**Editorial: the deceit of numbers**

When we see a claim that invokes numbers, our first thoughts are to accept it. If that claim, though, is in a newspaper or on social media we may take pause and, on reflection, ask: where did that number come from? How do they know? Is this genuine or a convenient number picked out of a host of competing surmises? The scepticism we express will, in part, be result of our suspicion of bias: political, ideological, campaigning. In short, we will simply dismiss the claim or at best interrogate it to explore its veracity.

Yet if an argument in a learned journal is corroborated by dint of some numbers, we are much more inclined to accept them: in some cases, would not accept a conclusion unless it was ‘proven’ quantitatively. Our assumption is that it is a refereed journal, therefore the numbers are credible. In the main, the only concerns we express is if there seems to be something faulty in the computations or perhaps something wrong with the interpretation. What we don’t do is fundamentally excavate beyond the surface deceit of numbers.

Numbers are deceptive. An enumerated argument is seen as more ‘objective’, ‘scientific’ and truthful than one that does not resort to numerical assertions. The problem with framing an argument around numbers is that the numbers take on a life of their own and become detached from the construct they are enumerating.

There have been many texts written in the last fifty years challenging the presumed objectivity of quantitative methods, ranging from Huff’s (1973) *How to Lie with Statistics* to Blauw’s (2020) recent *The Number Bias*, augmented by specialist critiques such as Morrison & Henkel (Eds.) (1970) excellent *The Significance Test Controversy*. Much of the point of these texts, despite the salience and perceptiveness of the critique of quantification, gets swept under the carpet of enumeration.

What are the weaknesses in quantification? There are several areas of deceit ranging from the operationalisation of the concepts, through sampling and inappropriate statistical techniques to misrepresentation and misleading interpretation of the data. Each will be explored.

**Operationalisation**

Operationalisation is the process through which abstract concepts are translated into measurable variables. In some disciplines this can be fairly easily achieved, heat is operationalised by measurement on a standardised thermometer. In other disciplines, such as sociology, criminology, economics, history, education and psychology (henceforth social science), operationalisation is far more complex: how do you operationalise alienation, poverty or satisfaction?

As explained in the *Social Research Glossary* (Harvey 2012–24a), operationalised concepts are related to theoretical concepts but are not coincident with them. Operationalisation is the process through which social phenomena are selected as indicators of social concepts. How this should work, is as follows. First, the concept is defined to make clear what is being studied and what needs operationalising. Second, appropriate indicators, which are observable and practically measurable entities, are selected by which to identify (and possibly measure) the concept under investigation. Sometimes a single indicator is used and sometimes a battery of indicators is used and combined in some way to form an index representing the concept being studied.

For example, an attitude questionnaire exploring student satisfaction involves a set of questions. The choice of questions is ultimately a subjective decision of the researcher based on preference and prior experience and supposedly mediated by the practical trial (and error) process of the pre-pilot and pilot surveys.

The key problem with operationalisation, then, is that of validity. How can one be sure that the operational measurement actually measures the theoretical concept? There are no ways in which validity can be ‘tested’: at best the operationalisation is painstakingly aligned with the theoretical concept through argument and explanation; at worst a convenient measure ‘will do’ (such as the percentage of overseas students enrolled as an indicator of internationalisation).

Of course, the scrupulous researcher will take considerable care in in breaking down concepts into component parts and ensuring that the best effort is made to find appropriate indicators at each stage. Unfortunately, vast amounts of social research use whatever indicators are convenient or are constructed without thorough analysis of appropriateness.

This process is exacerbated by the combining of indicators into an index. Again, the salience of each component indicator should be explored in detail and the weighting ascribed to each in the final index similarly exhaustively analysed. Most of the time, though, the index is a simple aggregation of indicators, often, ironically, validated by statistical analysis. For example, a satisfaction survey of 10 questions with fixed five-point ordinal scale responses will generate a single outcome that is just the summation of ten scores, ranging from 10 to 50, or the average score for the 10 questions, ranging from 1 to 5. This, irrespective of how important are the individual indicators that make up the index. So, for example, ‘the lecturer turns up on time’ is afforded the same importance as ‘the lecturer explains the concepts clearly’, both lose their conceptual origin as they are converted into a convenient number.

***Interchangeability of indicators***

One form of legitimating the indicator selection process rests on the pragmatic (and theoretically weak) notion of the interchangeability of indicators. This refers to the assertion by some quantitative practitioners who use multivariate analysis that, given there are a large number of potential acceptable items that could be used as indicative of a dimension of an operationalised concept, then any one indicator from that extensive pool is as good as any other indicator, that is, they are interchangeable.

However, any one item from the pool will not necessarily classify a given *individual* respondent in the same way as any other item. This is not important, though, provided the indicators all classify the subject *group* in approximately the same way. More important still, the whole point of multivariate analysis is to show relationships *between* different concepts, it is not interested in individual characteristics.

The argument runs as follows. An indicator X1 of concept C will be as good as any other theoretically sound indicator X2 of concept C for the purposes of multi-variate analysis because, empirically, the correlation of X1 with another dependent variable, Y, will be more or less the same as that of X2 with Y.

While individual people will be categorised in different ways in respect of concept C for each of the two indicators, the *group* as a whole will exhibit the same overall pattern for X1 as for X2. More important, the correlation of X1 with Y for the group will be the same (more or less) to that of X2 with Y.

Given sampling variation, this means that it does not really matter which, of a potentially large (or even infinite) number of possible theoretically sound indicators one chooses. In short, the argument obviates the need to worry about the subjective process of indicator selection. This, then, to some extent, appears to circumvent the problem of validity. (Harvey 2012–24b)

However, it is an illusory circumvention, justified by ‘common-sense’ exhortations, empirical examples and tenuous demonstrations devoid of any solid theoretical base. It is also to some extent a tautological argument as any indicator that is at variance with others can be disregarded as being unsound, or unreliable and therefore ‘invalid’. Interchangeability is a contrived way of legitimating the subjective element in what purports to be an objective process.

Commenting on the then proposed national student survey, Harvey (2003) explained how interchangeability of indicators was being misused:

I am all for institutions making their internal feedback available to prospective students. The proposed approach, though, is laughable in its pointlessness. The pilot, for example, assembled nine statements on teaching with which respondents might agree or disagree on a five-point scale. These are averaged and a teaching score generated ranging from one to five—a low score being more positive than a high one… What do the average scores show? What does 1.5 for teaching mean? Well, it means students quite strongly agree that teaching is... is what? Well, better than if it had scored 3.4, but maybe not quite as good as if it had scored 1.3. But what is it about teaching that this score represents? The whole scheme is based on the “interchangeability of indicators” thesis developed, pragmatically not theoretically, by sociologist Paul Lazarsfeld and colleagues in the early 1960s. It assumes that there is a concept called teaching and that any set of an unspecified subgroup of similar indicators is as good as any other for measuring the concept. Various statistical manipulations, such as factor analysis, “prove” this. But the whole process is based on an invalid presupposition—that the concept “teaching” is unidimensional. If it isn’t— and it isn’t—the average is meaningless.

Similarly, Bukodi *et a*l. (undated), referring to their work on children’s social origins and their levels of educational attainment explained that a major problem is that:

of determining how far differing findings are real or artefactual: … Some progress has been made in standardising the conceptualisation and measurement of educational attainment. But a potentially far more serious—and far less considered—problem remains with social origins. The assumption seems often to have been made, if only implicitly, of the ‘interchangeability of indicators’: i.e. it has been assumed that in whatever way social origins might be conceptualised and measured, much the same results would be obtained as regards associated differences in children’s levels of educational attainment.

In short, this non-theoretical legitimation needs to be approached with care and applied only where circumstances clearly show a conceptual mutuality, rather than being applied out of convenience. Furthermore, this is a device relevant to multivariate analysis and adopting it, even implicitly, in other circumstances would require substantial legitimation pertinent to the specific circumstance. This rarely occurs in submitted articles.

Moving on from operationalisation to sampling.

**Sampling**

The majority of social research is undertaken using samples; population analysis is relatively rare. Using samples raises substantial questions that, too often, are glossed over in submitted research. They include the size and nature of the sample, the extent of non-response and the sample’s representativeness of the population about which it is making claims.

Most statistical analysis requires random samples if the researcher intends to make generalisations from the results of the sample. A random sample is one in which all members of the population have an equal chance of being in the sample (an unbiased sample). To do this, there needs to be some form of sampling frame that identifies the population and enables the researcher to undertake a process that produces an unbiased sample.

A large proportion of samples used in social research are non-random. They are usually convenience sample, that is, samples of people who are easily contactable in one form or another. Sometimes, there are statements suggesting that although a non-random sample, the sample is representative because it appears to have similar parameters to the population. For example, a sample of academic staff at a university who bothered to reply to a questionnaire has approximately the same proportion of over 30s and a similar gender profile to all the academic staff at the institution. This, of course does not in itself make it representative because, among other things, age and gender may not be relevant variables correlated to the concepts of interest in the study.

Then, of course, there is the related issue of non-response. There is no way of knowing whether the people responding have views similar to those who do not. A survey of academics in a university asked about their views on quality processes, may result in a 30% response rate. Is that because a random sample of 30% responded, or because 70% had no interest in quality processes, or worse 70% were opposed to them but did not feel the managerial climate was conducive to criticising them? Without the non-responses it is impossible to know.

Sometimes the sample is a selective sample of informed people, such as those members of the professoriate closely involved in quality processes. Seeking informed opinions is perfectly legitimate but only if the results are projected as the views of a minority of activist rather than implying it is indicative of the view of the professoriate as a whole.

In thirty years of *Quality in Higher Education*, not one submission involving a sample could be said to be a truly random, unbiased sample. Does that matter? It depends on the claims made subsequently about the results of the study. Strictly speaking it means that no generalisations about the relevant population should be drawn and certainly no claims that anything has been proven should be asserted. Any conclusions should just be indicative suggestions, with explicit caveats about the limitations of the sample, its lack of representativeness and its size.

Talking of size, most samples are of whatever size the researcher can gather within the appropriate time and resources available. If there was a theoretical target number, few studies persist until they reach the target, often undertaking a one-off survey and using whatever responses they manage to glean. In many cases this is not enough to be particularly convincing nor of an adequate size to undertake meaningful statistical analysis. One submission to *Quality in Higher Education* undertook complex and entirely pointless statistical tests of a sample of 10.

Disappointingly, authors of submitted articles rarely even bother to justify the non-randomness of their sample and simply ignore this fundamental element of statistical analysis. They go ahead using statistical procedures on convenience non-random samples of selected respondents, thus rendering the statistical analysis null and void and entirely pointless.

**Scale of the data**

A common illusion practiced in social research is to create numbers from attitude surveys and then immediately forget the nature of the numbers. Attitude surveys often use so-call Lickert scales, which are ordinal scale: for example, very dissatisfied, through neutral to very satisfied; or strongly agree through to strongly disagree. Numbers are allocated to the ordinal scale and in many cases, researchers then treat them as interval and undertake statistical testing that require interval-scale or ratio-scale data. This manoeuvre tends to be legitimised, where it is acknowledged at all, by referencing publications that have argued for this sleight of hand. What they don’t examine is the nature of the sleight of hand. Instead, they assume it can be applied in any circumstance. A corruption compounded by ignoring the other conditions under which the statistical manipulation is legitimate. Let us explore this.

An interval scale has a consistent gap between the values, such as a measurement tape. The difference between 1 metre and 2 metres is the same as between 2 metres and 3 metres (a consistent 1000 millimetres). For a typical Likert scale of ‘very good’, through ‘good’ to ‘neutral’ to ‘bad’ to ‘very bad’, the difference between ‘very good’ to ‘neutral’ (2 units) is not necessarily conceptually the same as from ‘good’ to ‘bad’ (2 units). Once given numerical labels, the conceptual difference is immediately overlooked.

Carifio & Perla (2007) legitimated the use of interval scale statistical analysis on ordinal scale data, such as Lickert scales. They claimed that the F-test (a test of significant difference in variances) was robust and could be used on ordinal scale data such as Likert scales. They referred to Gene Glass’ study of ANOVA in which ‘Glass showed that the F-test was incredibly robust to violations of the interval data assumption (as well as moderate skewing)’. Researchers have used this so-called robustness to undertake ANOVA analysis of ordinal Lickert scales; although by no means do all those taking that position cite an appropriate source. Furthermore, those asserting such robustness then violate the whole process by not taking account of the conditions spelled out by Carifio & Perla. In essence, Carifio & Perla argued that the use of ANOVA on interval scale data works if there are *7 or more scale points* and that the *equality of variances assumption* is not violated. Furthermore,

it is really the correlation coefficient that is most effected by “scale” and “data” type, which is the real, core and key problem that is never mentioned or discussed by various experts on Likert scale, Likert response formats, and statistical analyses thereof…. F is not made of glass but correlation coefficients are to a great degree, and this particular empirical fact and its many consequences are one of the greatest silences in all of this literature. (Carifio & Perla, 2007, p. 110)

In short, using interval statistical techniques on ordinal-scale data is a fraught exercise that, still relies on complying with certain preconditions (number of scale points and equality of variances, for example) and is fragile in correlation analysis. Furthermore, the ordinal scale must be examined to ensure that *conceptually* (rather than numerically) the scale points are equidistant. *Quality in Higher Education* now does not accept articles that use interval data statistical procedures on ordinal scale data, such as the data from most Likert-like scales.

**Data presentation and interpretation**

Besides the problem of data scale, parametric tests tend to have other conditions. The often-used t-test of significance, for example, requires normally distributed populations: undertaking t-tests on clearly skewed population distributions is pointless, even more so if the sample is a non-random, biased convenience sample.

However, it is not just the inappropriateness of the statistical testing; there are further issues about how the results of statistical analysis are reported.

***Significance test interpretation***

Significance tests, although widely used, have been the subject of critique for decades. The concern, apart from their use in in appropriate circumstances, is on how the results are interpreted and reported. Significance tests test a null hypothesis, which says that observed differences (between samples or between a sample and an expected outcome) could be caused by chance because there is so much variation in the sample(s) that the populations from which the samples are taken could be the same. So, when that occurs, the conclusion is that there is no *statistical* difference. That is not the same as no *substantive* difference.

An observed difference between two samples may be statistically non-significant but that does not ‘prove’ the null hypothesis that there is no difference between the populations. Two issues here: first, statistical analysis is about probabilities. Something that is probably the case is not proof; it is only likelihood. Second, because the statistical analysis shows that the resultant probability value is larger than a predetermined threshold value (often p=0.05) does not warrant a conclusion that there is no difference. It simply shows that the observed substantive difference is insufficient, given the data derived from the sample (notably the variance), to assert that there probably is a difference. That is not the same as categorically asserting that there is ‘no difference’ or ‘no association’. The substantive difference may be quite clear and this should not be ignored even if available data suggests no statistically significant difference.

***Correlation and causality***

Confusing correlation with causality is a long-standing issue. Correlation shows that as X changes so does Y. But that is not the same as X causes Y. (Maybe Z ‘causes’ the change in both X and Y?) Blatant disregard for the difference between correlation and cause does not occur often but a shift tends to occur in explications of observed correlation when the results are summed up. Not infrequently, the results are rehearsed in conclusions implying the correlation indicates that X ‘impacts’ on Y, thus reintroducing the causal link by the back door.

There are other issues with presentation of statistics but, in essence, they all come down to an overemphasis on the numbers and an overstatement of the meaning of the numerical analysis.

**Conclusion**

In short, statistics are not what they seem. They are not objective nor truthful and they are only ‘scientific’ if science is defined narrowly by reference to numerical inference. Statistics don’t *tell* you anything; they have to be interpreted, they are framed within hypotheses, which themselves are underpinned by theoretical presuppositions, they are not ‘facts’. (A ‘fact’ is not incontrovertible and self-evident. A ‘fact’ depends on a theoretical context. Change the theory and the ‘fact’ withers away (Chalmers, 1982; Harvey 2012–2024c; Harvey 2012–2024d)).

Statistical analysis can be useful but they are by no means definitive and in many cases are misleading or obstructive because the statistics take over from the conceptual argument.

So don’t be deceived by numbers, don’t be taken in by them. Don’t assume that they actually show what they purport to demonstrate. The statistical analysis may aid understanding but just because a concept is quantified doesn’t mean it is valid. Don’t ever assume that statistics in social science prove anything. Take a close look at the statistical analysis but, more importantly, explore whether the source of the numbers is sound: is the operationalisation sensible or convincing? Is the sample viable or appropriate for statistical analysis? Does the statistical analysis reflect the scale of the data? Are the numbers being analysed conceptually detached from their origins? Have the results been interpreted in a non-misleading manner?

**References**

Blauw. S., 2020, *The Number Bias: How numbers lead and mislead us* (London, Sceptre).

Bukodi, E., *et al*., undated, ‘Social inequalities in educational attainment’, originally at <http://www.oisp.ox.ac.uk/res/education-and-social-policy/social-inequalities-in-educational-attainment.html> (accessed 24 January 2013, no longer online).

Carifio, J. & Perla, R.J., 2007, ‘Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes’, *Journal of Social Sciences*, 3(3), pp. 106–16.

Chalmers, A.F., 1982, What Is This Thing Called Science (Milton Keynes, Open University Press).

Harvey, L., 2003, ‘Scrap that student survey now’, *Times Higher Education*, 12 December 2003. Available at <http://www.timeshighereducation.co.uk/story.asp?storyCode=181752&sectioncode=26>, (accessed 4 January 2024).

Harvey, L., 2012–24a, *Social Research Glossary,* ‘Operationalisation’. Available at <http://www.qualityresearchinternational.com/socialresearch/operationalisation.htm> (accessed 4 January 2024).

Harvey 2012–24b *Social Research Glossary,* ‘Interchangeability of indicators’. Available at http://www.qualityresearchinternational.com/socialresearch/interchangeabilityofindicators.htm (accessed 4 January 2024).

Harvey, L., 2012–24c, *Social Research Glossary*, ‘Theory-laden observation’. Available at <http://www.qualityresearchinternational.com/socialresearch/theoryladen.htm> (accessed 4 January 2024).

Harvey, L., 2012–2024d *Researching the Real World*, part 1, sections 1.4 to 1.7. Available at <http://www.qualityresearchinternational.com/methodology/RRW1pt4Data.php> (accessed 4 January 2024).

Huff, D., 1973, *How to Lie with Statistics*(Harmondsworth, Penguin)

Lazarsfeld, P., Pasanella, A. & Rosenberg, M., (Eds.), 1972, *Continuities in the Language of Social Research* (New York, The Free Press).

Morrison, D.E. & Henkel, R.E., 1970, The Significance Test Controversy: A reader (Chicago, Aldine).