Journal of Philosophy of Education, Vol. 40, No. 4, 2006

The Quantitative-Qualitative Distinction and the Null Hypothesis Significance Testing Procedure

NIMAL RATNESAR AND JIM MACKENZIE

Conventional discussion of research methodology contrast two approaches, the quantitative and the qualitative, presented as collectively exhaustive. But if qualitative is taken as the understanding of lifeworlds, the two approaches between them cover only a tiny fraction of research methodologies; and the quantitative, taken as the routine application to controlled experiments of frequentist statistics by way of the Null Hypothesis Significance Testing Procedure, is seriously flawed. It is contrary to the advice both of Fisher and of Neyman and Pearson, the two sources from which it is drawn.

Discussions of research methodology tend to be dominated by a contrast, and often conflict, between two approaches, the quantitative and the qualitative. The first chapter of Burns (2000) is typical.[1] Research is divided, at least in the social science areas, between two competing methods. On the one hand, there is 'the traditional scientific approach' (p. 3), which 'holds that only a systematic, quantitative approach to generating and testing is adequate' (p. 4). Citing Kerlinger (1986), Burns distinguishes four ways of knowing: tenacity (i.e. holding on to what one already believes), authority, intuition, and science, of which only the last is self-correcting (p. 5). 'The four most important characteristics of science are control, operational definition, replication and hypothesis testing' (p. 5). 'Science is based on the form of logic known as deduction. The basic syllogism is: All Ps are Q/This is a P/Therefore this is Q' (p. 8).[2] The strength of the scientific approach is that it 'produces answers which have a much firmer basis than the lay person's common sense or intuition or opinion' (p. 9).[3]

On the other hand, there is 'the qualitative approach' (p. 10). 'The task of the qualitative methodologist is to capture what people say and do as a product of how they interpret the complexity of their world, to understand events from the viewpoints of the participants. It is the lifeworld of the participants that constitutes the investigative field' (p. 11). 'Qualitative descriptions can play the important part of suggesting possible relationship(s), causes, effects, and even dynamic processes in school settings.

Qualitative methods can highlight subtleties in pupil behaviour and response, illuminate reasons for action and provide in-depth information on teacher interpretations and teaching style' (p. 13). And qualitative reports are much easier to read and understand than reports of quantitative research (p. 14).

At the end of his discussion, Burns admits 'the practice of dichotomising and polarising social science research into quantitative and qualitative modes is overdone and misleading. . . . The contrast that can be supported is the dichotomy between naturalistic research and experimental research' (p. 14).

As a naturalistic description of the behaviour and lifeworld of people *engaged* in educational research, Burns's chapter has considerable validity. But as an account of research methods, it requires substantial correction. The most staggering point is the parochialness and ignorance of the scope of knowledge and of research implicit in the claim. Despite the invocation of Kerlinger's four ways of knowing, Burns's contrast is inapplicable to most of the vast tracts of knowledge outside the social sciences. Mathematical knowledge, for example, is produced neither by naturalistic nor by experimental methods.[4] The natural sciences do indeed use both naturalistic and experimental methods, but field studies are not narratives about the lifeworlds of rocks or pulsars or different kinds of algae, and experimental studies often do not fit Burns's requirements of controlling ('systematically eliminat[ing] the simultaneous influence of many variables', p. 6), operational definition, replication, and hypothesis testing.[5] Even most of those who study aspects of human society—historians, economists, students of linguistics, legal researchers, lexicographers, clinicians, psephologists, philosophers—use methods which bear little resemblance to either the quantitative/experimental or the qualitative/naturalistic models Burns describes. In Burns's case, the focus is clear. Despite his very general title, he is actually writing about research only in education, and only about certain kinds of research in that field; and his emphasis is heavily on quantitative/experimental studies. He devotes 344 pages to quantitative methods but only 164 pages to qualitative methods—ethnographic research, unstructured interviewing, action research, case studies, and historical research are the kinds described, and 50 of these pages are examples of research reports—and 40 pages to survey methods. These proportions are not unusual in textbooks on educational research.

It is, of course, possible to stipulate that all methods other than statistically analysed experiments are to be called 'qualitative', just as it is possible to divide all mammals into bats and non-bats. In both cases, the second class is so heterogeneous (mice, whales, kangaroos, tigers, and baboons differ greatly from each other) as to leave us wondering why the classification was made in that way. In particular, though some methods that are qualitative according to this definition are less than rigorous and their results are very shaky, others—most obviously proofs in Zermelo-Fraenkel set theory, including of course the formal probability theory underlying any use of statistics—are far more rigorous and have

conclusions far more certain than anything in empirical science. The phrase 'qualitative methods' on this suggestion is simply an index of the speaker's narrowness of outlook. The notion that by mentioning both quantitative and qualitative research one is being inclusive and covering all possibilities is a clear indication of ignorance of the scope of research methodology.[6]

Equally striking about the account is its neglect of ideas. If 'Research is a systematic investigation to find answers to a problem' (p. 3), it is surprising that there is no attempt to consider the differences between kinds of problems and the differences between kinds of answers to them (propositions), even at the simplest level. The answer to a clinical problem will commonly be a singular recommendation for what should be done in the existing context; the answer to a theoretical problem may be a context-free generalisation, the answer to another kind of problem may be different again. When Burns gives an example of the Barbara syllogism with singular minor, he remarks that 'the generalisation involved in the first proposition [the major premiss], i.e., all Ps are Qs, is difficult to prove in many cases' (p. 3), but he does not attempt any further taxonomy of propositions into generalisations and non-general propositions, let alone differentiation of each class into subspecies. There is no mention of other forms of traditional syllogism (though Bocardo or Fesapo are more congruent with Popperian methodology.) He describes the statement 'People are poor and starving because God wills it' as metaphysical, and says that such statements cannot be tested (p. 8), but he does not tell us what makes a statement a 'metaphysical' one. Apparently, those that mention God have this characteristic, but that need not be necessary for being metaphysical. There is no discussion of whether any metaphysical propositions however defined are falsifiable, nor of whether any other kinds of propositions are unfalsifiable. Nor is there discussion of whether falsifying propositions may themselves be falsifiable, the immunisation of propositions from falsification by adding conditions, the prospect that propositions may thereby be emptied of empirical consequences, how evaluative or particular or singular or existential or relational propositions fit into this picture with generalisations, and so on.

NULL HYPOTHESIS SIGNIFICANCE TESTING

Burns's fourth requirement for science, hypothesis testing, is the heart of the inferential use of statistics as he and many others understand it. However, statistics and probability are not used as a way of presenting data as relevant evidence for sorting through doubts and uncertainties and for indicating the degree of confidence in inferences warranted by the data. Rather, what is meant is the use of the Null Hypothesis Significance Testing Procedure (henceforth NHSTP). In the 1960s this was presented as the *sine qua non* of scientific research (Gigerenzer, 1993), and still is 'currently the cornerstone of many "quantitative" methods courses' (Gorard, Prandy and Roberts, 2002, p. 36), as indeed Burns's book

confirms. The NHSTP is given pride of place as an indispensable tool of quantitative research and treated with such awe and respect that its output the 'p-value' is regarded as more authoritative than a substantial effect size or even the inter-ocular impact of some graphical description of the data. Obtaining a low p-value below some threshold number is termed 'statistical significance' and is often seen not only in terms of *research* success itself but as the ticket to publication in journals (Gigerenzer, 1993; Sohn, 2000); Salsburg says that 'it provides Salvation: Proper invocation of the religious dogmas of Statistics will result in publication in prestigious journals' (1985, p. 220). Frick, though he concedes at least its insufficiency, still defends it on the grounds that it helps categorise findings 'as being acceptable or not to enter the corpus of claims in psychology' (1996, p. 388).

The main steps of the NHSTP are as follows:

- Specify the null hypothesis, H_0, which is to be opposed; no one would go through with the NHSTP unless they wished to oppose the H_0.
- Calculate a p-value (which these days is invariably done by a computer program).
- If the p-value is smaller than the preset level of significance, á (commonly set at 0.05 though some times 0.01—both for historical reasons rather than from any epistemic significance), then reject H_0. Otherwise, rigorously speaking, one *fails to reject* (rather than *accepts*) H_0 at that level of significance. In the event of a rejection of H_0 then the alternative hypothesis, H_A, the one the researcher wishes to put forward, is claimed as acceptable and publishable.

In addition, it is sometimes recommended that *power* (which is the frequentist probability of correctly rejecting a false H_0—i.e. a long run rate as opposed to an epistemically evaluative comment on any single case), should also be calculated.[7] The preset level of significance, α, is also not a comment on the case at hand but the long run error *rate* of wrongly rejecting H_0—but under identical circumstances.

The p-value is the probability, in the frequentist sense, of obtaining a value of the test-statistic that is at least as extreme as that calculated from the data under H_0. (With some computer programs stars replace the p-value, where the smaller the p-value the more stars, implying that some rejections are of higher rank than others.)

The Null Hypothesis Significance Testing Procedure (NHSTP) is presented, taught and practised as an uncontroversial dichotomous decision rule and also as '*the* monolithic logic of scientific inference' (Gigerenzer, 1993, p. 324, his emphasis); indeed, it is what *inferential* statistics is taken to be all about: a simple mechanical procedure that almost anyone can perform and use to make a research claim. However, it is not uncontroversial, in its origins or in its continued use. Of the NHSTP, Meehl wrote 'I suggest to you that Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the

almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology' (1978, p. 817), and Wang wrote of it that 'The tyranny of the N-P [Neyman-Pearson] theory in many branches of empirical science is detrimental, not advantageous, to the course of science' (1993, p. 21).[8] However, neither Fisher nor Neyman and Pearson is to blame for the way the NHSTP is used, though it draws from the ideas of both. It is rather textbook writers and social research educators who have put together and perpetuated this hybrid procedure (Gigerenzer, 1993; Hubbard and Bayarri, 2003); and still very few textbooks mention that it is problematic (Gliner, Leech and Morgan, 2002).

Not only would neither Fisher nor Neyman and Pearson have agreed with the NHSTP as currently taught and practised but each accused the other party of the same sin, the mechanical and otherwise thoughtless use of statistics (Gigerenzer, 1993). Neyman (1957) noted that Fisher consistently used 0.01 as a conventional level of significance regardless of the research context and problem (p. 12) even though he [Fisher] opposed the final accept-or-reject outcome of the use of the Neyman-Pearson theory (Gigerenzer, 1993, p. 321). Fisher for his part argued that while the N-P theory was possibly useful with repeated sampling for testing long sequences of hypotheses as in industrial quality control contexts, it was *irrelevant* to scientific advance (Hacking, 1965). While Neyman and Pearson certainly saw their [N-P] theory as meaningfully applicable in such industrial contexts and Neyman referred to the ideas of N-P theory, relating it to the wider context of 'the concluding phase of scientific research' (1957, p. 14), as a matter of inductive behaviour involving 'an act of will to behave ... in a particular manner' (p. 12), nevertheless they clearly recognised the need for judgment: as Pearson wrote: '*Of necessity*, as it seemed to *us*, we left in our mathematical model a gap for the exercise of a more intuitive process of personal judgment in such matters—to use our terminology—as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities' (1962, p. 395, emphases added). According to the originators of the N-P theory it is only after the exercise of meaningful judgments that an accept-or-reject decision at some preset level, α, is sensible; and adopting such behaviour over a *long run* of experiments in terms of error *rates* can be used cost-effectively to control quality in some industrial process (such as, for example, ensuring that the diameter of ball bearings on a production line is within acceptable limits). Fisher saw the p-value as an indication or measure of inductive inferential evidence for being able 'to argue from consequences to causes, from observations to hypothesis' (1935, p. 3), but he also said 'no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon' (p. 13).

The NHSTP as commonly practised, however, dispenses with such niceties as judgment and repeated testing: it matters only whether or not a

single experiment produces a statistically significant result. The rigid dichotomous decision to reject or accept (or, more rigorously stated, not to reject) the null hypothesis (which is usually a straw dummy anyway), then follows mechanically, but not by any argument concerning the epistemic state of a research claim. The calculations involved in the NHSTP make no reference whatsoever to any alternative hypotheses nor does the decision, to reject or not, have any epistemic significance of itself. Almost anyone can do it (Gigerenzer, 1993), but the NHSTP procedure rigidly demands that if a significant result is obtained then the null hypothesis, H_0, *must* be rejected—no ifs or buts. There is no room for *almost* significant, though, 'surely, God loves the .06 nearly as much as the .05' (Rosnow and Rosenthal, 1989, p. 1277)—and as for power, 'surely God loves the .79 nearly as much as the .80 recommendation for power' (Ernest and McLean, 1998, p. 60). Neither Fisher nor Neyman and Pearson would have endorsed this rigidity. The N-P theory does not require that a manufacturing process should be rejected outright if a sample is drawn that gives a significant result. One should behave only *as if* the null hypothesis is false and check the machinery, because false alarms are likely in the long run (Gigerenzer, 1993); and in checking the machinery a vast number of collateral premisses are then appealed to in judging whether the machinery is in order or is in need of replacement or repair. Fisher, too, not only said no single experiments could suffice for the demonstration of natural phenomena (1935, p. 13), but also held that it was only after 'fundamental thinking has been accomplished' that exercises 'Constructive imagination, together with much knowledge based on experience' that a 'problem be given a mathematical form' and that it was 'nothing but an illusion to think that this process can ever be reduced to a self-contained mathematical theory of tests of significance' (Fisher, 1939, p. 223). Fisher (1939) highly commended 'Student' (the pen-name of W. S. Gossett) for possessing not only these qualities but also the required pertinacity to perceive and solve problems. Indeed such pertinacity or 'good tenacity' is among what Lakatos (1973, p. 95) held as among the essential qualities of a good researcher. As well, Fisher (1933, p. 46) was quite aware that it was necessary that the statistician be appreciative of the limitations of any available data.

However those who use the NHSTP in social and educational research do not follow Fisher's or Neyman's and Pearson's advice; rather the p-value of a single study is routinely treated as sanctifying or damning some hypothesis, and the results are published without replication and without examining the 'machinery' that produced the result to see if the data are anomalous or limited in any way.

Rigidity that does not admit that it may well be more reasonable to suppose it is the data that are anomalous or limited, but requires instant rejection of the null hypothesis upon obtaining statistical significance, together with the view that science progresses through accumulating publicly verifiable empirical facts and through mechanical procedures and experiments which anybody can repeat at will, are of course the characteristics of a kind of positivism (Polanyi, 1962) which disavows

the intuitions and judgements of the researcher. And this is characteristic of the practice of the NHSTP. Thankfully physicists in the early 1900s were not bothered with the need for 'rigour' of this sort, or Einstein's theory of relativity might never have caught on (Polanyi, 1962; Carver, 1993). As Rozeboom (1960) has noted with reference to the NHSTP, research does not progress through decision-precipitating rules, but through inferential judgements by which we appropriately adjust our beliefs and opinions. It is quite absurd to make an epistemic claim to a general proposition on the basis of a single experiment, as the NHSTP attempts to do.

It is perhaps along these lines that the real philosophical difference underlying the quantitative-qualitative divide is drawn: positivism as contrasted with rationalistic realism, bats as opposed to non-bats, NHSTPism versus (real) research. Research worthy of the name requires an intellectual commitment (Polanyi, 1962), judgement (Fisher, 1939; Pearson, 1962; Rozeboom, 1960), the necessary kind of pertinacity or good tenacity (Fisher, 1939; Lakatos, 1973), and a constructive imagination (Fisher, 1939). The house of knowledge has many mansions. Even poetry may use numbers and John von Neumann put his faith in 'the influence of men with an exceptionally well-developed taste' to avoid 'degeneration' in mathematics (1947, p. 196).

As for a *probabilistic modus tollens*, claimed as the justifying logic of the NHSTP, that just doesn't work either, (Cohen, 1994; Royall, 1997; Sober, 2002); rare events do happen (Hacking, 1965); a frequentist low probability, a mathematical construct, does not automatically convert to the inferential notion of implausibility; and randomisation is not a cure either because the nice mathematical properties of a random variable do not extend to the given data sample in hand. To maintain the illusion that research, even just within education, is to be thought of as divided between quantitative and qualitative is no longer sustainable.[9]

Correspondence: Jim Mackenzie, School of Policy and Practice, Faculty of Education and Social Work, University of Sydney, New South Wales, Australia.
Email: j.mackenzie@edfac.usyd.edu.au

NOTES

1. There are many textbooks providing introductions to research methods for students of education and the social sciences. Burns (2000) is typical of the genre and is widely used. We have focused our remarks on it for the sake of definiteness.
2. This form is recognisable as the argument schema known in the traditional formal logic of the Middle Ages as Barbara with singular minor. An argument is deductively valid just in case it is logically impossible for all the premisses to be true and the conclusion false. Traditional syllogistic can represent some, but not all, valid deductive arguments.
3. Since Burns will later cite Popper with approval for endorsement of the scientific approach, it may be worth noting Popper's own account of this 'firm basis', which differs starkly from Burns's: 'The empirical basis of objective science has thus nothing "absolute" about it. Science does not rest upon solid bedrock. The bold structure of its theories rises, as it were, above a swamp. It is like a building erected on piles. The piles are driven down from above into the swamp, but not

down to any natural or "given" base; and if we stop driving the piles deeper, it is not because we have reached firm ground. We simply stop when we are satisfied that the piles are firm enough to carry the structure, at least for the time being' (Popper, 1959, §30, p. 111). Burns does not cite this work, though it is of course the central text for Popper's views on scientific method.

4. Mathematical research is certainly not quantitative in the sense defined.
5. And even when they seem to, this may be a misleading appearance of the conventions for writing up research insisted upon by editors of scientific journals. '[T]he scientific paper is a fraud in the sense that it does give a totally misleading narrative of the processes of thought that go into the making of scientific discoveries' (Medawar, 1963, p. 38).
6. One recalls the boast by a rural radio station that it plays both kinds of music, Country *and* Western.
7. Burns defines: 'power is the ability of a statistic to correctly reject the null hypothesis when it is false' (2000, p. 160).
8. The N-P [Neyman-Pearson] theory refers to the theory developed by Jerzy Neyman and Egon Pearson. The N-P theory departs from Fisher's idea of hypothesis testing (of considering the probabilistic behaviour of just one hypothesis), in that it also recognises the need to include the alternative hypothesis *together* with the possibility of the errors of wrong decisions about which of the hypotheses could be true. Under the N-P theory the error of rejecting a true null hypothesis is called an error of the first kind (type I error), while the error of accepting or failing to reject a false null hypothesis is called an error of the second kind (type II error). The N-P theory uses the criterion of what it calls the critical region. If the value of the test statistic, as calculated from the sample, falls into that region then the null hypothesis, according to which such an event is deemed improbable or infrequent, is regarded as rejectable—albeit with the possibility of a type I error. In addition the N-P theory also discusses the idea of the *power* of a test, which is the probability or (limiting relative) frequency with which a test would correctly reject a false null hypothesis.
9. Work for this paper was hindered by the inadequate funding of Australian academic libraries.

REFERENCES

Burns, R. B. (2000) *Introduction to Research Methods*, 4th edn (Sydney, Longmans Pearson Education Australia).

Carver, R. (1993) The Case Against Statistical Significance Testing, Revisited, *Journal of Experimental Education*, 61, pp. 287–292.

Cohen, J. (1994) The Earth is Round ($p < .05$), *American Psychologist*, 49, pp. 997–1003.

Ernest, J. M. and McLean, J. E. (1998) Fight the Good Fight: A Response to Thompson, Knapp, and Levin, *Research in the Schools*, 5.2, pp. 59–62.

Fisher, R. A. (1933) The Contributions of Rothamsted to the Development of the Science of Statistics. Annual Report of the Rothamsted Station, pp. 43–50. (Reprinted in his *Collected Papers*, ed. J. H. Bennett, vol. 3 (Adelaide, University of Adelaide Press), pp. 84–91).

Fisher, R. A. (1935) *The Design of Experiments* (Repr. Edinburgh; Oliver & Boyd, 8th edn, 1966).

Fisher, R. A. (1939) 'Student', *Annals of Eugenics*, 9, pp. 1–9 at ⟨http)//www.library.adelaide.edu.au/digitised/fisher/165.pdf⟩ (Reproduced with permission of Cambridge University Press).

Frick, R. W. (1996) The Appropriate Use of Null Hypothesis Testing, *Psychological Methods*, 1.4, pp. 379–390.

Gigerenzer, G. (1993) The Superego, the Ego, and the Id in Statistical Reasoning, in: G. Keren and C. Lewis (eds) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (Hillsdale, NJ, Lawrence Erlbaum Associates), pp. 311–339.

Gliner, J. A., Leech, N. L. and Morgan, G. A. (2002) Problems with Null Hypothesis Significance Testing (NHST): What do the Textbooks Say?, *Journal of Experimental Education*, 71.1, pp. 83–92.

Gorard, S., Prandy, K. and Roberts, K. (2002) An Introduction to the Simple Role of Numbers in Social Science Research. ESRC (Economic and Social Research Council) Teaching and Learning Research Programme, Research Capacity Building Network, Occasional Paper Series, Paper 53 at ⟨http://www.cf.ac.uk/socsi/capacity/Papers/roleofnumbers.pdf⟩.

Hacking, I. (1965) *Logic of Statistical Inference* (Cambridge, Cambridge University Press).

Hubbard, R. and Bayarri, M. J. (2003) Confusion Over Measures of Evidence (p's) versus Errors (α's) in Classical Statistical Testing, *The American Statistician*, 57.3, pp. 171–182.
Kerlinger, F. (1986) *Foundations of Behavioral Research* (New York, Holt).
Lakatos, I. (1973) Lectures on Scientific Method, in: M. Motterlini (ed.) *For and Against Method* (Chicago, University of Chicago Press).
Medawar, P. (1963) Is the Scientific Paper a Fraud? Repr. in his *The Strange Case of the Spotted Mice and Other Classic Essays on Science* (Oxford, Oxford University Press), 1996, pp. 33–39.
Meehl, P. E. (1978) Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology, *Journal of Consulting and Clinical Psychology*, 46.4, pp. 806–834.
Neumann, J. von (1947) The Mathematician, in: R. B. Heywood (ed.) *The Works of the Mind* (Chicago, University of Chicago Press).
Neyman, J. (1957) Inductive Behavior as a Basic Concept of Philosophy of Science, *International Statistical Review*, 25, pp. 7–22.
Pearson, E. S. (1962) Some Thoughts on Statistical Inference, *Annals of Mathematical Statistics*, 33, pp. 394–403.
Polanyi, M. (1962) *Personal Knowledge: Towards a Post-Critical Philosophy* (London, Routledge & Kegan Paul).
Popper, K. R. (1959) *The Logic of Scientific Discovery* (*Logik der Forschung,* 1934) (London, Hutchinson).
Rosnow, R. L. and Rosenthal, R. (1989) Statistical Procedures and the Justification of Knowledge in Psychological Science, *American Psychologist*, 44, pp. 1276–1284.
Royall, R. (1997) *Statistical Evidence—a Likelihood Paradigm* (Boca Raton, FL, Chapman and Hall).
Rozeboom, W. W. (1960) The Fallacy of the Null-Hypothesis Significance Test, *Psychological Bulletin*, 57, pp. 416–428.
Salsburg, D. S. (1985) The Religion of Statistics as Practiced in Medical Journals, *American Statistician*, 39, pp. 220–223.
Sober, E. (2002) Intelligent Design and Probability Reasoning, *International Journal of Philosophy of Religion*, 52, pp. 65–80.
Sohn, D. (2000) Significance Testing and the Science [Comment], *American Psychologist*, 55.8, pp. 964–965.
Wang, C. (1993) *Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety* (New York, Marcel Dekker).