaire (GHQ-12) a widely used measure of psychiatric morbidity. Using a large sample taken from the British Household Panel Survey it is demonstrated that the scale is a single dimension if correlated errors are taken into account (Hankins 2008b), and that the reliability of the GHQ-12 reported in many studies may be considerably over-estimated (Hankins 2008c). It is argued that inappropriate analysis limits the potential of PROMs such as the GHQ-12 in health services research.
1. Introduction and overview

1.1 Patient-reported Outcome Measures in health services research

A Patient-reported Outcome Measure (PROM) can be defined, broadly, as quantitative information elicited from a patient, usually in the form of a structured interview, task or questionnaire completed by the patient or the interviewer (Marshall et al. 2006). In practice, PROMS usually take the form of a questionnaire and it is this method of data collection that is critically discussed in this thesis.

PROMS are increasingly used in health services research to quantify diverse aspects of patients’ experiences and perceptions including but not limited to attitudes, beliefs, satisfaction, mood, morbidity, pain, functional limitations and quality of life. In practice the expression „PROM“ not only refers to the methodology of eliciting measurements by self-report but also to the instruments themselves. For example, patient satisfaction in UK general practice may be measured using the Improving Practices Questionnaire (IPQ) or the General Practice Assessment Questionnaire (GPAQ) (Hankins et al. 2007). Similarly in general practice severity of depression may be measured using the Beck Depression Inventory (BDI) or the Patient Health Questionnaire (PHQ-9) (Cameron et al. 2008), while in secondary care settings the Hospital Anxiety and Depression Scale (HADS) may be used. All are examples of the methodology of PROM in that they are intended to quantify a domain of interest (in these examples, patient satisfaction and severity of depression) using patient self-report and hence the questionnaires themselves are referred to as PROMs. In addition, PROMs are increasingly used as aids for, or the basis of clinical decisions; in part this derives from increased exposure of clinicians to PROMs through their use in research (Cameron et al. 2008).
One obvious reason for the increasing use of PROMs in health services research is that questionnaire measurements are relatively easy to obtain at little cost compared to measurements obtained by other methods such as assay or diagnostic interview by a health professional. This ease of administration should of course be balanced against the quality of measurement obtained. On the other hand, patient report might be the only method for eliciting data on some outcomes such as satisfaction or mood. If this is the case then a lower quality of measurement might be acceptable, but should be quantified to allow accurate interpretation of the results.

In addition, government initiatives have increased the attractiveness of PROMs to researchers and health professionals. UK General Practitioners (GPs), for example, are awarded bonus payments for conducting annual surveys of patient satisfaction using the Improving Practices Questionnaire (IPQ) or the General Practice Assessment Questionnaire (GPAQ) (Hankins et al. 2007). Current NICE guidelines for the management of depression in primary care (Clinical Guideline 23; 2007) require an assessment of the severity of depression using a PROM and UK GPs are again awarded bonus payments for applying such measurements in the management of the depression treatment pathway (New GMS Contract 2006/7). In the US, the Food and Drug Administration (FDA) has published guidelines for the use of PROMs in clinical trials (US Department of Health & Human Services FDA 2006) which authorise the use of PROMs as outcome measures in their own right: the guidelines stress that PROMs are only legitimate for this purpose if they are rigorously validated.
1.2 Classical Test Theory (CTT) and Patient-reported Outcomes

All of the examples above share a common theory of measurement. Developers of PROMs in Health Services Research use the psychometric methodology developed over most of the 20th Century, beginning with Spearman (1904). Despite several advances in the latter half of that century (most notably Item Response Theory (Rasch 1960/1980; Mokken 1971) all PROMs in use for health services research have been developed using the measurement model derived from Spearman. This model and the accompanying methodology are referred to as Classical Test Theory (CTT; Kline 2000), and because of the importance of this model to PROM in health services research, the key concepts and definitions are reviewed in Section 2.

The methodology of PROMs originated in the domain of psychometrics for the sole purpose of measuring abilities such as intelligence and knowledge. It was then adopted by mainstream psychologists for measuring a much wider range of psychological attributes such as attitudes, perceptions and beliefs, before being adopted for health services research for a yet larger range of attributes such as „patient satisfaction”. The publications collected together for this thesis illustrate a number of issues arising from this „drift” in the application of the methodology. The first is that PROMs have been uncritically adopted for HSR, with little or no consideration of basic concepts such as validity and reliability. The second is that an important component of PROMs, discrimination, has also been neglected. The third is that lack of awareness of some of the basic assumptions of CTT have resulted in inappropriate analysis of data, resulting in misleading conclusions. These are to some extent overlapping issues since most of
the conclusions drawn concern the degree of measurement error entailed. They are also not limited to HSR but apply to the majority of studies in psychology using CTT.

Before considering CTT in detail and how this applies to each publication, it should be helpful to illustrate the extent to which PROMs are relied upon as outcome measures and the potential for misleading analyses to influence clinical practice. A recent meta-analysis Kirsch et al. (2008) reported on the results of thirty-five clinical trials of the six most widely-prescribed anti-depressants. The authors concluded that these anti-depressants were only as effective as placebo for all but the most severe levels of depression, “reaching conventional criteria for clinical significance only for patients at the upper end of the very severely depressed category” (italics added). The study was widely reported in the mainstream and health services media and clearly has the potential to influence clinical practice. Whilst attracting some critical comments because of the unusual methodology (not least the unconventional method of deriving the trial effect sizes), this study is a good illustration of the uncritical adoption of PROMs and lack of consideration of measurement error.

The primary outcome for all of the clinical trials was a PROM, the Hamilton Rating Scale for Depression (HRSD: Hamilton 1960). The outcomes of these trials and the meta-analysis therefore depend on the quality of measurement provided by the HRSD and the criteria used to determine what constitutes a clinically relevant difference between the treatment and control groups. For the latter, Kirsch at al. used the NICE guidelines for the management of depression, which state that a difference of 0.5 standard deviations in an individual score is a „clinically meaningful difference“.
The NICE guidelines, however, do not account for the fact that an individual score obtained from a CTT measure, such as the HRSD, is an *estimate* of that individual's position on the scale. The degree of precision of that estimate depends on the reliability of the PROM, which will vary from sample to sample. A „clinically meaningful” difference should therefore be specified in terms of the measurement error actually observed in the study, and should certainly be larger than the difference expected from measurement error.

To put this in more concrete terms, a recent review of the psychometric properties of the HRSD (Bagby et al. 2004) reported that reliability varied from study to study, ranging from 0.48 to 0.92 with a median of 0.75. If we consider the median value of 0.75, then the standard error of measurement is 0.5 standard deviations, and the 95% confidence interval around the score is ± 1 standard deviation. Hence, the NICE guideline that a difference of 0.5 standard deviations is „clinically meaningful” ignores the fact that a difference of ± 1 standard deviation is expected simply from measurement error.

In addition to this, the meta-analysis failed to take into account that each study would report HRSD scores with varying degrees of error, and that this error might vary as a function of severity of depression. For example, it is possible that the HRSD was more reliable in the samples with extremely high levels of measured depression, since these patients might be expected to respond to the HRSD with more consistency. The key finding of the meta-analysis, that the anti-depressants were only effective at the highest levels of depression, might therefore be due to the greater degree of measurement error at lower levels of depression.
This single example reveals several fundamental misunderstandings of how PROMs work, and the conclusions that may safely be drawn from them. To summarise:

1. National (NICE) guidelines for the management of depression propose a criterion for clinically meaningful change which is smaller than the expected change due to measurement error
2. The NICE guidelines also fail to acknowledge that measurement error will vary from study to study, so that no universal criterion can be defined
3. The findings of this meta-analysis are consistent with variation in reliability across samples, as opposed to a substantive clinical finding
4. In general, meta-analyses of studies in which the primary outcome measure is a PROM will underestimate the pooled effect size unless measurement error is taken into account

1.3 The problem of measurement error in health services research

Measurement error has been acknowledged in the psychometric literature since Spearman, and is recognised to some extent psychological research. In the latter, a large degree of measurement error is usually considered acceptable:

“In the early stages of research...reliabilities of .70 or higher will suffice...it can be argued that increasing reliabilities much beyond .80 is often wasteful of time and funds” (Nunally 1978, p.245)

However, the potential for harm (economic or personal) is very remote in basic psychological research: funding is limited and very little of the output will apply directly to an individual. In contrast, health services research attracts large amounts of public
funding and the results may be applied directly to the treatment or management of an individual. Hence it should be expected that PROMs should be rigorously evaluated before use:

“In contrast to the standards in basic research, in many applied settings a reliability of .80 is not nearly high enough...in those applied settings where important decisions are made...a reliability of .90 is the minimum that should be tolerated and a reliability of .95 should be considered the desirable standard” (Nunally 1978, p.245-246)

As will be made clear in the following section, however, the accurate assessment of reliability depends on a number of assumptions: if these assumptions are not true, then an apparently reliable PROM might in fact have large measurement error. In addition, it is argued that the current methods of obtaining highly reliable PROMs do so at the expense of the ability of the PROM to discriminate between individuals and may therefore impinge on the validity of the PROM.

Section 2 gives an outline of CTT and its assumptions, and introduces the work of George Ferguson, an early researcher in psychometric theory. In Section 3 the thesis discusses the following publications as they apply to the issues identified in Sections 1 and 2:

a) The basic requirements of reliability and validity for PROMs in health services research


3.3 Hankins M (2008) The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result of negative phrasing? *Clinical Practice & Epidemiology in Mental Health* 4:10


b) **Developing the theory of test discrimination for PROMs in health services research**

3.9 Hankins M (2008) How discriminating are discriminative instruments? 
   *Health & Quality of Life Outcomes* 6:36 & Published Correspondence

3.10 Hankins M & Llewellyn CD (2008) Is the Satisfaction with Cancer 
   Information Profile (SCIP) valid for tailoring information for patients with 
   head and neck cancer? *BMC Cancer* 8:164
2. Classical Test Theory (CTT)

2.1 The measure model and its assumptions

As stated above, the predominant methodology for the development of PROMs in HSR is Classical Test Theory (CTT; Kline 2000). CTT is based on a mathematically simple model of the measurement of human abilities by a series of tests. The ability of each person is estimated by the summed score or the mean of these tests, and the score of each test is assumed to be a linear function of the ability. In terms of questionnaire design, CTT develops measures of attributes such as attitudes, beliefs, satisfaction and quality of life using a series of items intended to measure the attribute of interest. The estimated quantity of the attribute is once again the summed score of the items, or the mean of the item scores.

For example, the twelve-item General Health Questionnaire (GHQ-12; Goldberg 1988) estimates the degree of severity of psychiatric impairment using twelve items, each with four response options (see Table 2 of paper 3.3). An individual's response to each of the twelve items is assumed to be a linear function of the degree of his or her psychiatric impairment, and hence the summed score of responses is an estimate of the degree of psychiatric impairment. Since the items are scored 0-1-2-3 in increasing order of impairment, this gives a scale score of between 0 (no impairment) to 36 (maximum possible impairment). This metric is entirely arbitrary, of course, since any set of numbers might have been chosen to score the response. The statistical treatment of item and scale scores, however, is based on correlational analysis (in particular the variance explained) for which the metric should be irrelevant. Hence linear transformation of item scores is permissable: shifting the scoring frame from 0-1-2-3 to
1-2-3-4 will not change the size of correlation between items nor between the scale score and some other variable. Non-linear transformations such as multiplication by a second variable, however, do change the size of correlations between items, a topic discussed in papers 3.6 and 3.7.

The CTT measurement model can be represented algebraically as follows:

\[ O_i = T + e_i \] (2.1)

Where \( O_i \) is the score for item \( i \) (i.e. the data obtained by the questionnaire), \( T \) is the true score (the attribute to be measured) and \( e_i \) is the error in measurement of item \( i \).

The model makes two important assumptions of the data:

a) Measurement error \( e_i \) is random for all items, i.e. uncorrelated with any other variable, with a mean (across all measurements) of zero (the assumption of uncorrelated error);

b) Apart from error, the variance of the observed score, \( O_i \) is due to variance in one variable, the true score, \( T \) (the unidimensional assumption)

These assumptions are the basis of many of the statistical tests associated with CTT, in particular the estimation of scale reliability and factor analysis: the extent to which a PROM is valid depends on the extent to which they hold.

2.2 The observed score, the true score and reliability

The measurement obtained by a CTT measure is known as the observed score, \( X_+ \): this is simply the sum of the individual item scores \( O_i \):

\[ X_+ = \sum_{i=1}^{n} O_i \] (2.2)
in which \( i \) indexes each item from 1 to \( n \) and \( O_i \) is the score obtained on item \( i \).

The theoretical reliability of the observed scores is \( r_{xx} \), the degree of measurement error in estimating the true score from the item scores:

\[
    r_{xx} = \frac{\text{var}(T)}{\text{var}(X_+)}
\] (2.3)

With no measurement error, \( r_{xx} = 1.0 \), since without error the variance in the observed score \( X_+ \) will equal the variance in the true score \( T \). As measurement error increases, the proportion of variance in \( X_+ \) due to \( T \) decreases and \( r_{xx} \) tends towards 0.0. In practice an estimate of between 0.70 and 0.95 would be expected, indicating that measurements contained between 30% and 5% measurement error.

2.3 Estimates of reliability \( r_{xx} \)

Equation 2.3 gives the definition of the theoretical reliability \( r_{xx} \), the so-called "self-correlation". \( r_{xx} \) is a theoretical quantity, the correlation between two administrations of the questionnaire to the same participants under "ideal" conditions of (a) no change in the construct measured between administrations, (b) no recall of the first administration at the second administration and (c) all other sources of error held constant. Since \( T \) is unknown (it is what we are trying to estimate from the measurements), \( r_{xx} \) cannot be computed directly.

\( r_{xx} \) can, however, be estimated from the data. Traditional methods simulate the two theoretical administrations of the questionnaire. For example, test-retest reliability is derived from two administrations of the questionnaire, but cannot guarantee that the ideal conditions hold. Parallel forms reliability estimates \( r_{xx} \) from the correlation between
two equivalent forms of the questionnaire, guaranteeing condition (b) holds but not (a) and (c). Split-half reliability essentially simulates parallel forms by estimating reliability from the correlation of one half of the questionnaire with the other half. Hence, because all the items are completed at the same administration, conditions (a) and (c) are usually met, but not (b). The split-half method is attractive because it may be derived from cross-sectional data, but estimates vary according to which items are selected for each "half". By far the most common estimate, Cronbach’s coefficient Alpha (Cronbach 1951), derives the most robust estimate by essentially taking the average of all the possible split-half correlations.

2.4 The Standard Error of Measurement

As mentioned in 1.2, one use of the reliability coefficient is to estimate the standard error of measurement (SEM). The equation for the SEM is as follows:

\[
SEM = Std \, Dev \times \sqrt{1 - r_{xx}^2}
\]  

(2.4)

The SEM provides an estimate of the range of \( X \) that would be expected from a given true score, \( T \), on repeated administrations of the PROM. For example, if the sample standard deviation for a given PROM was 2.0 points, and the reliability estimated as 0.8, then the SEM would be 0.6 SD, or 1.2 points. For an individual with a score of 20 the probable range of \( T \) would be 20 ± SEM, or 18.8 to 21.2 points.

The 95% confidence limits for \( T \) may also be calculated from the SEM \( (1.96 \times SEM) \). Hence for the example above the observed score of 20 points would indicate that the 95% confidence interval for \( T \) would be between 17.6 and 22.4 points.
2.5 Reliability and validity

The reliability coefficient is an important diagnostic of the quality of measurement for a number of reasons. A low reliability coefficient means that a large amount of the variation in the observed score is variance due to error, not variance due to the true score. Measurement error of this sort cannot be ignored since it limits the degree to which the observed score \( X_+ \) may be correlated with other variables. In terms of hypothesis testing this attenuates the observed effect size and favours the null hypothesis.

Definitions of questionnaire validity vary, but all methods of assessing the degree to which the questionnaire measures what it is supposed to measure depend on the correlation between the observed score, \( X_+ \), and another corroborating variable. The size of this correlation is inversely proportional to the reliability: a perfectly unreliable questionnaire will not correlate with anything, and hence cannot be validated. Since the validity of questionnaire data ultimately depends on the degree to which it measures the true score, low reliability invalidates the measurement. Reliability is therefore a necessary but not sufficient condition for validity.

2.6 Unidimensionality and factor analysis

The family of statistical methods referred to as „factor analysis“ is commonly used to assess the dimensionality of CTT measures. CTT assumes that the observed score, \( X_+ \), is the sum of item scores when all items measure the same true score, \( T \). For a given individual, variation in item scores is assumed to be due to random error alone. If the item scores vary systematically, this suggests that variation in item scores is due to
another source of variation: the items measure more than one attribute and the assumption of unidimensionality is not met.

The estimates of reliability discussed in 2.3 are based on the assumption that variation in the observed score $X_\theta$ is due to variation in the true score, $T$, and random error. If another systematic source of variation is present, then $r_{xx}$ may be over- or under-estimated. Hence it is important to establish that scales are unidimensional.

Factor analysis refers to a collection of methods developed by psychometricians to establish the dimensionality of a dataset. The common core of these methods is to represent the correlation matrix as the product of one or more over-arching variables, or factors. In the context of CTT, factor analysis is applied to the correlation matrix of item scores; the intention is either to confirm that a single factor "explains" the correlation matrix, or to explore how many factors are required to explain the correlation matrix. In either case, items loading above an arbitrary threshold on a factor are assumed to form a unidimensional scale.

Factor analysis, however, assumes that the correlation matrix is free of measurement error. If items have large measurement error, they will have small correlations with other items, and will not load on any factor. If, however, some items share the same measurement error (i.e. they are biased), then they will tend to correlate with more with each other than with other items, and this will be interpreted as a factor. Hence, a set of questionnaire items might contain a subset of items with a common bias but still measure a single construct; factor analysis is not capable in principle of distinguishing this kind of questionnaire from one measuring more than one construct.
2.7 Reliability and the assumptions of Classical Test Theory

The assumptions of CTT outlined in 2.1 influence the degree to which the reliability coefficient accurately estimates $r_{xx}$:

a. Equation 2.1 assumes that the measurement error on each item will be random. One consequence of this is that *errors will be uncorrelated across items*. The assessment of unidimensionality by factor analysis depends on this being true.

b. Estimates of reliability are also based on the assumption of uncorrelated errors: reliability may be over- or under-estimated when errors are correlated.

c. Equation 2.2 means that the scale score $X_s$ is derived from the simple sum of item scores, or any linear transformation of item scores. Since the estimation of the reliability of the scale score is based on this assumption, non-linear transformations of item scores invalidate such estimates.

2.8 Classical Test Theory and the Theory of Test Discrimination

The measurement model and its implications outlined in sections 2.1 to 2.6 focus on unidimensionality and reliability, and indeed most researchers schooled in psychology and using CTT are aware of the need to demonstrate that measures are unidimensional and reliable. It is argued in this thesis that the assumptions on which these analyses rest are often not met, with misleading results.

A second strand to this thesis is a neglected area of thinking within psychometric theory. To a large extent this thesis follows the work of George Ferguson. Ferguson was an exponent of Guttman scaling and developed several aspects of Guttman’s "deterministic" scaling approach. He made two important contributions to psychometric...
theory, neither of which has received much attention in modern times. His main contribution was his theory of test discrimination, developed over the period 1941 to 1949, and further developed in papers 3.8, 3.9 and 3.10 of this thesis. The second contribution was his recognition that factor analysis, the principal analytic method of psychometricians, would produce two or more factors from unidimensional data. These "false" factors were a simple artefact of the frequency distributions of the variables: variables with similar frequency distributions would form different factors even for unidimensional data. This phenomenon is explored in papers 3.3, 3.4 and 3.5, where the frequency distributions of item scores are shown to be influenced by bias on negatively-phrased items.

Ferguson’s work seems to have fallen out of favour in the psychometric research community around 1949, the point at which he developed his theory of test discrimination, a fully ordinal approach to psychometric theory. This occurred at a point in time when psychometricians were strongly motivated to demonstrate that psychometric tests gave interval or ratio data (Michell 2009) and at this time an ordinal theory of tests was distinctly unwelcome (despite earlier warnings by Boring (1920) that psychometric data were best considered ordinal data).

It is not widely recognised that early researchers in psychometrics were greatly concerned that test and measurements yielded a normal distribution (Guilford 1954). This concern was based on the entirely false premise (current at the time) that most, if not all, human attributes were distributed normally in the population: hence if a test produced a normal distribution of scores, this was seen as evidence of its validity. Jackson & Ferguson (1943) pointed out that, first, there was no reason to assume that
human mental abilities were normally distributed and second, that the distribution of test scores was entirely under the control of the test designer, since the distribution of the summed score is completely determined by the number of items, the difficulty of the items and the inter-item correlations.

Jackson & Ferguson proposed that, since the distribution of test scores followed directly from the design of the test, the test should be designed to yield the most useful distribution of scores rather than some arbitrary (i.e. normal) distribution. The definition of „useful” in this context was that the test should discriminate important differences between persons. If the purpose of the test, for example, was to discriminate the highest performing respondents from the rest, then the only important discrimination was whether a respondent was in the former or the latter group: no discrimination was required within each group, and hence the optimal test design was one that made this discrimination and no others. A test designed for this purpose would have a J shaped distribution of scores. At the other extreme, if all between-respondents discriminations were of interest (as is more common in modern-day research), then the test should discriminate equally well all along the scale: the optimal distribution of scores would therefore be rectangular (i.e. flat), with each possible discrimination between respondents made with equal frequency. From this perspective the pursuit of a normal distribution is questionable, since the normal distribution makes few discriminations in the middle of the range. Ferguson also explicitly linked his theory of discrimination (Ferguson 1949) with purely ordinal measurement ambitions, arguing that psychometric assessment, being intended to discriminate rather than quantify, was a matter of rank ordering respondents from lowest to highest. Hence the only meaningful data from such
measurements was derived from purely ordinal relations (greater than, equal to, less than) among respondents.

Seemingly independently of Ferguson, Willard Thurlow was developing a similar theory, presented at conference in 1948 (Thurlow 1948). Ferguson published his paper “On the Theory of Test Discrimination” in 1949, in which he briefly presented his arguments and coefficient Delta, an index of the degree to which a test discriminated between respondents. Thurlow published the same index and some variations of it the following year (Thurlow 1950). His paper dealt with the concepts in far more depth than Ferguson’s: like Ferguson, he noted that psychometric measurement was usually intended to rank order, rather than objectively quantify, respondents. He also noted the paradox that reliability and scale variance could be maximised simply by having 50% of respondents “pass” a test and 50% “fail” but that this would provide minimal discrimination because within each 50% no further discriminations were made. Hence reliability was not the sole determinant of the quality of a scale intended for measurement of continuous attributes:

“If we were trying to predict a pass-fail criterion...this would be satisfactory. But if we are trying to predict a continuous criterion, this situation would not be satisfactory. In terms of the discriminations furnished by the test, it is evident that a test could have high reliability...and still discriminate poorly” (Thurlow 1950 p.281)

While Ferguson assumed that coefficient Delta would generalise to different measurement purposes, Thurlow reviewed the then current measurement models (CTT, Guttman scales and Thurstone equal-interval scales) and demonstrated the utility of
Delta for each. He also proposed a variant coefficient of „valid discrimination“ which took into account measurement error (i.e. are the discriminations made greater than those expected by measurement error?).

Despite Thurlow’s earlier presentation at conference and greater depth of treatment, the coefficient derived by both Thurlow and Ferguson became known as Ferguson’s Delta. Consistent with the origins in testing, rather than measurement (i.e. items were dichotomous „pass“ or „fail”), neither author considered the more general case of a scale derived from Likert-type (polytomous) items. Most psychometric textbooks fail to mention Ferguson’s Delta, the notable exceptions being Guilford (1953), Stevens (1951) and Kline (2000).

In paper 3.8 I argue that the concept of test discrimination, and in particular coefficient Delta, was not further explored because of the assumption that Delta applied only to measures with dichotomous items, which was largely true in 1949 but not in the present day. Adapting Delta for use with polytomously-scored items proved relatively simple, and widened the scope of the statistic for use with Likert-type scoring favoured in health services research. The argument is further developed in paper 3.9, in particular exploring Thurlow’s proposal that reliability and discrimination are not entirely compatible concepts. Finally, in paper 3.10 the theory of test discrimination is applied to an existing PROM, the Satisfaction with Cancer Information Profile (SCIP), demonstrating utility in developing discriminating health services research measures.
3. Discussion of Publications Included in the Thesis

3.1 The Satisfaction with Information about Medicines Scale (SIMS): A new measurement tool for audit and research

This paper is included to illustrate (a) a “standard” approach to PROM validation and (b) how the basic assumptions of CTT assumptions may be ignored, with the potential for misleading results.

The Satisfaction with Information about Medicines Scale (SIMS) is a PROM developed at the University of Brighton for measuring the degree to which patients are satisfied with the information received about their medication. At the time of writing the paper has been cited 34 times.

Reliability was estimated using Cronbach’s Alpha, and separate estimates are given for different conditions, since each patient population could vary in measurement error.

Test-retest reliability was also used in one sample (patients receiving anticoagulants): once again, sub-samples (stable and unstable patients) were treated separately since they represent two distinct populations that could vary in measurement error. Validity was assessed as the correlation of the SIMS score with other constructs thought to be associated with satisfaction.

The reporting of the reliability analysis takes the now-standard approach of declaring the reliability “acceptable” or “good” if the reliability coefficients are larger than some arbitrary threshold (in this paper, this is 0.6). As argued in sections 2.2 and 2.4, the
threshold itself is unimportant; what is important is the degree to which measurement error influences the decisions to be made from the data.

As may be seen from Table 2, the reliability coefficients ranged between 0.81 and 0.91 for the full-scale SIMS. From these estimates the standard error of measurement ranged between 0.3 and 0.4 standard deviations. The implication of this is that differences in satisfaction or changes in satisfaction of these magnitudes are entirely within the range expected from measurement error.

Table 2 also gives estimates for two SIMS subscales: these are uniformly lower than for the full-scale SIMS. This is almost certainly due to an artefact introduced in the calculation of Cronbach’s Alpha. Alpha varies as a function of the inter-item correlations and the number of items. Hence, cutting the full scale SIMS (17 items) into two shorter scales (9 items and 8 items) would be expected to reduce the value of Alpha even if there was no loss of reliability.

More problematic for the introduction of two subscales in that principal components analysis (PCA) was used to assess the dimensionality of the scale but is not reported in the paper. This was due to constraints on the size of the Methods and Results sections, and the expectation that PCA was too complex a method for the general readership of the journal. Hence, the two subscales of the SIMS („Action and Usage” and „Potential Problems”) are derived from a factor-analytic method. Although each scale is presented as a separate dimension of patient satisfaction, the only basis for this is the interpretation made of the two-factor structure resulting from PCA. Hence it is possible
that there is only a single dimension of patient satisfaction and that the two subscales identified are products of the analysis.

Although the Discussion section notes that the SIMS could be used for „tailoring“ information provision, the ability of the SIMS to discriminate between different levels of satisfaction has not been demonstrated. In particular, the reliability coefficients suggest highly correlated item scores, which may militate against discriminations, since all items effectively agree. This issue is explored in paper 3.10 which develops the SIMS for use with patients with head and neck cancer.

3.2: Measuring patient satisfaction for the Quality Outcomes Framework

In 2004 the contract for the provision of UK General Practice services (the New GMS Contract) introduced the „Quality Outcomes Framework“ (QoF). Under QoF, general practitioners (GPs) were awarded additional annual payments for meeting quality indicators set out in the contract. The value of each quality indicator is determined by a points system, each point being worth approximately £120, depending on the size of the practice. In total 1050 points were available, which translates into potential payments of approximately £126000 per practice per year.

Quality indicators are grouped by domain and valued according to public health priorities and the evidence base for best practice: hence, a practice can gain 5 points for keeping a register of patients with chronic obstructive pulmonary disease (COPD), or 2 points for keeping a register of patients receiving drug treatment for epilepsy. What is perhaps surprising is that one of the largest points allocations (70 points) is for conducting a survey of the „patient experience“, using one of two questionnaires
mandated in the contract: the Improving Practices Questionnaire (IPQ, Greco et al. 2000) or the General Practice Assessment Questionnaire (GPAQ, Mead et al. 2008). Both are essentially CTT-derived measures of patient satisfaction.

When I moved to the Division of Primary Care & Public Health at Brighton & Sussex Medical School in 2005, I became interested in this adoption of psychometric methods by mainstream general practice, not least because of the implication for public funding. An estimated 8500 practices are covered by the QoF and over 95% are awarded the full 70 point allocation for conducting the patient experience survey. A conservative estimate is that the surveys cost over £65,000,000 per year, and so it might be expected that the survey tools (the IPQ and GPAQ) and methodology employed were extremely rigorous. I originally intended to conduct a literature review to assess the validity and reliability of the IPQ and the GPAQ, and comment accordingly. In particular it seemed important to be able to estimate the reliability of the questionnaires, since practices would ultimately be compared with each other or required to „improve“ over time. A systematic search of the peer-review literature, however, revealed very little evidence of any validation at all, with only one paper describing the IPQ, and none at all describing the GPAQ.

At this point it became very difficult to write the original literature review, since there was no literature to review. However, it seemed important to report on the findings, so the range of the review was extended to include the precursor to the GPAQ, the GPAS (General Practice Assessment Survey, Roland 2000). This was justified on the grounds that, despite some differences, it would be reasonable to expect the reliability and
validity of the GPAQ to be estimable from the GPAS. This decision allowed the inclusion of three additional papers.

The resulting publication was abridged at the request of the editor, as the main readership - UK General Practitioners - was not expected to understand the more technical aspects of psychometric methods. A number of technical details were therefore omitted from the publication. These included the item selection method employed and the likely bias introduced in the IPQ by having only one negative response choice on the five-point scale (poor, fair, good, very good, excellent), as well as the nonsensical principal components analysis which revealed two subscales: unsurprising since the IPQ was constructed by joining together two existing questionnaires. Also omitted was the discovery that both the IPQ and the GPAQ had been modified since the original GMC contract and did not match those mandated by the contract: hence multiple versions of both questionnaires exist, none with sufficient validation.

The results were presented at conference in 2006 and published in 2007. Following publication, and in response, the GPAQ development team published reliability data for the GPAQ (version 2, Roland et al. 2008). This paper revealed that the original GPAQ had been modified because one of the items had been scored in the opposite direction to the others, confusing respondents, a problem that should have been addressed had the GPAQ been validated before use. As a result, all data for this item collected in the previous two surveys were discarded.
The resulting correspondence (appended to this paper) not only confirmed the findings, but also reinforced the impression that these measures had been adopted with very little understanding of the requirements for valid measures of patient-reported outcomes in health services research. Reliability and validity data are the absolute bare minimum requirement to understand how well the measures are working, and hence the use to which the data might be put. The data obtained, at enormous expense, are consequently of little value, but are currently published by the NHS (http://www.qof.ic.nhs.uk/).

3.3 The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result of negative phrasing?

3.4 The reliability of the twelve item General Health Questionnaire (GHQ-12) under realistic assumptions

Papers 3.3 and 3.4 attempt to resolve two related issues, and so both will be discussed in this section. Once again, it demonstrated that the use of PROMs in health services research is compromised when the assumptions of CTT are not met. In these papers, it is shown that (a) PROMs may be misidentified as multi-dimensional when they are not, (b) spurious dimensions may interpreted as substantive dimensions (“reification”), and (c) reliability coefficients may be over-estimated.

The twelve-item General Health Questionnaire (GHQ-12: Goldberg et al. 1988) is a widely used PROM, measuring the degree of „non-psychotic psychiatric disorder“ . The measurement properties of the GHQ-12 have been widely studied and reported; as a result there is a large literature exploring the dimensionality of the GHQ-12. At the time
of writing this paper, the consensus view was that the GHQ-12 comprised three
dimensions, and hence should not be used as a unidimensional PROM.

Paper 3.3 explores these findings and concludes that the apparent three dimensional
structure is more likely a reflection of the misapplication of CTT, and the dimensions are
the product of the analytic strategies employed. First, as Ferguson noted, factors may
emerge from unidimensional tests simply because of differences in the distribution of
item scores. Second, the GHQ-12 factors identified split the scale into two or three sets
of similarly phrased items (positively and negatively phrased). Third, item wording,
particularly negative phrasing, has been found to produce additional factors from
unidimensional measures.

This paper argues that the multidimensional factor structures reported reflect a failure of
the assumption that item measurement errors are uncorrelated (see section 2.1).
Although agnostic on the mechanism for the wording effect, it is argued that, whatever
the mechanism, any bias so introduced would be detectable from (a) increased variance
on the negatively phrased items and (b) correlated errors on those items. Since these
are a priori hypotheses, and supported by the data, this paper provides the strongest
evidence that the GHQ-12 is unidimensional.

The intention of the paper was to demonstrate that the conventional CTT method of
factor analysis would give misleading results if the assumptions of CTT were not met.
Having demonstrated that the GHQ-12 was unidimensional, it then became apparent
that I had identified but not resolved a further issue: if the assumption of uncorrelated
error was not met, then the reliability coefficient (Cronbach’s Alpha) was not valid, since
it is based on the same assumption. It also became apparent that the issue of positively and negatively worded items in the GHQ-12 had been previously addressed by Huppert et al. (1988), who suggested recoding the negatively phrased items to eliminate the bias. Hence in paper 3.4 I compared the fit of Huppert et al.’s scoring method with the previously-examined methods, with essentially the same results. Moreover, when reliability was estimated taking into account the correlated errors, it was found that Cronbach’s Alpha consistently over-estimated the reliability of the GHQ-12. Of particular concern was the decrease in reliability for the Likert-type scoring from 0.90 to 0.73, increasing the 95% confidence limit for T from ± 4.2 points to ± 6.6 points, and the very low reliability of Huppert et al.’s scoring method.

3.5 Time orientation and health-related behaviour: measurement in general population samples

Although papers 3.3 and 3.4 demonstrated the principle that inattention to the assumptions of CTT could lead to invalid conclusions about construct dimensionality, it was not clear how well these results would generalise to other PROMs. An opportunity to examine a similar problem (the extent to which factor structure could be explained by correlated errors) arose with another study of the performance of questionnaire measures of time-orientation.

Once again, taking into account the correlated errors on negatively-worded items demonstrated that the measure was unidimensional, rather than multidimensional as previously assumed. This raises the possibility that many PROMs are affected by the problem of correlated errors, and that researchers should adopt methods of analysis that incorporate this feature of the data to resolve problems of dimensionality. It also
remains to be seen if other constructs identified purely as the result of factor analysis are similarly spurious.

One serious error in this paper arises due to the delay in publication process. The analysis was conducted after that of paper 3.3 but before the development of thinking that justified paper 3.4. Hence in paper 3.5, although it is demonstrated that the assumption of uncorrelated measurement errors was not met, reliability is estimated using Cronbach’s Alpha, which rests on this assumption. Hence the value cited in the paper may be over-estimated, which in turn threatens the validity of the following analyses.

3.6 Statistical guidelines for studies of the theory of reasoned action and the theory of planned behaviour

3.7 The expectancy-value muddle in the Theory of Planned Behaviour - and some proposed solutions

These two papers will be considered together since their role within this thesis is to demonstrate how misapplication of CTT can result in nonsensical results. The argument developed in these papers concerns the use of "multiplicative composites" in the construction of scale scores. It is demonstrated in these papers that the theoretically-derived method of "weighting" one set of scores by another leads to uninterpretable findings, and in this section I re-examine the viability of "multiplicative composites" as the basis for CTT scale development.

The main argument of this section therefore concerns the violation of the fundamental assumption of CTT: that scale scores are derived from summed item scores, the only
permissible transformation of which is linear (Section 2.1). The specific context is two psychological theories widely used in health services research: the theory reasoned action (TRA; Fishbein & Ajzen 1975) and the related theory of planned behaviour (TPB; Ajzen 1985). Since TPB has largely superseded TRA, the discussion will focus on this theory.

The TPB models the likelihood of a given behaviour as the direct result of one or more psychological attributes (intention and perceived behavioural control), and further that intention is itself determined by three preceding constructs: the attitude towards the behaviour (attitude), the perception of social norms for the behaviour (norms), and perceived behavioural control (PBC).

The problem arises because of the origins of the TRA and TPB as social cognition models. These assert that attitude, norms and PBC are derived from the weighing up of „expectancy“ (how likely an outcome is) and „value“ (the perceived worth of the outcome). This „expectancy-value“ model is derived from economic theory, and while a case may be made for psychological plausibility (for example, someone will attend to a highly likely outcome of great value more than a less likely outcome with lower value), the implementation of TPB measures has taken this literally: two sets of item scores are obtained, one for „expectancy“ and the other for „value“. These pairs of item scores are then multiplied together and the products added to give the scale score $X_+$. 

In CTT terms this can only be sensibly interpreted as the measurement of two dimensions, „expectancy“ and „value“, since there are two separate sets of items. Hence from equation 2.1:
\[ E_i = T_E + e_i \]

The „expectancy” true Score \( T_E \) is estimated by the item scores \( E_i \) pertaining to „expectancy” and:

\[ V_i = T_V + e_i \]

the „value” true Score \( T_V \) is estimated by the item scores \( V_i \) pertaining to „value”.

On this basis it would be perfectly legitimate within CTT to treat „expectancy” and „value” as two separate unidimensional measures of distinct aspects of attitude, social norms and perceived behavioural control and sum the E and V items to form two E and V scales:

\[ X_{E+} = \sum_{i=1}^{n} E_i \]  \hspace{1cm} (3.1)

\[ X_{V+} = \sum_{i=1}^{n} V_i \]  \hspace{1cm} (3.2)

However, the specification of the TPB model requires the „expectancy” and „value” item pairs to be multiplied and then added to form the scale score. Hence, in contrast to equation 2.2 the measurement obtained is:

\[ X_+ = \sum_{i=1}^{n} E_i \times V_i \]  \hspace{1cm} (3.3)

Of note here is that further analysis of this summed score is not possible with CTT methodology for a number of reasons. First, the computation of reliability and factor analysis makes no logical sense: the composite score has been derived from (by definition) a two-dimensional source and cannot in principle generate a unidimensional measure. Second, the estimation of reliability, unidimensionality and validity based on
the item scores and summed score is only legitimate if the summed score is a linear transformation of the item scores. As can be seen from equation 3.3, the summed score \( X \) is the sum of the product of two sets of item scores (known as a multiplicative composite score or a sum of products score). Unless one set of item scores is constant, this is a non-linear transformation of the other set. It has been argued that equation 3.3 is a simple weighting of \( E \) scores by \( V \) scores (which would be legitimate), but once again this requires that one set of scores is constant.

Attempting to measure TPB constructs using multiplicative composites is therefore incompatible with CTT, and, in addition to the arguments presented in papers 3.6 and 3.7, a large number of TPB studies in health services research are based on invalid measurement data (e.g. see Godin & Kok 1996 for review).

### 3.8 Questionnaire discrimination: (re)-introducing coefficient \( \delta \)

This paper was the first to develop Ferguson’s and Thurlow’s coefficient of discrimination, Delta, for Likert-type scaling, and hence allow its use for assessing the discrimination of modern questionnaire measures. The purpose of the paper was threefold: to generalise the coefficient, to re-present the argument for examining the discrimination of scales, and to provide an empirical example of the utility of the coefficient of discrimination. Further innovations were the development of the 95% confidence interval for Delta using the bootstrap method, a recent innovation unavailable to Ferguson and Thurlow, and the implementation of Delta and the bootstrapped confidence interval in the statistical package R.
The publication of this and subsequent work in open-access journals was a conscious decision to support the open-access publishing movement as well as to ensure wider dissemination of the concepts. At the time of writing the paper has been accessed over 3000 times. Consistent with this I developed the software in R, an open-source statistics package freely available at no cost, and used British Household Panel Survey (BHPS), a large database freely available to non-commercial researchers.

There is a typographical error on page 3, bottom of column 1: m is the length of item scale, not the scale itself.

The arguments for the theory of discrimination have been presented in section 1 and the empirical findings will be discussed here, since they illustrate very clearly that discrimination is a property distinct from reliability and validity. The illustration itself also sheds some light on the longstanding discussion of the most appropriate scoring method of the widely-used PROM, the 12-item General Health Questionnaire (GHQ-12), and this question is revisited in more detail in papers (6) and (7).

Given the need to provide a convincing empirical demonstration of the need for an index of discrimination, and being limited to a cross-sectional dataset, I decided to examine the effects of different scoring methods (dichotomous or Likert-type) on the discrimination of the GHQ-12. One of the difficulties in determining a definitive scoring method for the GHQ-12 has been that both methods give equally reliable scales (typically Alpha>0.85), which are highly correlated (typically r>0.9), i.e. both methods have similar measurement error, and both correlate similarly with other measures (to the extent that this validates one scoring method, it validates the other equally well).
Hence, by standard psychometric assessment they are indistinguishable, but by considering the distribution of scores and the resulting discriminations, some distinction may be made. As noted by Ferguson, however, which of the two scoring methods is better depends on the intended use of the scale: if the intention is to discriminate between [no symptoms at all] vs. [any symptom at all], analogous to a diagnosis, then the dichotomous scoring method achieves this purpose. If the purpose is to measure the degree of psychiatric morbidity along a continuum, for example as a covariate, then the Likert-type scoring achieves this better. It should be noted that the enhanced discrimination of the Likert-type scoring method does not automatically generate more discriminations, as discussed later (paper 3.1), it is possible for a dichotomously-scored scale to be as discriminating as a Likert-type scale. The denominator of coefficient Delta takes into account the number of available scale points, and expresses the actual discriminations made as a proportion of these. Hence, Likert-type scoring generates more possibilities for discriminations, but unless these possibilities are actually observed in the data, a low coefficient of discrimination will result.

This paper illustrates the empirical value of Ferguson’s and Thurlow’s index of discrimination, and also confirms relevance to health services research.

3.9 How discriminating are discriminative instruments?

The domain of health-related quality of life (HRQL) is one area of research where CTT PROMs are explicitly developed to make discriminations of the Thurlow/Ferguson type. In this paper I refer to the dominant framework in this literature, developed primarily by Guyatt et al. (1992), as the „McMaster framework“.
Like Ferguson, Guyatt’s approach to HRQL is functional, in that he defines PROMs as fulfilling one or more distinct functions: evaluation of change over time, diagnosis against some external criterion, or discrimination between individuals. He argues that PROMs intended for the first function („evaluative“) are required to be both reliable and „responsive“, and in this departs from CTT, since he argues that „responsiveness“ is a distinct property of the PROM. In CTT, the ability of a PROM to detect changes in the measured attribute is one way to assess validity (i.e. if the PROM measures the correct attribute, it should also register change when that attribute changes). Almost all of the development of this framework, however, has been in developing various indices of „responsiveness“, and the issue of discrimination has been ignored.

I argue in this paper that a framework defining some PROMs as „discriminative“ requires some index of how many discriminations are made. As noted by Thurlow, a test might be perfectly reliable, yet not discriminate well; in addition, a test may appear to discriminate well, but be unreliable. Hence the concepts of reliability and discrimination are distinct, and establishing reliability tells us very little about the discriminations actually made.

At the request of the editors several empirical examples were included: these demonstrated the role of Delta in establishing whether a questionnaire discriminates, and in particular provided the first empirical investigation of the tension between maximising reliability and maximising discrimination. The example demonstrates that one common method of maximising reliability (excluding items to increase Alpha) decreases discrimination, and vice versa. This is an area in urgent need of further research, since it is clear that estimates of reliability based on highly correlated items
will result in scales constructed of items that essentially agree, i.e. each additional item in the scale fails to add further discrimination.

The argument is defended in subsequent correspondence (appended to the paper) and appears to fill an important gap in the McMaster framework.

3.10 Is the Satisfaction with Cancer Information Profile (SCIP) valid for tailoring information for patients with head and neck cancer?

The arguments developed in papers 3.8 and 3.9 were applied in this paper to the development of the Satisfaction with Cancer Information Profile (SCIP), which is itself a PROM developed from the SIMS (paper 3.1).

The justification for this paper was the intention to use the SCIP in a further study of the provision of „tailored“ information to patients with head and neck cancer. The concept of targeted information based on a measured lack of information entails that the measure can in fact discriminate between different informational needs. The SCIP had been previously validated and had been shown to be reliable, but as argued in 3.8 and 3.9, a reliable PROM does not necessarily discriminate in this way.

The results further justify consideration of PROM discrimination in addition to the traditional CTT indices of validity and reliability. The study also extends the theory of test discrimination to the examination of discrimination at the levels of item and sub-scale. The utility of this approach is demonstrated in the different outcomes for the two parts of the SCIP (A and B): while SCIP-A was found to be discriminating at the item, sub-scale and full scale levels, SCIP-B was found to be relatively indiscriminate at the item level.
4.0 Conclusion

Papers 3.1 to 3.7 share a common theme: that the assumptions of a widely-used measurement model are frequently ignored, with serious consequences for the generation and interpretation of meaningful data. Although these consequences have often been acknowledged in the psychological research literature, the use of Classical Test Theory in health services research raises the stakes, especially when clinical diagnosis and treatment decisions are based on such potentially misleading data. Under ideal conditions, patient-reported outcome measures are only estimates of the patient’s position on a continuum, and mostly crude estimates at that. Hence, “clinical cutoff scores” or criteria for “clinically meaningful change” should take this into account, and researchers and clinicians in general should be aware of the sizable uncertainty in these measurements; this should be quantified by the standard error of measurement, derived from a reliability coefficient greater than 0.90. It is certainly unacceptable to publish PROMs or data derived from PROMs without this basic test of data quality.

It has been demonstrated in this thesis, however, that ideal conditions do not always apply and that this can affect the quality of the reliability coefficient. Hence, even when the reliability and the standard error of measurement are reported, they may be over- or under-estimated. The assumption of uncorrelated errors was found not to hold in papers 3.3 to 3.5 and it remains to be seen how many other PROMs violate this assumption, and to what extent this threatens the validity of the data.

Papers 3.8 to 3.10 demonstrated the utility of an under-researched development of CTT, the Theory of Test Discrimination. Despite the initial promise of this approach, the
interaction between reliability and discrimination requires further elaboration.

Discriminations made by unreliable PROMs are not useful, and so attention once again must be drawn to the generally poor state of reliability estimation in health services research.
5.0 References


Spearman C (1904) “General Intelligence”, Objectively Determined and Measured American Journal of Psychology 15, 201-293.


Appendix 1: Published works and correspondence

a) The basic requirements of reliability and validity for PROMs in health services research


3.3 Hankins M (2008) The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result of negative phrasing? *Clinical Practice & Epidemiology in Mental Health* 4:10


b) Developing the theory of test discrimination for PROMs in health services research

3.9 Hankins M (2008) How discriminating are discriminative instruments? *Health & Quality of Life Outcomes* 6:36 & Published Correspondence

Declaration

I declare that the research contained in this thesis, unless otherwise formally indicated within the text, is the original work of the author. The thesis has not been previously submitted to this or any other university for a degree, and does not incorporate any material already submitted for a degree.

Signed:

Dated: 02\textsuperscript{nd} June 2009

All published works for which I am the sole author were entirely my own work; those for which I am the first author are substantially my own work, and solely my own work with reference to material discussed in this thesis. For the remainder, the work discussed in this thesis was my unique contribution to the publication (declarations from the lead authors of these works follow).
Dear Matthew,

This is to confirm your contribution and co-authorship the paper:


You made a substantial contribution to the psychometric analysis of the two measures of time orientation, the Consideration of Future Consequences Scale and the Zimbardo Time Perspective Inventory including advising on the use of Principal Components Analyses, the implications of reverse scoring of items and the use and meaning of Ferguson's delta to assess discrimination of a measure. You conducted the Structural Equation Modelling of the two measures and wrote the sections of the paper presenting the Structural Equation Modelling and describing discrimination. Finally, you commented on drafts and revisions of the paper.

Best wishes

Rachel

Rachel Crockett MSc PhD CPsychol, Visiting Research Fellow, Section of Health Psychology, Institute of Psychiatry, King's College London, 5th Floor Bermondsey Wing, Guy's Campus, London Bridge. SE1 9RT

Phone: 020 7188 5422/0192 Fax: 020 7188 0195
From: Rob Horne [mailto:rob.horne@pharmacy.ac.uk]

Sent: 14 November 2008 14:44

To: A.Mandy@brighton.ac.uk

Subject: Matthew Hankins


Dear Anne


I am writing to confirm that Matthew Hankins made a significant contribution the above paper. Matthew was responsible for the psychometric analysis of the data, the interpretation of the factor structure and drafted/amended the methods and results sections relating to the psychometric analysis.

With best wishes

Rob

Rob Horne
Professor of Behavioural Medicine
Head of Department of Practice & Policy
Director, Centre for Behavioural Medicine
The School of Pharmacy, University of London
School of Pharmacy, Mezzanine Floor, BMA House
Tavistock Square, London WC1H 9JP
Email: rob.horne@pharmacy.ac.uk Web: www.pharmacy.ac.uk
Direct line: + 44 (0) 207 874 1293
From: David French [mailto:aa3767@coventry.ac.uk]

Sent: 18 November 2008 04:28

To: Hankins Matthew

Subject: statement for Matthew Hankins

To whom it may concern:

I can confirm that Matthew Hankins contributed sufficiently for co-authorship of the paper "The expectancy-value muddle in the theory of planned behaviour - and some proposed solutions" that was published in the British Journal of Health Psychology in 2003.

Yours sincerely

David P French PhD CPsychol,
Reader in Health Behaviour Interventions,
Applied Research Centre in Health and Lifestyle Interventions,
Faculty of Health and Life Sciences,
Coventry University,
Whitefriars Building,
Priory Street,
Coventry CV1 5FB.

Tel +44 (0)2476 795 429
Fax +44 (0)2476 795 987.

http://www.coventry.ac.uk/cu/hls/staff/a/2928