Measuring is more than assigning numbers

Gorard, Stephen; Walford, G; Tucker, E; Viswanathan, M

Measuring is more than assigning numbers

Stephen Gorard
The School of Education
The University of Birmingham
s.gorard@bham.ac.uk

Measurement as an unconsidered concept

Measurement is fundamental to research-related activities in social science (hence this Handbook). In my own field of education research, perhaps the most discussed element of education lies in test scores. Examination results are measurements, the number of students attaining a particular standard in a test is a measurement; indeed the standard of a test is a measurement. The allocation of places at school, college or university, student:teacher ratios, funding plans, school timetables, staff workloads, adult participation rates, and the stratification of educational outcomes by sex, social class, ethnicity or geography for example, are all based on measurements. Good and careful work has been done in all of these areas (Nuttall 1987). However, the concept of measurement itself remains under-examined, and is often treated in an uncritical way. In saying this I mean more than the usual lament about qualitative:quantitative schism or the supposed reluctance of social scientists to engage with numeric analysis (Gorard et al. 2004a). I mean that even where numeric analysis is being conducted, the emphasis is on collecting, collating, analysing, and reporting the kinds of data generated by measurement, with the process of measurement and the rigor of the measurement instrument being somewhat taken for granted by many commentators. Issues that are traditionally considered by social scientists include levels of measurement, reliability, validity, and the creation of complex indices (as illustrated in some of the chapters contained in this volume). But these matters are too often dealt with primarily as technical matters – such as how to assess reliability or which statistical test to use with which combination of levels of measurement. The process of quantification itself is just assumed.

Although this chapter considers all of the issues above, what I want to do here is rather different. The first section discusses the general basis on which we allocate numbers to things and call the former ‘measurements’. The next section reconsiders the usual classification of levels of measures, leading to a third section suggesting a flaw in the logic of creating complex indices (such as attitude ‘measurements’). The next section looks at the kinds of errors that occur in measuring, leading to discussion of an even more serious logical flaw in the way in which we routinely handle errors in our measurements. The chapter ends by suggesting a fundamental change to the ways in which measurements are routinely handled in social science. Thus, the chapter illustrates the limitations of a range of fundamental notions that are generally taken for granted in social science measurement. In doing so, it is intended to strengthen social science measurement by encouraging scepticism and parsimony, and by discouraging the obfuscation of unsustainable positions through increasing complexity of purportedly technical solutions to measurement problems.

What is measurement?

When we say that we are measuring something, we generally imply a series of related preliminary steps that we have come to take for granted, and so to ignore. The danger in this lies in researchers remaining unaware of these pre-technical steps, and so mistaking the allocation of numbers to things as though it were measurement, leading to pseudo-quantification. For example, when we create a measure from scratch, the thing we seek to measure must exist in the sense that we can observe it, whether directly or indirectly, otherwise our measurement of it is pointless.[1] It must be identifiable by some means and identified by us. And we must know something about its properties and behaviour, so that we can relate these to our measures of it. Put another way, even where the thing we seek to measure is a concept, like length, this concept must have manifest qualities that we can measure, like the way in which sticks of different length might protrude to different extents (Gorard et al. 2002). We must also have a measuring scale, based on numbers or their equivalent, and this scale must have a standard so that other users can comprehend and compare our measurements. The scale must have divisions or units that are related as directly as possible to changes in the properties and behaviour of the thing being measured, so that we can associate these changes with gradations in the measurement scale. We should also have a clear estimate of the accuracy of our measure, in the sense of how well changes in the scale can track changes in the properties and behaviour of the thing being measured. In other words, we need to have a good idea of the size of the error component in our measure.

Without any one of these components, measurement is largely an illusory exercise deceiving either the producer or the user. If the thing we are measuring does not exist, the idea of measurement makes no sense. If we know that it exists somehow and somewhere, but it is not readily identifiable by the qualities it makes manifest then we cannot hope to measure it at present. If we can identify it but have no idea in fact of how it varies over time, place or other context then we cannot set up a measurement, because we cannot associate the measurement scale with these variations. We may later use an established measurement technique to measure new examples of the phenomenon, but at the outset we have to be able to associate existing variations in the phenomenon with gradations of our measure. Once we have met these preliminary requirements (among others perhaps) we have identified what seems to be a palpable quantity that is, in theory, capable of numeric summary. Next we need a suitable scale. If we do not have a scale capable of variation then we cannot associate scale variation with variation in our tentative quantity. If the scale does not have agreed and comprehensible steps within it, then we cannot convey variation in the quantity by variations in the scale. If we cannot directly associate variation in the quantity with variation in the scale – 'calibrate' our measure – then we are in danger of being misled or being misleading in our measurements. So, I emphasise this for what follows, we need two separate things – the object of measurement or its manifestation in the real world, and the measurement of it from the measuring scale.

These assumptions are largely unexamined when we first learn about and use simple measurements, such as the number of people in a room, and they remain so as we

[1] A genuine measurement is not the same as the thing being measured. A person's height exists independently of any measuring scale. If the height did not exist then the measuring scale would be useless (or perhaps worse than useless). This is a key point for this chapter.

begin to create and use more complex scales such as how tall the people in the room are, and even the much more complex scale of how much time they have been in the room for. However, all such everyday measures meet the requirements above. We can learn to identify people and separate them from each other, from their surroundings, and from the remaining contents of the room. We can sense when a room is more or less crowded with people, and begin to associate that sense with the numeric scale of 'digits' (for small numbers we might literally draw an analogy between how many people and how many fingers). This scale is shared, comprehensible, and reliable. The scale goes up or down by one for every new person entering or leaving the room. Barring computational mistakes, the scale is quite error-free, requiring no instruments, estimation, or fractions. It is a real number measuring scale.

Similarly, we can observe that the tops of some peoples' heads are further from the floor when standing upright, and that this characteristic remains relatively constant while they are in the room. We could even get people to stand in a line representing more or less of this characteristic of having their head further from the floor when standing upright. From such activity we could create an ordinal scale from shortest (less of this characteristic) to tallest (most). We could then select an item with a fixed amount of this characteristic (the shortest person, or a piece of furniture perhaps) and create a truer scale by using this item as a standard. The quantity of the characteristic can then be counted as a multiple of this standard. If the item used as a standard is commonly agreed or widespread, then this becomes the basis for a real number measuring scale, such as a ruler. It is more complex and so more prone to error than a scale of the number of people in a room – not least because tallness varies with posture, diurnally, and with age. But as long as we know these complications (or learn them once we have established the scale), we can take them into account when measuring tallness and when using the results for practical purposes. Time, or how long someone had been in the room, is more complex again and so even more prone to error, but it is similarly possible to convert to a quantity, and so arrive at a consensual, reliable and useful scale to measure this characteristic.

A key point in all three examples is the analogous behaviour of the scale used to measure and of the thing being measured. We can see when people come or go and how this relates to changes in the number of people counted in the room. We can see how much one person's height protrudes beyond another, and how this relates to variation in a scale (perhaps by standing both people and the scale standard against a wall). It is the isomorphism between the numbers and the observations that gives us a measurement, and this depends on our correct identification of a quantity to be measured in the first place (Berka 1983). I repeat, if people do not exist, or are not separated from their surroundings, or cannot be seen to protrude in relation to each other, then our quantities do not exist.

The examples so far are deliberately simple and everyday in scale, in order to illustrate the derivation of measurements from observations of the thing being measured (whether direct or indirect). Even these simple examples are imperfect, and require a certain number of assumptions that could be debated. The isomorphism between physical characteristics and measuring scales relies on imperfect judgement, for example. In social science, however, many researchers seek to go beyond relatively simple examples of creating measurements, such as number of pupils in a school. As soon as they do so, their measurement problems and concerns will multiply. This means that they should be much more careful, explicit and humble in their approach.[2] It is not the mere assignment of number to things that leads to a quantity. No amount of 'quantification' will, by itself, establish a good and useful measurement. Obviously, anyone can simply assign a number to an object or to the imagined gradations within a concept. In this case we are using numbers as convenient labels – such as the serial numbers allotted to examination candidates for anonymity. What we must not do is mistake this process for real measurement.

Unfortunately, in ignoring the underlying premises of measurement, social science research has become saddled with quite a range of quantified items that are *not* clearly good and useful measures. In many examples in popular use, such as attitude scales and the like, we cannot calibrate the purported scale to the real thing as we are not sure what the real thing is. The object of measurement may not exist, may not be identifiable or may not behave in an analogous manner to the numbers used as measures. For example, the problem with the well-known claim that 'intelligence is what IQ tests measure', for example, is not so much whether IQ tests measure intelligence, but whether they measure anything (Prandy 2002). We need to have a quantity identifiable by its qualities against which we can compare our measuring scale. If IQ is *really* only what IQ tests measure then IQ cannot be a real quantity, and so any attempt to measure it is liable to considerable confusion. The same might apply to 'attitude' measures.[3]

Explicitness is key to forming and using a good measure – and one that is reliable in the sense that it can reach the same result when used properly in the same situation by more than one researcher. If there is a dispute about the number of people in a room, it is possible to settle the matter by reference to the identifiable and separate things being measured. We have the measure (such as seven people) and we have bodies in a room that we can line up, and count repeatedly, until a consensus view is arrived at or a judgement becomes more general. Of course, there can still be errors of measurement but in principle they can be resolved by direct comparison of the figures and manifestation of the quality being measured. Resolving a dispute about the heights of people is more complex than about how many people there are, but still

[2] A measuring scale for a latent variable is inherently more problematic than for a simple measure like the number of people in a room. Both scales might be useful, and both might have problems, but the latter is the more likely to have serious problems, and the more likely in extreme to be so inaccurate as to be pseudo-quantification. I realise that there are a large number of latent measurement scales in popular use in social science. There is not the space here to discuss the validity of all of these scales individually. Nunnally (1975, p.9) says '…the best general approach… is to develop an instrument that correlates as highly as possible with a particular factor or intellect or personality. This requires that the items be homogeneous with respect to content… Thus, a good item for a vocabulary test is one that correlates well with the total score… methods of item analysis and test construction became clear: namely one should select those items that correlate well with the test as a whole and throw out those items that do not…'. This clear explanation from an expert in psychometrics lays bare the tautology involved in such purported measurements. The point about the spurious nature of this internal reliability is dealt with later in the chapter. Here, the key point for readers to note is that this passage only addresses the measure itself, and does not consider at all the correlation (isomorphism) between the measure and the thing it is supposedly measuring.

[3] To clarify, it is the isomorphism (correlation) between the behaviour of the measuring scale and the thing being measured that allows proper quantification. If IQ scales work well for the purposes intended, perhaps by accurately predicting performance in school or occupation, then they are a true, but probably still problematic, measure. IQ here is intended illustration the broader principle being proposed, that a true measuring scale must measure something other than itself.

possible. We can get others to compare the ruler (standard of height) with the lengths of the people. In each example, we have explicit, stable and easily re-counted evidence of a phenomenon and we have a separate system of measures. We can compare the two. The problem with extant efforts to operationalize IQ or attitudes is that we may have only the measurements. How then do we check for reliability or calibration or measurement error?

The rest of the chapter looks in more detail at this question, at the kinds of measurements used in social science, and assesses their suitability in terms of the ideas outlined so far.

Are there levels of measurement?

One of the aspects of measurement that is routinely proposed by social science researchers (instead of the importance of isomorphism perhaps) is that there are four levels of measurement (Stevens 1992). These are termed ratio, interval, ordinal, and nominal. The claim is usually made that we need to be aware of these four because their characteristics affect how we should use and analyse them. Is this true, or is it better to go back to the ideas of measurement outlined so far, such as the need for isomorphism between measurement and measured, and so avoid the terrible fate that apparently awaits anyone who confuses their levels of measurement?

A ratio measure is one which meets all of the criteria described in the last section. One cannot have a ratio measure of a non-identifiable quantity, for example. There is a direct analogy between numbers on the scale and properties or behaviour of the thing being measured. Most significantly, when there are four people in a room this is more than two people in a room; it is exactly twice as many. When someone is six standard units tall, they are three times as tall as someone two units tall, and so on. Conceptually, a further key point here is that the room can be empty, and the absence of a person will yield a measure of no units tall. Some scales can also have negative units, and these work in the mirror-opposite way to the positive units. Most people encounter such numbering systems in their everyday life, and most deal with them perfectly well without knowledge of measurement levels (or put another way any problems they have with such numbers are not usually about levels of measurement). Many of the figures used in social science, such as the examples at the start of the chapter, are similarly uncomplicated in this regard. These include the number of students and teachers in an establishment, the amount of government funding per pupil at school, the proportion of female pupils, the rate of participation of adults in part-time study, or the percentage of students deemed to have passed a test. All of these figures can be understood without reference to levels of measurement. Of course, many of them are much more complex and thus error-prone than a simple head-count. But it is not their ratio nature that makes them so. The complexity comes from real-life complications such as how to treat part-time or temporary staff, student mobility, missing data, or the definition of participation (see Gorard 2008).

An interval measure is usually described as being like a ratio measure in all respects but without a genuine zero point (such as no people in the room, or no passes in a test). Thus, an interval measurement must also be of an identifiable quantity, and the scale and quantity must be able to vary proportionately and in tandem. However, it is difficult to think of a real-life example of such a measure. A temperature scale, such as Centigrade, is the most commonly cited example but this is not a scale often used in social research. Economic indices like the FTSE100 may be another, or they may have a zero even if it is unlikely to be used. And the distinction makes little difference to how we use ratio and interval scales in practice, apart from recalling that 20 degrees is not twice as hot (whatever that means) as 10 degrees. We may add or subtract interval numbers (to find that one room is 10 degrees hotter than another, for example), and I have never seen a statistical methods resource that suggests different models or tests for ratio and interval values, despite claiming that we need to know about the distinction so that we do not use the wrong kind of analysis.

So-called nominal measures are, in fact, not measures at all but categories of things that can be counted.[4] The sex of an individual would, in traditional texts, be a nominal measure. But sex is clearly not a quantity – although each sex could be allocated a number for shorthand. Even where the categories are expressed as numbers for some reason, these numbers cannot be used for arithmetic, and are not measures since there is no isomorphism between the quantity and its measure. There is no quantity, other than the frequency of individuals in each category of the variable ‘sex’ – i.e. how many females and how many males. Frequencies can be added, subtracted, multiplied and divided just like ratio measures. Again, drawing a distinction here is pointlessly confusing. A frequency is a ratio measure. A category is not a measure at all. A category is a quality. Again, in everyday life there is little confusion over this.

The remaining level of measurement is termed ordinal, meaning that the scale has an order but no standard unit of measurement. An ordinal scale can be isomorphic with something observed, as long as the thing observed has an order to it but does not really increase or decrease along a continuum or in regular jumps. An ordinal scale should *not* be used, and would not then be isomorphic, simply because of ignorance about the way in which the thing of interest varies. It should not be simply a sloppy interval measure. Ordinal ‘measures’ are quite common in current education research, and they have so many associated problems and issues – the so-called parametric strategy, and the latent nature of what they purport to measure, amongst others – that I leave further discussion of these for the next section. Suffice to say here that a true ordinal scale like a nominal scale is not really a measure at all (as defined above), and the most widespread (and so commercially successful) scales based on these premises are treated and analysed in the same way as ratio and interval scales anyway.

Problems with ordinal ‘measures’

[4] Ratio and interval measures do not have to be continuous in their scaling, as should be clear from the examples used thus far. It is not the theoretically continuous nature of underlying temperature that makes Centigrade an interval scale. In everyday terms, the number of people in a room is based on a measuring scale involving only whole units. The saltatory nature of the scale used does not mean that it is not a ratio scale. So-called nominal measures are quite different, as the example of sex should make clear. The ideas of ratio and interval make no sense with this example, largely because the points on the scale (such as male or female) are not really measures at all, as defined so far in the chapter. They are merely categories. It makes no sense to ask whether male is some ratio of female. The measures involved are the frequencies of cases in each category, not the categories themselves. And frequencies are clearly on a ratio scale, despite what most textbooks might claim.

Ordinal numbers are the same as nominal numbers in referring simply to categories of things that can be counted, and that can be treated accordingly. Again, the identification of a different level of measurement seems unnecessarily complex. The distinction is based on the claim that some sets of categories have an intrinsic order to them, and these are termed ordinal, whereas some categories do not, and these are termed nominal. The given example of an ordinal value might be the grades achieved in a school examination or test. But for most analyses, this order makes no difference beyond the obvious. In calculating the number of students achieving a certain combination of grades, such as A to C, a teacher is not going to add in the frequencies of students in grade D, for example. In the same way with a nominal scale representing full and part-time employment and no employment, an analyst can calculate the number of all employed by adding the first two categories. They are not going to add in the unemployed category, except by mistake. Knowing the level of measurement here makes no practical difference. Whether a grade C is genuinely different to a grade D, and where the cut off point should be if so, is a difficult issue. But it is not one that is related to the level of measurement.

It is also important to realise that an intrinsic order is available for just about any set of categories. The serial numbers for TV channels on an old-fashioned analogue TV set could be treated as a nominal value having only arbitrary meanings, or could be placed or programmed in their order of appearance when flicking through with the channel ‘advance’ button, or the historical order in which they started broadcasting, or their transmission frequencies, and so on. All of these make sense, and appear to be at least as well ordered as many social or occupational class categories that commentators try to order in terms of skill or prestige (Rose 1996, Erikson and Goldthorpe 2002, Lambert 2002). Many categorical variables in research are not intrinsically either ordinal or nominal. Instead, it depends how one looks at them.

When a set of categories are clearly intended to be ordinal, such as the order in which competitors finish a sprint race, any ensuing analysis will usually have available the much better and more genuine measures that underlie the order, such as the times taken. It is difficult to imagine a situation in which it would be better to analyse by using the ordinal data than the evidence used to create that order in the first place. Of course, if there is no good evidence to create the order in the first place (no sprint times) then there is no good order anyway.

The main reason why ordinal categories are treated separately to other categorical data by some researchers is because they want to claim that these are actually more like real numbers than they are like nominal categories. That is, analysts want to treat a numerical order of categories as though they were real numbers that can be added, subtracted, multiplied and divided in a normal way. They want to be able to claim that a grade ‘A’ in an examination is worth 10 points and so five times as good as a grade ‘E’ scoring two points, for example. Therefore, if a grade B is worth 8 points, then a grade A minus a grade E is a grade B, and so on. This approach is common in statistical work, but it is of dubious value and rigour.[5] Categories such as examination

[5] This is not to suggest that an examination grade could not be given a numeric name. The point is that whether they are letters or numbers, such grades are neither ratio nor interval in nature. A grade 2 at GCE O level in the UK was in no real way twice as much of anything as a grade 1, and the difference both in frequency and exchange-value between a grade 1 and a 2 was not the same as between a 3 and a 4. The grades are categories rather than measures. The only measures they are based on are the

grades are thresholds superimposed onto theoretically continuous levels of variation. They are not equal interval, either in distance on a scale or in the proportion of cases within any grade. The difference in terms of underlying marks or performance between two students with different grades in an exam could be much less than the difference between two students with the same grades. The orders involved are not real numbers, and should not be treated as if they were. In everyday life, there is little confusion on this point. Only in traditional statistics is this confusion created, so that analysts can convert exam grades to points, and so on, so that they can then use a more convenient or more sophisticated analytic process. However, these analytic processes should only be available when researchers are in fact using real numbers (and often not even then, see below). What these analysts are doing is sacrificing clarity and even intellectual honesty in the name of technical sophistication.

In education and psychology research, a fairly common approach is to use questionnaires to measure latent variables like peoples' attitudes or their social 'capital'. A standard method to operationalise such constructs is to use scales of agreement and disagreement linked to items expressed as statements – such as Likert scales. The scale might vary from strongly agree, and slightly agree, to strongly disagree. These scale categories are then routinely converted to numbers (perhaps 1 for strongly agree, 2 for slightly agree and so on), despite the lack of equal interval between the points on the scale. Each item is then intended to be a component of the measurement of a construct (or composite index). The set of items involved in the construct are completed by respondents, and then tested for reliability by looking at the correlations between the responses to each item, treating the scaled ordinal responses as real numbers. There are so many things to query with such an approach that it is hard to know where to start, particularly given that it is so widespread in use.

How is it possible to measure something, in the sense of establishing an analogous relationship between the object measured and the numbering scale, when the object measured is latent? This is very different from a concept of length emerging from differential protrusions (see above). With length we can test the concept directly. With latent constructs we only have the constructs. Converting agreement scales to equal interval numbers means that the numbers and the scale would not be isomorphic anyway. Yet, ironically the approach relies much more heavily on the real quality of the measurements involved because we do not have a direct explicit comparator on which to model our measurement.

Also we are no longer dealing with a single measure, but a composite formed from many. Does this make the overall composite measure better (or worse)? The notion that several questions or indicators can be combined to produce a better answer than just one is premised on sampling theory. It assumes that the variability of each indicator is equal to every other and that this variance is due solely to random error.

underlying scores used to allocate the grades. To convert the scores into letter grades and then convert the grades into numbers and then use those numbers as though they were scores is a good illustration of the mess researchers get into, and widely accept, once the desire for quantification over-takes the principles of establishing a measure. Better by far to use the underlying scores (and what kind of measure these are is debatable, but not dealt with here). In the absence of these scores researchers must be careful when working with grades. They are not like the number of people in a room, since the underlying score for an individual with an A grade may be nearer to an individual with a B grade than two individuals with A grades are to each other. We can, however, use the frequencies of individuals in each grade category, or the probability of reaching a threshold grade, for example, as real numbers.

This seems very unlikely (Anderson and Zelditch 1968). How do we know that each item should have a specific weighting in creating the construct? Usually all items are given equal weight for no good reason (Nunnally 1975, p.8). The reliability of constructs is assessed in terms of correlation between items in the construct. If this is low, then non-matching items are rejected. But a high correlation is not, in itself, justification for a construct or latent measure. It could just mean that the researcher is asking the same question over and over again, and that one item would do instead. Nunnally (1975) suggests that in practical terms 'even a reliability of 0.90 is not high enough'. If the individual measures taken to create the overall construct are intended to be measuring the same thing to such an extent (0.9) then they will be repetitious (see below). If they are not intended to measure the same thing then how can we be justified in averaging them, or otherwise generating the composite? This is very different to contrasting two distinct measures, such as the number of pupils per teacher. It is more like those invalid UK league tables of Higher Education used in the press where the number of students with a first class degree is added to the number of computers and the library spend per annum etc.

One claim in response to this is that asking similar questions repeatedly leads to greater accuracy since random errors in responses are eliminated. Is this actually true? When we conduct a survey to create constructs, we must assume that a majority of respondents are able to answer any question correctly. If, on the other hand, a majority of respondents are not able to answer a question correctly then the survey is doomed from the start, and the use or non-use of constructs is irrelevant. 'Correctly' here means that respondents give the most appropriate response, that they would wish to retain on reflection, and which is supported by any available independent evidence. For example, someone born in the UK might correctly respond that they were born in the UK, retain this answer if allowed time to reflect on it, and could support it by reference to their birth certificate. A 'mistaken' answer, on the other hand, would be one that was inconsistent with other evidence, including the respondent's answers to other items within the construct.

When we conduct a survey to operationalise constructs, we assume that where respondents do not answer questions correctly then their errors are random in nature. We assume that errors are random for the sake of this argument because that is the assumption underlying the statistical treatment of constructs and their reliability. Of course, some incorrect responses will be systematic error (that is subject to non-random bias), but I will argue in the next section that these cannot be treated by statistical means, and that these techniques make the use or non-use of constructs irrelevant. For example, adult respondents might rarely overestimate their age, and more commonly understate their age, meaning that any measurement of age could have a slight bias. Techniques based on sampling theory cannot estimate or adjust for this bias.

In creating a construct we can only gain accuracy by asking different versions of the same question when two or more responses from the same individual differ. If we ask the same or very similar questions several times we gain nothing if the answer to each version of the question is consistent with every other. For example, if an individual responds correctly with their age to one question and their valid date of birth to another question then the response to one of the questions alone is sufficient. We only learn something from repetition of the question when the second response differs substantially from the first.

Therefore, it follows that most people who change their answer between the first usually correct answer and the second answer will be making a mistake on the second answer. Some respondents will give the same answer both times, whether correct or incorrect, but these are unaffected by the repetition. If the errors are truly random then a certain proportion of those who gave the correct answer to the first question will give an incorrect answer to the second version of the same question. A similar proportion of those who gave an incorrect answer to the first question will give a correct answer to the second version of the same question. But there will be more people giving a correct answer than an incorrect answer first time (see the first assumption). Therefore, for those answers that differ, the answers to the second question will be more likely to be in error than the first answer.

Therefore, asking a question more than once reduces rather than increases the accuracy of the evidence we collect from the survey. This conclusion has been confirmed by simulation.[6] Why do many researchers not notice this problem? I suspect because, unlike a simple measure like length, they are dealing with something that they cannot otherwise calibrate. In such situations, we can have no idea of the ‘correctness’ of responses, other than their consistency. We are stuck with a circular definition like IQ – the construct is what the construct measures. Or perhaps researchers are confused about the issue of reliability, within items of the construct (rather than across test-retest situations). Where items in a construct have high internal consistency or high reliability they are merely asking the same thing repeatedly (in the same way that if we know someone’s date of birth we also know their age). Where they do not have this high internal consistency, they are not asking the same question in effect, and so are not the subject of this discussion.

The suggestion that in building constructs we create a more accurate measure of the underlying variable fails in almost every way. Overall, surveys are better at collecting 'factual' information like date of birth than 'imaginary' data such as attitudes to car ownership.[7] However, because constructs are usually based on imaginary things that we cannot confirm apart from with the use of similar survey instruments with exactly the same limitations, and since we have no way of knowing the real answer, we can never test empirically what we gain in accuracy by asking questions several times in different ways. We do know that there are clear opportunity costs. Asking what is

[6] For a simple example, using an Excel spreadsheet, create a large set of random numbers (perhaps 1 or 0) in one column. Treat this as the underlying theoretically ‘true’ answer sought by an instrument (of course, this is the essential knowledge we lack in practice and which makes latent measurement so much more dangerous than everyday measurement). Choose an overall ‘accuracy’ of response level such as 90%, and create a second column identical to the first except that each number has a 10% chance of randomly flipping from 0 to 1 or vice versa. Treat this as the response to the first item in the instrument. Create a third column from the first in the same way. Treat this as the response to the second item in the instrument. Both sets of ‘responses’ will be around 90% accurate in isolation. What do we gain or lose by aggregating the two item responses? Try it. You will find, however you resolve differences between items (does 0 followed by 1 count as half, for example?), that the combined result is less than 90% accurate. The more items you add, the less accurate the aggregate result becomes.

[7] An individual’s views on cars might be more properly sought via in-depth conversation or, better yet, observation.

effectively the same question again and again leaves less room in a questionnaire of fixed length for questions on other things. The repetition also leads to boredom among respondents, which may affect response rates and the integrity of responses. Using only the best single item (perhaps the one with the highest loading in a factor analysis generated when developing the questionnaire) allows us to reduce the length of the questionnaire in order to increase completion and response rates (once the measure has been developed for use). This, it seems to me, is a far surer path to quality than using constructs in areas other than assessment. Constructs such as multiple item tests *can* work better in assessment for a number of reasons, most notably because they are more properly measurements given that the composer can have a good idea beforehand of what the correct response should be to each item. In examinations there is, in essence, a correct answer (or answers) for which marks are awarded, and these can be moderated for consistency between examiners. This is not to defend assessment *sui generis* but to point out that, in its own terms, it seeks to find out whether the candidate can correctly describe an answer that is already known. Attitude scales, on the other hand, generally seek to find out something so far unknown. The ultimate arbiter of the attitude is the respondent not the examiner.

I repeat, the problem with constructs such as attitudes and preferences is that there is no explicit thing – e.g. the correct answer, or the number of people in the room – with which the purported measure can be compared. Therefore, it is hard to see attitude scales and the like as real measures of anything very much. Even in those situations where they correlate highly with something explicit (a human behaviour for example) there is a danger of tautology. For example, people who choose to go swimming might be more likely to report wanting to go swimming, and so on. In this case, where we have the behaviour itself it is a better measure than the report of preference or intention – which is a weak proxy. It is also often, when tested in non-tautological conditions, an inaccurate proxy. For example, students' reported attitudes to science at school are weakly but inversely related to their subsequent choice of studying science subjects or not (Gorard and See 2009).

The behaviour of errors in measurement

Knowing whether we have the correct figure is a key component of measuring. Whatever kind of measurement we use, and however good the scale we devise, there are likely to be some errors in the figures that emerge. It is a condition of good and safe measurement that we have some idea of the scale and possible impact of such errors, and I focus on this issue here because it is usually ignored in traditional resources in favour of concerns over sampling variation (see next section). A measure is intended to react isomorphically with the quantity of which it is a measure. Measurement error is an indication of the failure of that isomorphism. As such, it has nothing to do with sampling variation. Measurement error can come from mis-specification of the scale, mis-calibration, mis-reading, mis-recording, and errors in transcription or recall. One common source of simple mistakes comes from copying, such as when transferring a long list of numbers from paper to computer or calculator. Intriguingly, errors are also introduced merely by entering numbers into a computer or calculator even when the numbers are entered entirely correctly. A computer stores all of its numbers in binary, and only allocates a specific number of binary digits to each number. The process of converting base ten numbers into binary numbers, therefore, automatically introduces small errors to some numbers, even where they start out as perfect representations of the reality we are measuring.[8] We have no reason to assume that all such sources of error are random in nature; in fact we have a considerable body of evidence to show that they are non-random (Gorard 2006a). Researchers are more likely to mis-read or mis-type data in such a way as to support their favoured ideas, for example. A ruler that was calibrated to be too short would consistently over-estimate lengths, rather than provide any kind of random variation around the 'true' length. Thus, the techniques of statistical analysis based on random error and sampling theory are of no general use in assessing the general quality of a measurement. Standard errors, confidence intervals, and significance levels, for example, do not, and cannot begin to, help us estimate the importance of such errors in our measurements.

In research situations there are many sources of genuine error that are unaddressed by traditional statistical analyses concerned only with sampling variation. Many purported measures in common use in social science probably contain a very high

[8] It is well-known that some numbers in denary (decimals) are irrational, like pi, and that others would take an infinite number of decimal places, like the ratio 1/3. Use of figures such as these in analysis (other than mathematics) involves putting up with a minor representational error in practice. The size and therefore the impact of that error is best considered in relative terms, as a fraction of the number in which it is an error. We do not always know this relative error precisely, but we can estimate its upper bound (for example, in representing the number 1/3 as a 6 figure decimal fraction: 0.33333, the maiximum error would be bounded by 0.00001). It should also be well-known to all readers that both the absolute and relative representational errors in figures alter when the number base is altered. The unwieldy decimal of the ratio 1/3 is handled easily and with no error term in number base three (as 0.1), for example. The apparently simple decimal fraction 1/10 (or 0.1), on the other hand, cannot be represented accurately in one byte (8 binary digits). The closest we could achieve in binary would be 0.0001100 which is equivalent to 0.09375 in decimal, or 1/16 plus 1/32. We introduce an error of more than 6% simply by storing the decimal 0.1 on the computer! Extending the number of bits used to represent the decimal ratio 1/10 reduces the relative error but will never eliminate it (the absolute error in 8 bits is .00625 or 1/10 of 1/16 and so the sequence simply repeats *ad infinitum*). There are, by definition, an infinite number of such potential problems when using a computer or calculator, but they are systematic to the number bases involved. They are certainly not random, and cannot be dealt with by any technique predicated on them being random. Three further points should be stressed for readers who have not considered these elementary concerns for measurement before. First, the example of a never-ending binary number is used as an illustration. These kinds of representational errors can occur though with any value held in a limited number of bits (as all must be). Second, these representational errors are, of course, additional to errors of any kind in the measuring process itself. Third, the initial relative error in any measurement changes during analysis so that, on average, the relative error will increase half of the time on every single arithmetic step of a calculation. Yet analysts almost never check the behaviour of these systematic errors and their propagation, relying on the false assurance of rituals based on random sampling theory. The relative error can quite easily grow to such proportions that we say the analyis is 'ill-conditioned', meaning that the manifest result is terribly sensitive, and owes more to the error than to the substantive measurements involved (Gorard 2001). It is very possible that whole areas of social science are based on such ill-conditioned models. The field of school effectiveness is one such candidate, in which assessment scores at two points in time are subtracted from each other to leave most of the 'gain' score susceptible to high relative error propagating from the inevitable errors in both sets of assessment scores. One 'analyst' - a top professor of school effectiveness no less - once remonstrated with me that school effectiveness cannot be so flawed since they had found one school in an area, which had had four successive years of positive value-added. If I tell readers that there were at least 16 schools in the area, that the 'effective' school had been identified *post hoc*, and that no other school in the area had the same four year record, then I hope all of you can see why this claim is so ridiculous. If not, I point out that there was also one school with four successive years of negative scores, one with three negative followed by one positive, one with two negative followed by a positive and a negative, etc.

proportion of measurement error.[9] These include attitude scales, the categorisation of ethnic groups or occupational classes, the allocation of national examination grades, or definitions of poverty. What they all have in common is a high level of imprecision – some, like attitude scales, more because of the vagueness of what is being measured, and some, like the allocation of national examination grades, more because the size of the operation leads to mistakes and imperfect moderation between assessors. While we can take steps to minimise the chances of the second kind of error, we cannot guarantee their absence.

Why do errors matter so much? Surely, we can just use the best available figures, analyse them accurately, and be aware that our results will not be perfectly accurate. Well – consider the following example. We calculate the mean score in a school test for a group of boys, and find the answer 60%. We also calculate the mean for a group of girls, and find the answer 70%. Let us imagine that both means are around 90% accurate. An error component of one part in ten may seem reasonable when considering either figure in isolation. But this means that the real boys' score actually lies between 54% and 66% (i.e. 60 plus or minus 6). The real girls' score lies between 63 and 77 (i.e. 70 plus or minus 7). If we subtract the girls' mean from the boys', the manifest answer is 10. But the true answer could be anywhere between 23 (77-54) and -3 (63-66). Remember, these are not probability calculations. Confidence intervals are no help here. The difference -3 is as likely as any other answer. Therefore, despite both figures initially being 90% accurate, when we subtract the two figures we genuinely have no idea whether the result is positive or negative. The actual range for the answer, from -3 to +23 (or 26), is nearly three times the size of the surface answer (10). The range of measurement error involved is far greater than the range of the answer itself. The relative measurement error in the original measures has been propagated by the ensuing simple calculation.

This propagation of initial measurement errors is usually ignored in training texts, because it remains largely unaddressed by traditional statistical analysis. Even where we have relatively accurate initial figures to work with, the process of conducting arithmetic with numbers changes the size of the error component relative to the numbers. Even such a simple technique as above, when subtracting two numbers, manages to convert an error of just 10% to an error of 260% of the figures involved. What has happened is that the initial numbers were so close that the subtraction makes them almost negligible in the result, leaving the answer to be made up almost entirely of error. This happens in real-life calculations all of the time. You may be able to

[9] In quite simple everyday kinds of measurement that we might use in social science, even where completion of a questionnaire, for example, is compulsory by law, and where we have access to some way of judging, it is clear that we have high levels of missing data. According to official statistics in the UK the most prevalent cell in any analysis of the social class of students in higher education is 'unknown' (Gorard et al. 2007). In subsequently assessing the proportion of students in any social class, the measurement error introduced by this missing data (of the order of 30%) must be added to any concerns we have about the class categories themselves (such as their timeliness or their equal appropriateness for all ages and sexes, and their threshold effect), and to any errors made by those 70% of respondents who did answer, plus errors in entering or coding the responses, plus any representational errors (see above). I think it is fair to say that this relatively simple measure contains a very high proportion of measurement error. Of course, with psychometric data and similar it is not so easy to estimate the measurement error (one of the main points made in this chapter), but let us not allow this inevitable ignorance to fool us into working on the basis that the relative error will be any less than in the social class example.

imagine what happens to the error components in more complex analyses. Value-added analyses of school performance, for example, routinely involve finding the difference between two sets of similar-sized numbers each with high levels of initial error. Knowing about the propagation of errors, how much faith would you want to put in the results?

Two important points emerge from this consideration. First, there is no standard acceptable amount of error in any measurement. The relevance of the error component is a function of scale, and of the use to which the measurement is put. Second, the relative size of the error in any result is not determined by the accuracy of the original measurements. It depends on the precise steps in the ensuing computation. Of course, it helps if the initial readings are as accurate as possible, but whether they are accurate enough depends on what is done with the readings, and how the error component propagates as it is used in calculations. It is important to recall that every statistical, arithmetic, or mathematical operation conducted with measurements is also conducted with their error components. If we square a variable, then we also square its error component and so on. The more complex the calculations we conduct the harder it is to track the propagation of errors (even if we are aware of them at source), and so make an informed judgement of the ratio of error to final result. In extreme cases, the manipulation of variables leads to results almost entirely determined by the initial errors and very little influenced by the initial measurements. When answering a typical analytic question, such as ‘is there a real difference between two figures’, we need to take into account, in a way that traditional analysis simply ignores, the likely scale of any errors. Why is this not done as standard?

Misplaced emphasis on random errors

One of the main reasons why consideration of real measurement error issues, such as bias and propagation, are routinely ignored is the false belief that measurement error is handled by statistical testing, and estimation. The basis of statistical testing derived from sampling theory is the calculation of a conditional probability. This probability becomes the p-value used for significance testing, and also for standard errors, confidence intervals and often in deciding which variables to retain in complex statistical modelling. The calculation of the probability is grounded in various assumptions, such as a random or randomised sample and complete measurement of all cases in the selected sample (de Vaus 2002). The probability is intended to help decide whether two sets of measurements (perhaps two sub-samples) could have come from the same overall group (see Gorard 2010 for more discussion of this).

> Application of NHST [null hypothesis significance testing] to the difference between two means yields a value of p, the theoretical probability that if two samples of the size of those used had been drawn at random from the same population, the statistical test would have yielded a statistic (e.g., t) as large or larger than the one obtained. (Nickerson 2000, p.242)

If the p-value calculated in this way is less than a certain level, traditionally 5%, then the null hypothesis (often the ‘nil’ null hypothesis of no difference) is rejected. Standard reputable texts on methods for statistical analysis are in agreement over the conditional nature of null hypothesis testing, even though they may all express this condition in different ways (Carver 1978).

> The final step in the testing process is to see whether the proportion from the random sample is sufficiently far from the proportion assumed by the null hypothesis to warrant the rejection of the null hypothesis. (Fielding and Gilbert 2000, p.248)

> The null hypothesis is presumed true until statistical evidence, in the form of a hypothesis test, indicates otherwise — that is, when the researcher has a certain degree of confidence, usually 95% to 99%, that the data does not support the null hypothesis. (Wikipedia – 29/05/08)

The probability of the data/sample encountered in any new research is conditional upon the null hypothesis (rather than the other way around), and is routinely used to accept or reject the null hypothesis, where the null hypothesis is usually one of no difference. Why does this matter?

The logic is a modified form of the argument of *modus tollendo tollens*, or denying the consequent. If the null hypothesis is assumed true we can calculate the probability of observing data as (or more) extreme than the data we did observe. If this probability is very small it suggests that the null hypothesis is not likely to be true. Of course, the formal *modus tollens* is not based on probability but certainty. The argument goes:

> If A then B
> Not B
> Therefore, not A

This argument is indisputable, and its soundness can be proved by logic trees or truth tables. However, when converted to a likelihood argument, the argument says:

> If the null hypothesis (H) is true then we can calculate the probability of observing data (D) this extreme
> The probability of D (given H) is very small
> Therefore, the probability of H is very small

This second argument is more complex than the simple *modus tollens*. Unlike the formal logic version, it allows for a false positive result (or Type I error) when H is rejected incorrectly because the small probability of observing D has occurred despite H (as it will on a small percentage of occasions). By definition a Type I error (incorrect rejection of the null hypothesis) is only possible if the null hypothesis is true. The truth of the null hypothesis itself is not derivable from p since p only exists assuming the null hypothesis to be true. Many specific events, when closely described, can be deemed low probability, and the occurrence of a low probability event is not, in itself, any evidence of it not being due to chance (as this term is usually interpreted, Gorard 2002). Sitting down to a game, the gamer would not declare a die biased just because it rolled two threes in succession. Nor would they do so if they were dealt three cards in the same suit from a standard pack. Yet both of these gaming events are less likely than the 5% threshold used in traditional statistical analysis.

This likelihood version of *modus tollens* also allows for a false negative result (or Type II error) where the probability of D is found to be not very small, and so H is not rejected even where it is not true. In general, users of statistics tend to understand these limitations of the method, either formally or intuitively. They may reason that as long as the conditional probabilities of D and H are clearly linked, so that a low value for p(D|H) means a low value for p(H|D) and *vice versa*, the approach is useful and valuable. As Nickerson (2000, p.251) says:

> …to many specialists, I suspect, it seems natural when one obtains a small value of p from a statistical significance test to conclude that the probability that the null hypothesis is true must also be very small.

But this is not so at all. Reversing D and H leads to two statements that are clearly not equivalent in real life (or indeed in science, logic, maths, or statistics). The probability of carrying an umbrella if it is raining is not the same as the probability of it raining if one carries an umbrella. Modus tollens does not work with likelihoods.

The p-value for a statistical test is the probability of getting the data we did given that the null hypothesis of no difference between sample and population (or two sub-samples) is true. For the p-value to be calculated the null hypothesis must be assumed as true. By definition, therefore, the p-value by itself cannot be used to judge or even help judge the likelihood of the truth of the null hypothesis. Yet what analysts seem to want from statistical testing is precisely that ability to judge the probability of the null hypothesis. And this is what they usually report the p-value as being, and how they use it in practice. The widespread practice of significance testing rests on a crucial confusion between these two conditional probabilities.

The assumptions of significance testing only really work when the prior probability of the null (and alternate) hypothesis is 50%. In this unlikely circumstance the probability of the data observed given the null hypothesis - p(D|H) - is the same as p(H|D) and so the null hypothesis could be plausibly retained or rejected with likelihood p(D|H). Even then, of course, it is still not clear that one study leading, however convolutedly, to $p(H|D)<0.05$ should be able to over-ride completely a prior p(H) of 0.50 in this manner. In other circumstances, indeed in most conceivable circumstances, even this is not possible, which means that statisticians not only have to claim a series of rather unrealistic assumptions such as a perfect random sample with full response, no dropout and no measurement error, they also have to claim that all of their null hypotheses have a 50% likelihood before any data is collected or analysed. In other words, statistical analysis must assume no prior knowledge of any kind.

Nickerson (2000, p.247) provides numerous examples of specialists in statistics routinely ignoring this distinction between pD|H and pH|D, and shows that false beliefs about these conditional probabilities are ‘abundant’ in the literature. These examples include social scientists and statisticians ‘of some eminence’, such as Fisher himself. This ‘belief that p is the probability that the null hypothesis is true’ (Nickerson 2000, p.247) is an even more fundamental error than the less common but still prevalent mistaken belief that 1-p is the probability of any specific alternate hypothesis being true.

As analysts, we can convert the probability of getting the data we did given the null hypothesis into the more useful probability of the null hypothesis given the data we obtained, by using Bayes' Theorem. But to do this requires us to know p(H) – the unconditional probability of the null hypothesis being true in the first place. So, apparently tautologically, in order to find the empirical probability of the null hypothesis we must know how likely it is beforehand. Analysts might reasonably use their *a priori* subjective judgement of the null hypothesis being true to substitute for p(H). In which case the new p-value is an estimate of how far that subjective view of the null hypothesis might be affected by the new empirical evidence (see Gorard et al. 2004b). But almost none of the material published in mainstream social science research journals, follows such a procedure. Most analysts simply appear to misinterpret the conditional probability of the new data as though it were the probability of the null hypothesis conditional upon the new data. Falk and Greenbaum (1995) claim that the lack of a clear relationship between pD|H and pH|D discredits the whole logic of significance testing, which answers a question we would never knowingly ask, but leads us to conclude we have some kind of answer to the questions we might actually want answered – such as how probable is the hypothesis, how reliable are the results, and what is the size of the effect found?

There have been many criticisms of statistical testing over decades (Gorard 2006a), and many examples of misuse of the method such as widespread acceptance of p-values based on non-probability samples, or dredging datasets via multiple use of a technique whose probability calculations are predicated on one-off use (Wright 2003). What is demonstrated in this chapter is far more fundamental. It is that the easy assumption that a low value for p(E|H) means a low value for p(H|D) is false. *Modus tollens* does not work with probabilities – or, expressed differently, it requires a further probability in order to make it work at all. Since at least the writing of Jeffreys (1937) it has been well established that a small value of p(D|H) (such as less than 0.05) can be associated with a probability of H that is actually near 1. A null hypothesis significance test is therefore:

> …based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research. (Rozeboom 1960, p.417)

Unfortunately, reflective consideration of errors in measurement and analysis has been almost completely overshadowed by a mechanical procedure designed for dealing only with random variation. This procedure is almost impossible to use correctly as intended in social research (as opposed to agricultural trials, for example), and even when used correctly has a logical flaw. It deals only with random variation, which our consideration of genuine measurement errors shows to be only a small part of the problem. And as now revealed, it is a fatally flawed way even of dealing with random variation. Nothing like this based on sampling theory and analysis of random variation can save ill-designed number scales applied to obscure concepts from the charge of pseudo-quantification. On the other hand, rejecting significance tests makes the use of numbers less problematic and so encourages their use in research (Gorard 2006b). We should continue where possible to randomise our cases to treatment groups in trial designs, and to select samples at random in passive designs where population figures are not possible. We do this because it helps minimise systematic bias, not because it means we can then use significance tests and everything that stems from techniques predicated on random errors.

Rethinking the use of measurement

If accepted, the argument here suggests that confusion between the measurement of observable events and the habit of assigning numbers to imagined events (including perceptions, attitudes, and intentions) has possible dangers. These dangers include the opportunity costs of conducting research with flawed techniques when the time, money, effort, and access to research sites could have been used to better effect. They include the vanishing breakthroughs that occur when insecure research knowledge is rolled out into policy or practice (Harlow et al. 1997). Perhaps almost as importantly, the possible dangers include the ethical and methodological distortion of new researchers by their mentors. Let the pendulum swing back a little towards scepticism about the easy allocation of numbers to things, and about the replacement of the basic pre-technical steps in creating a measurement by increasingly complex models and techniques. Let us think a little more (but a little less defensively) about the real process of measurement. Perhaps we can then help build the capacity to find and use appropriate measures in social science, that will be of genuine help to the societies we are ostensibly doing the research to benefit.

References

Anderson, T. and Zelditch, M. (1968) *A basic course in statistics, with sociological applications*, London: Holt, Rinehart and Winston

Berka, K (1983) *Measurement: its concepts, theories and problems*, London: Reidel

Carver, R. (1978) The case against statistical significance testing, *Harvard Educational Review*, 48, 378-399

de Vaus, D. (2002) *Analyzing social science data: 50 key problems in data analysis,* (London: Sage)

Erikson, R. and Goldthorpe, J. (2002) Intergenerational inequality: a sociological perspective, *Journal of Economic Perspectives*, 16, 3, 31-44

Falk, R. and Greenbaum. C. (1995) Significance tests die hard: the amazing persistence of a probabilistic misconception, *Theory and Psychology*, 5, 75-98

Fielding, J. and Gilbert, N. (2000) *Understanding social statistics*, London: Sage

Gorard, S. (2001) *Quantitative Methods in Educational Research: The role of numbers made easy*, London: Continuum

]Gorard, S. (2002) The role of causal models in education as a social science, *Evaluation and Research in Education*, 16, 1, 51-65

Gorard, S. (2006a) Towards a judgement-based statistical analysis, *British Journal of Sociology of Education*, 27, 1, 67-80

Gorard, S. (2006b) *Using everyday numbers effectively in research:* Not *a book about statistics*, London: Continuum

Gorard, S. (2008) Who is missing from higher education?, *Cambridge Journal of Education*, 38, 3, 421-437

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, (forthcoming)

Gorard, S. and See, BH. (2009) The impact of SES on participation and attainment in science, *Studies in Science Education*, (forthcoming)

Gorard, S., Prandy, K. and Roberts, K. (2002) *Introduction to the simple role of numbers*, Occasional Paper 53, Cardiff University School of Social Sciences

Gorard, S., Rushforth, K. and Taylor, C. (2004a) Is there a shortage of quantitative work in education research?, *Oxford Review of Education*, 30, 3, 371-395

Gorard, S., Roberts, K. and Taylor, C. (2004b) What kind of creature is a design experiment?, *British Educational Research Journal*, 30, 4, 575-590

Gorard, S., with Adnett, N., May, H., Slack, K., Smith, E. and Thomas, L. (2007) *Overcoming barriers to HE*, Stoke-on-Trent: Trentham Books

Harlow, L., Mulaik, S. and Steiger, J. (1997) *What if there were no significance tests?*, Marwah, NJ: Lawrence Erlbaum

Jeffreys, H. (1937) *Theory of probability*, Oxford: Oxford University Press

Lambert, P. (2002) Handling occupational information, *Building Research Capacity*, 4, pp.9-12

Nickerson, R. (2000) Null hypothesis significance testing: a review of an old and continuing controversy, *Psychological Methods*, 5, 2, 241-301

Nunnally. J. (1975) Psychometric theory 25 years ago and now, Educational Researcher, 4, 7, 7-21

Nuttall, D. (1987) The validity of assessments, *European Journal of Psychology of Education*, 11, 2, 109-118

Prandy, K. (2002) Measuring quantities: the qualitative foundation of quantity, *Building Research Capacity*, 2, 2-3

Rose, D. (1996) Official social classifications in the UK, *Social Research Update*, 9, 1-6

Rozeboom, W. (1960) The fallacy of the null hypothesis significance test, *Psychological Bulletin*, 57, 416-428

Stevens, J. (1992) *Applied Multivariate Statistics for the Social Sciences*, London: Lawrence Erlbaum

Wright, D. (2003) Making friends with your data: improving how statistics are conducted and reported, *British Journal of Educational Psychology*, 73, 123-136