The future of quantitative educational research methods: bigger, better and, perhaps, Bayesian?

Matt Homer
School of Education
University of Leeds
Leeds LS2 9JT
Email for correspondence: m.s.homer@leeds.ac.uk

Abstract

In this paper, I discuss some of the key ongoing developments taking place in quantitative educational research methods, and consider the likely future changes over the next decade or so. There are a range of issues and developments that together are shaping what is do-able and what is seen as acceptable in terms of robust quantitative approaches to research. Amongst the most important of these are a greater awareness of methodological problems with previous approaches, ongoing methodological innovations, better computer power, greater availability of secondary and 'big' data, and a stronger emphasis on demonstrating causal effects rather than just correlations in research. Some of these changes are promoted by current funding polices, but others are, perhaps, more a result of the unique time we are living; where technological change, the relative ease of generating data (e.g. via online surveys), and the existence of a range of large scale secondary data sets together mean that the importance of applying appropriate quantitative research methods is more and more recognised and valued. However, whilst generally these developments are positive, much of what goes in is still to an extent contested, and I will consider some of the important current arguments, and touch on the likely challenges that remain. I start with the fundamental assumptions of inferential statistics.

What use is a p-value?

A common approach to teaching about hypothesis testing is (or, hopefully, used to be) to focus on the importance of the 5% level. If we are, say, comparing two groups on an outcome, we do a test (e.g. the independent sample t-test). If the result shows $p<0.05$ we have a 'significant' result and conclude that the two groups differ on the outcome. However, it is well known that this process is flawed in a number of crucial ways.

1. The logic is the wrong way around – traditional hypothesis testing tells you about the likelihood of obtaining the observed data *given the null hypothesis*, but actually the researcher is more interested in the opposite; the probability of the null hypothesis being true *given the data*. There are ways around this logical problem (see comments on Bayesian methods later) but these are not straightforward (Gorard, 2010; Neale, 2015; Ziliak & McCloskey, 2008).

2. The 5% level is entirely arbitrary. Formally, this is the level that we decide we are happy with making a Type I error (i.e. the mistake of finding a 'significant' result even though there was really no difference in our data – also known as a 'false positive'). We should not reify this 1 in 20 value, but rather think of p-values on a continuum (Cumming, 2011). Then decisions become more nuanced rather than focussing on which side of an artificial boundary your result lies.

3. If you have a big enough sample, everything is 'significant'. For example, you might get a 'small' correlation of, say, $r=0.1$, but because the sample is 1000 this is 'statistically significant'. A better interpretation would be that only 1% of the variation in one of the variables is accounted for by the other (since r-squared=0.01) – in other words, not a lot. This example shows that the p-value is not a good measure of practical importance. Instead we should use appropriate effect sizes or confidence intervals (Coe, 2002; Cumming, 2011). Effect sizes in particular are not sample size dependent and so give a standardised measure of the 'effect' under study.

4. Is the sample really random? We often forget that the whole of inferential statistics is based on the idea of a random sample from a population – but in education, certainly, we rarely have a truly random sample (Gorard, 2003). When you send out a questionnaire, who responds? Do they have the same views as those that don't respond? Probably not, and yet we can't force people to respond to our research instruments and so the issue of bias is always with us.

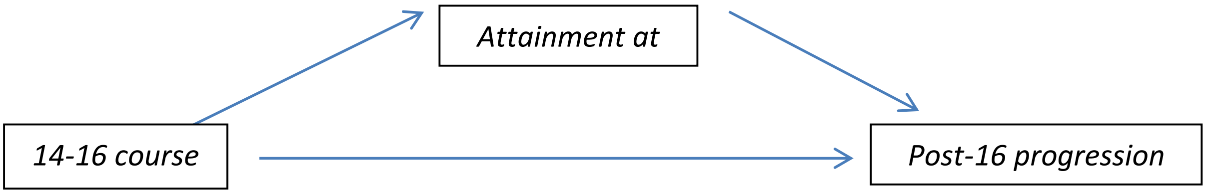
Given these problems, there are some who say that p-values are therefore useless (Gorard, 2010, but see also the response by Neale, 2015), and even that they should be 'banned' as has recently been the case in the journal Basic and Applied Psychology (Trafimow & Marks, 2015). My view is that p-values certainly have some use – they are, however, unfortunately, often misused and can and do distort research practise (Ioannidis, 2005; Young & Karr, 2011). So it's OK to report a p-value, but useful inferences about what the research is telling you rest on so much more than this. My feeling is that educational research is generally quite advanced compared to some other disciplines in this regard, and hopefully in the future good practice in this area will continue to spread.

How do we best demonstrate causality?

Demonstrating causality is becoming more and more important for many (quantitative) researchers, and, increasingly, for funders (e.g. the Education Endowment Foundation[1] and the Department for Education), despite the philosophical and practical objections that can be made (Morrison, 2012). The randomized control trial – borrowed from clinical science - is often frustratingly quoted as the 'gold standard' here (Haynes et al., 2012), but there have been and remain many criticisms of this approach in education (Thomas, 2013). Whilst there is clearly a place for such a research design, when, for example, studying a focussed intervention whose key outcomes can best be measured in a what is essentially a test, there are many researchers who would argue that its supposed supremacy is (at best) an illusion. Anecdotally, I also feel that the language of 'experimentation' understandably alienates some researchers – 'trial', 'control', 'intervention', 'randomization', and so on, so those in the RCT camp could take greater care to 'sell' their approach. More importantly, the privileging of this type of research (including by funders) is highly problematic for many researchers, many of whom are otherwise highly competent in quantitative approaches (Norman, 2003). Of course, the arguments for and against RCTs in medical research interventions are different from those in the social sciences, and the whole issue is, perhaps, less contested.

In some scientific disciplines, for example in Biostatistics, there is a growing strand of work attempting to use observational data to try and estimate causal effects. This is in part because there is already a lot of such secondary data to hand, and also because of some of the problems with carrying out RCTs in health and related settings (for example, cost, practical and ethical problems, and doubts about the generalizability of findings). These newer methods include propensity score matching (Bai, 2011), mediation/moderation models (Field, 2013, Chapter 10), and more and more emphasis on latent variables that include measurement error and structural models (Kline, 2011) . For example, suppose we are interested in a new course at KS4 (14-16), and how it impacts on attainment at KS4 and progression to post-16. The classic approach would probably involve modelling influences on progression at an outcome in terms of *14-16 Course* and *Attainment at 16* (perhaps using logistic regression). However, a better approach would be a causal model as shown in the diagram:

[1] https://educationendowmentfoundation.org.uk/apply-for-funding/



The key difference here is that a successful course might impact on progression directly (e.g. through enhanced student engagement), but also indirectly through higher attainment at 16. In 'standard' techniques making such distinctions is not possible. Note also that using a causal diagram forces you as a researcher to make your thoughts (i.e. your theory) explicit with regard to the causal nature of what is going on. This is really important and useful.

Hopefully, these types of approach will gain ground in educational research as confidence and knowledge of such approaches spreads, although I wouldn't underestimate the difficulty of capacity building in these areas, and in correctly implementing and interpreting the findings for some of these techniques. Perhaps the Q-Step initiative[2], to upskill undergraduate quantitative skills across the UK, will have sufficient medium to long term impact to make this less of a concern as some of these graduates develop into post-graduate quantitative researchers.

I would, however, also like to add some caution about the efficacy of the use of observational techniques in 'determining' causality. You can have a causal theory and successfully test it on the data you have, but if you as a researcher haven't sufficiently 'controlled' the data but merely 'observed' it, then you simply don't know what other 'lurking' variables might be present that are being ignored in your (therefore incorrect) causal model (Murnane & Willett, 2010, Chapter 3). This problem isn't going away.

One final point on study design and 'robust' research - we probably need more replication studies in education – and to move away from the idea that a single study 'proves' anything at all (Makel & Plucker, 2014). Given the issues there have been historically with studies not being replicated , this might seem sensible or even obvious, but it will require a bit of a philosophical shift on the part of funders and journals, away from the 'what is new in this?' model of judging research proposals, and towards a more nuanced view of real world research and the necessity of trying things out more than once, particularly for 'important' findings, perhaps via different research teams (Warne, 2014). I re-visit some of these issues again when discussing meta-analysis later in this paper.

[2] http://www.nuffieldfoundation.org/q-step

Where is the data going to come from?

I have already alluded to the growth in the use of secondary data sets in educational research (Smith, 2008; Gorard, 2012). We have the national pupil database in England[3] containing all student qualification/attainment data from national tests, GCSEs, A-levels and other qualifications from ages 4 to 19, as well as student and school characteristics. This is freely available to researchers and provides a rich resource for tracking students and assessing influences on participation and attainment. There is also higher education data available from HESA[4] which can be linked to the NPD, although there are sometimes fees for its use. As an interesting aside, there is a nice philosophical argument to be had here about whether it is acceptable to use inferential statistics on such datasets given that you have access to the full population rather than a sample from which you are generalising (Gibbs et al., 2015; Gelman, 2009). I would argue that Gibbs and colleagues, who don't like treating populations as samples in census data, probably suffer from a lack of imagination here – just because we have the whole population doesn't mean we are just interested in them in themselves. There is always a hypothetical super-population available based on the view that the current census data is just one realisation of what could have been observed.

The NPD is one of a number of large national 'administrative' data sets containing millions of individual records, but there are also large international comparative datasets – for example, the Programme for International Student Assessment (PISA)[5] and Trends in International Mathematics and Science Study (TIMMS)[6] covering core subject areas (e.g. reading, science and mathematics – but also contain affective measures). Data from these international surveys is again available to researchers, and there is a growing literature consisting of analysis within and between particular countries, often focussing on particular subject areas (e.g. science education). I am not going to get into to the argument here about PISA and what many might see as the problematic and pernicious effects that the international 'league tables' might have on educational policies within countries (Kreiner & Christensen, 2013; Schuelka, 2013). The point is that the use of all of these secondary sources, including those curated by the UK Data Service[7], is only going to grow over time in part because the ESRC and other funders recognise the efficiency of using pre-existing datasets for new research purposes. Further, there is also the growing linkage of data (Harron et al., 2015). For example, I know of that tentative steps are being made to link health data to NPD data to investigate the impact of (say) early acute trauma on later educational outcomes. There are, of course, ethical and other concerns with such moves but if these can be adequately addressed then data linkage is going to open up big and important new areas of educationally-related research.

---

[3] https://www.gov.uk/guidance/national-pupil-database-apply-for-a-data-extract

[4] https://hesa.ac.uk/bespoke-data-service

[5] http://www.oecd.org/pisa/

[6] http://timssandpirls.bc.edu/

[7] https://www.ukdataservice.ac.uk/

Another potential and ever growing area for generating quantitative data is the more or less ubiquitous online survey. The freely available tools for implementing such surveys (e.g. Bristol online surveys – available through the University of Leeds[8]) are now sufficiently user-friendly and convenient to be accessible to all researchers. Provided you can gain access to your target sample (e.g. via email), these tools provide an incredibly easy way to 'do' quantitative research on some groups, including even quite young children (Lloyd & Devine, 2010). Issues of representativeness and bias can obviously be a problem, as can lack of respondent engagement – where low response rates are quite common. However, with careful design – sufficient piloting and keeping survey length to a minimum – the use of these tools is bound to grow in the future.

What are the innovative methodological approaches likely to be?

Some 'modern' techniques have become quite familiar in educational research in the last 10 years or so. For example, multi-level modelling (Goldstein, 1995) is often employed on clustered data (e.g. pupils in classes in schools), in part to correctly account for the lack of independence in the observations. In fact, some argue that these approaches are over-used and there are simpler approaches that will often suffice (Gorard, 2007; Hutchison & Schagen, 2008; Huang, 2016). Nevertheless, in educational contexts it is important to account for the structure in the data and this will become accepted practice where it hasn't already. I now discuss some other 'modern' quantitative approaches whose uses are likely to grow.

Meta-analysis

Meta-analysis is where a 'typical' intervention effect size across a range of studies and contexts is estimated – as per John Hattie's influential work on 'what works' in education (Hattie, 2008). This I find highly problematic. There are technical objections to meta-analysis and to Hattie's work in particular (Higgins & Simpson, 2011; Terhart, 2011), but in practical terms one has to ask what use to a policy maker is an average across widely varying contexts – for example, what does the effect of class size on educational outcomes in (say) Singapore tell us about its likely impact in Leeds? The fact that effect sizes for particular interventions vary across contexts surely indicates that the 'mechanism' is mediated by a wide range of other influences that are highly specific to the particular location. Regardless of the methodological objections to meta-analysis, the growth of experimental studies in education across the world, and certainly in the UK, is likely to lead to the demand for ways of summarising across them. Where contexts are reasonably similar, this probably makes sense.

[8] https://it.leeds.ac.uk/info/173/database_and_subscription_services/206/bristol_online_survey_accounts

Missing data

Less controversial methodological areas that are growing in importance include better ways to handle missing data (Pampaka et al., 2014) compared to (say) just doing a complete case analysis – which usually produces biased results . More user-friendly software plays a part here. For example, SPSS has a range of options to impute missing data, depending on exactly what assumptions about the nature of the missingness are reasonable. However, care needs to be taken in sufficiently justifying any assumptions that are required when using these imputation methods.

Confirmatory and latent variable techniques

There will be a move away from descriptive and simpler methods towards more confirmatory methods that model and 'test' the data better. Examples of such a shift include employing multi-level approaches (or similar) in preference to ordinary least square regression for clustered data, confirmatory factor analysis in addition to (or instead of) exploratory factor analysis (Kline, 2011), latent class analysis as opposed to cluster analysis as a way to group cases (Hagenaars, 2009), and Rasch modelling over more 'standard' item response theories in psychometric measurement (e.g. assessment and affective scales) (Bond & Fox, 2007; Panayides et al., 2009).

Re-sampling methods

There will also be growing use of re-sampling and other computational approaches (Bai & Pan, 2008) – where the sample itself is re-sampled, for example to use bootstrapping (Wood, 2004; Boos & Stefanski, 2010) to estimate standard errors (i.e. uncertainty) for effects where the usual distributional assumptions (e.g. normality) might not apply. The use of simulation studies , which are relatively common in some other scientific fields, might be applied more in education – perhaps to investigate, for example, whether patterns of attainment observed in real data match that generated some hypothesised underlying causal mechanism.

Bayesian approaches – updating estimates based on new data

Finally, there might be a shift toward Bayesian methods (Kaplan, 2014), where the whole philosophy of statistical inference is turned on its head. Suppose we want to know the 'size' of the 'gap' in educational outcomes between students from high and low socio-economic backgrounds. The standard approach (sometimes called 'frequentist') is to use research data to estimate a 'fixed' value for this difference, possibly also including some measure of the uncertainty in this value (i.e. an associated confidence interval). The Bayesian approach views the world differently, conceptualizing the 'gap' itself as following a distribution (i.e. a pattern) – which we don't know and can't ever know completely. When new data becomes available it serves merely to update our understanding based on our previous view of this distribution. So in a Bayesian world, the current state of the knowledge depends on the new data in combination with the previous state of knowledge.

However, this is a challenging area, not least because the available software and the mathematical/philosophical underpinnings can prove a bit of a barrier, but in the end it might well become more widely employed in statistical modelling in educational research. Bayesian approaches have many advantages, for example the equivalent of a 95 confidence interval under Bayes (a 'credible interval') means exactly what it says (e.g. '*the true value is in this range with 95% certainty'*) - compare with the convoluted definitions of confidence intervals under 'frequentist' (i.e. traditional) statistical approaches (e.g. '*Were this procedure to be repeated on multiple samples, the calculated confidence interval would include the true value 95% of the time'*).

One disadvantage of a Bayesian approach, and one that can make those of us used to traditional statistical approaches uncomfortable, is that if you have no 'previous' data to go on then you have to make a 'subjective' choice for this (this is the 'prior' in Bayesian terminology). However, Bayes does do inference the 'right way around', so a p-value in a Bayesian analysis does tell you the likelihood of the null hypothesis being correct – again compare with logic 'the wrong way around' under traditional approaches (as mentioned earlier).

Conclusion

This piece has argued that over the next decade there will be more educational data available, that it will often be 'bigger', and that the use of better, more sophisticated quantitative methods in the future will continue to grow to make the best of these data sources. A key area where this is already happening but will develop further is in the estimation of causal effects from non-experimental research.

However, one should be careful not to argue for more complex methods *per se* where simpler methods will do. Any developments in the application of quantitative research approaches in the future should rest, as they always should have, upon the needs of the research questions to hand. If simple approaches will do, then fine. The researcher is at the heart of all this, and the underlying theoretical and philosophical assumptions 'in play' when carrying out the research need as much consideration (if not more) than the (statistical) methods employed.

References

Bai, H. 2011. A comparison of propensity score matching methods for reducing selection bias. *International Journal of Research & Method in Education*. **34**(1),pp.81–107.

Bai, H. and Pan, W. 2008. Resampling methods revisited: advancing the understanding and applications in educational research. *International Journal of Research & Method in Education*. **31**(1),pp.45–62.

Bond, T.G. and Fox, C.M. 2007. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* 2nd ed. Psychology Press.

Boos, D. and Stefanski, L. 2010. Efron's bootstrap. *Significance*. **7**(4),pp.186–188.

Coe, R. 2002. It's the effect size, stupid: what effect size is and why it is important. [Accessed 7 March 2014]. Available from: http://www.leeds.ac.uk/educol/documents/00002182.htm.

Cumming, G. 2011. *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Field, A.P. 2013. *Discovering statistics using IBM SPSS statistics: and sex and drugs and rock 'n' roll*. London: Sage Publications.

Gelman, A. 2009. How does statistical analysis differ when analyzing the entire population rather than a sample? *Statistical Modeling, Causal Inference, and Social Science*. [Online]. [Accessed 16 November 2015]. Available from: http://andrewgelman.com/2009/07/03/how_does_statis/.

Gibbs, B.G., Shafer, K. and Miles, A. 2015. Inferential statistics and the use of administrative data in US educational research. *International Journal of Research & Method in Education*. **0**(0), pp.1–7.

Goldstein, H. 1995. *Multilevel statistical models* 3rd ed. London: Arnold.

Gorard, S. 2010. All evidence is equal: the flaw in statistical reasoning. *Oxford Review of Education*. **36**(1),pp.63–77.

Gorard, S. 2007. The dubious benefits of multi-level modeling. *International Journal of Research & Method in Education*. **30**(2),p.221.

Gorard, S. 2012. The Increasing Availability of Official Datasets: Methods, Limitations and Opportunities for Studies of Education. *British Journal of Educational Studies*. **60**(1),pp.77–92.

Gorard, S. 2003. Understanding Probabilities and Re-Considering Traditional Research Training. [Accessed 16 November 2015]. Available from: http://www.socresonline.org.uk/8/1/gorard.html.

Hagenaars, J.A. 2009. *Applied Latent Class Analysis* 1 edition. Cambridge; New York: Cambridge University Press.

Harron, K., Goldstein, H. and Dibben, C. 2015. *Methodological Developments in Data Linkage* 1 edition. Chichester, West Sussex, United Kingdom: Wiley-Blackwell.

Hattie, J. 2008. *Visible learning: a synthesis of meta-analyses relating to achievement*. London : New York: Routledge.

Haynes, L., Service, O., Goldacre, B. and Torgerson, D. 2012. *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials* [Online]. London: Cabinet Office. [Accessed 7 May 2014]. Available from: https://www.gov.uk/government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials.

Higgins, S. and Simpson, A. 2011. Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement. By John A.C. Hattie. *British Journal of Educational Studies*. **59**(2),pp.197–201.

Huang, F.L. 2016. Alternatives to Multilevel Modeling for the Analysis of Clustered Data. *The Journal of Experimental Education*. **84**(1),pp.175–196.

Hutchison, D. and Schagen, I. 2008. Concorde and discord: the art of multilevel modelling. *International Journal of Research & Method in Education*. **31**(1),p.11.

Ioannidis, J.P.A. 2005. Why Most Published Research Findings Are False. *PLoS Med*. **2**(8),p.e124.

Kaplan, D. 2014. *Bayesian Statistics for the Social Sciences* 1 edition. New York: Guilford Press.

Kline, R.B. 2011. *Principles and practice of structural equation modeling*. New York: Guilford Press.

Kreiner, S. and Christensen, K.B. 2013. Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*. **79**(2),pp.210–231.

Lloyd, K. and Devine, P. 2010. Using the Internet to Give Children a Voice: An Online Survey of 10- and 11-Year-Old Children in Northern Ireland. *Field Methods*. **22**(3),pp.270–289.

Makel, M.C. and Plucker, J.A. 2014. Facts Are More Important Than Novelty Replication in the Education Sciences. *Educational Researcher*. **43**(6),pp.304–316.

Morrison, K. 2012. Searching for causality in the wrong places. *International Journal of Social Research Methodology*. **15**(1),pp.15–30.

Murnane, R. and Willett, J. 2010. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford ; New York: OUP USA.

Neale, D. 2015. Defending the logic of significance testing: a response to Gorard. *Oxford Review of Education*. **41**(3),pp.334–345.

Norman, G. 2003. RCT = results confounded and trivial: the perils of grand educational experiments. *Medical Education*. **37**(7),pp.582–584.

Pampaka, M., Hutcheson, G. and Williams, J. 2014. Handling missing data: analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education*. **0**(0),pp.1–19.

Panayides, P., Robinson, C. and Tymms, P. 2009. The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal*. [Online]. [Accessed 18 November 2009]. Available from: http://www.informaworld.com/10.1080/01411920903018182.

Schuelka, M.J. 2013. Excluding students with disabilities from the culture of achievement: the case of the TIMSS, PIRLS, and PISA. *Journal of Education Policy*. **28**(2),pp.216–230.

Smith, E. 2008. *Using Secondary Data in Educational and Social Research* 1 edition. Maidenhead; New York, NY: Open University Press.

Terhart, E. 2011. Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies*. **43**(3),pp.425–438.

Thomas, G. 2013. No one can control for a sense of when 4-3-3 might turn the game. *Times Higher Education*. [Online]. [Accessed 16 November 2015]. Available from: https://www.timeshighereducation.com/comment/opinion/no-one-can-control-for-a-sense-of-when-4-3-3-might-turn-the-game/2001371.article.

Trafimow, D. and Marks, M. 2015. Editorial. *Basic and Applied Social Psychology*. **37**(1),pp.1–2.

Warne, R.T. 2014. Two Additional Suggested Reforms to Encourage Replication Studies in Educational Research. *Educational Researcher*. **43**(9),pp.465–465.

Wood, M. 2004. Statistical inference using bootstrap confidence intervals. *Significance*. **1**(4),pp.180–182.

Young, S.S. and Karr, A. 2011. Deming, data and observational studies. *Significance*. **8**(3),pp.116–120.

Ziliak, S.T. and McCloskey, D. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: The University of Michigan Press.