The limitations of quantitative social science for informing public policy

John Jerrim[1]

Robert de Vries[2]

1 UCL Institute of Education

[2]University of Kent

July 2015

Abstract

Quantitative social science (QSS) has the potential to make an important contribution to public policy. However it also has a number of limitations, many of which are unknown or poorly understood by those not familiar with quantitative methodology. The aim of this paper is to explain these limitations to a non-specialist audience and to identify a number of ways in which QSS research could be improved to better inform public policy.

Key words: Independent verification; replication; publication bias; statistical uncertainty; public policy.

Contact Details: John Jerrim (J.Jerrim@ioe.ac.uk), Department of Quantitative Social Science, Institute of Education, University of London, 20 Bedford Way London, WC1H 0AL

Acknowledgements: I would like to thank Gemma Moss, David Budge and John Micklewright for helpful comments on a draft presentation of this work.

Policymakers make extensive use of quantitative evidence to inform and justify policy decisions. However, there are a number of issues in how this evidence is produced and used by such groups. In this paper we focus specifically on the 'supply side' of how quantitative social science (QSS) evidence is created, and the most pressing challenges associated with this process.

Many of the issues we consider have been extensively discussed within various disciplines (e.g. Ioannidis 2006; Dirnagel and Lauritzen 2010; Franco et al 2014). However, by bringing these together, we hope to illustrate how they combine to potentially harm the policy making process. In doing so, we hope to inspire a change in QSS research practice, and to encourage policymakers (and the public) to engage more thoroughly and critically with QSS evidence.

Introduction

QSS has a major role in the policy-making process worldwide (see Johnson and Antill [2011] and Parsons et al [2014] for recent examples from the UK). High impact studies receive significant media coverage, with academic evidence used to identify important social issues and inform appropriate policy responses (see examples below). There are many reasons why policymakers might be particularly drawn to quantitative evidence. First, numerical findings and statistics may seem more certain and 'scientific' than qualitative observations and interviews. Indeed, Allen and Preiss (1997)[1] note that people find messages supported by quantitative evidence more persuasive than narrative arguments alone, while journalists judge quantitative studies to be more accurate and newsworthy than qualitative research (Schmierbach 2005). A further attraction is that QSS often simplifies complex social problems into a single set of numbers. Some suggest this may explain the prevalence of international performance indicators (Kelley and Simmons 2015), such as the widely cited

[1] However, as Brownson et al (2009) note, the combination of qualitative and quantitative evidence has a more persuasive impact than either type of evidence alone.

Programme for International Student Assessment (PISA) rankings of children's educational attainment, which are widely cited by policymakers across the globe.

To the authors' knowledge, no research has investigated the effect of these attitudes on policymakers' use of QSS. However, even a cursory examination of UK parliamentary debates yields several examples of policymakers treating QSS evidence as an independent body of neat, certain facts. The following quote illustrates this practice clearly:

> '*Evidence shows that cohabiting parents are four times more likely to have separated by the time their child is three years of age, and by their child's fifth birthday, more than one in four of those who cohabit have split up. For married parents, however, the break-up rate is fewer than one in 10…It is not any form of prejudice; it is the evidence behind the Government's wish to recognise marriage in the tax system*'. (House of Commons Debate, 21 October 2014, c194WH)

Yet this does not reflect the reality of QSS research. The social world is incredibly complex and dynamic, and evidence, even on quite basic social questions, is rarely black and white. Instead it is often provisional, qualified, and uncertain.

In the following sections, we discuss four limitations of the QSS evidence base, and suggest ways these could be addressed to make academic QSS a more effective resource for policymakers. We also briefly discuss difficulties at the QSS evidence-policymaking interface via a specific case study. Many of the topics covered are likely to have relevance for other disciplines (e.g. epidemiology, genetics) and alternative approaches (e.g. qualitative social science). However, as our expertise is in QSS, this is our point of focus.

Transparency and verifiability

Policymakers often deploy QSS evidence by citing figures from academic papers. As already noted, these figures are commonly treated as exact, concrete, and final. Yet such estimates are often the result of a long and complex process, with many assumptions and caveats embedded in the analysis.

To begin, the data must be downloaded and the various files merged together. This dataset must then be 'cleaned', before key variables are recoded into the desired format. Any difficulties with the data should be documented and investigated, including missing information and possible errors in measurement. Descriptive statistics should then be produced before any 'modelling' takes place (however, often it is only this last step that is reported in detail in QSS papers).

It is vital that the steps outlined above are carried out correctly; mistakes (e.g. miscalculating new variables) are easy to make, and can lead to serious difficulties in the modelling that follows. Leaving aside true errors, there are also many judgement calls which can have a dramatic effect on the results. To take a concrete example, the effects of income inequality have been a topic of recent political discussion. However, 'income inequality' can be measured in different ways[2]. This includes choice of index (e.g. Gini coefficient), income concept (e.g. gross or net), and unit of analysis (e.g. household versus individual). These choices can dramatically change how countries compare in terms of inequality (Solt 2009), and therefore inequality's apparent association with important social outcomes.

For policymakers to have confidence in the analysis process, transparency is vital. It should be easy for independent researchers to verify how figures have been produced, and whether small adjustments in methodology (e.g. using one inequality measure in place of another) has

[2] This is also an area where errors have been found in QSS work, undermining the credibility (and the general public's trust) in results (see Giles 2014).

any effect on results. Just as schoolchildren must show workings in their mathematics homework, one might presume academics are required to show how they produce their figures. However, this is not the case. For results to be truly independently verifiable, at least two conditions must hold:

(i) The data must be publicly available and free to download.

(ii) The exact analysis steps must be publicly available.

Unfortunately most QSS research falls down on at least one of these counts. Although the ESRC has made substantial progress in making data publicly available, many important resources remain locked behind large fees[3] or restrictive access agreements (e.g. the Twins Early Development Study, TEDS, and data produced by the Universities and Colleges Admissions Services, UCAS).

Even where data are freely and publicly available, the goal of true independent verifiability remains out of reach. This is due to academics' 'program code' not being freely available. 'Program code' refers to computer instructions which, when applied to the data, produce the final reported figures. These instructions can run to thousands of lines, with each potentially containing a choice (or mistake) influencing the results. The methods section of an academic paper cannot plausibly cover all of these instructions in detail. It is therefore commonplace to reduce this code to a few general sentences. As Anderson et al (2005:101) note '*an applied article is only the advertising for the data and code that produced the published results*.' In other words, for most QSS, only the paper (the 'advertisement') is available and forms part of the evidence base. The actual substantive research (i.e. the analysis steps and program code) is not.

[3] Datasets like the Avon Longitudinal Study of Parents and Children (ALSPAC) can cost thousands of pounds (see ALSPAC 2013:3). This is a clear barrier to researchers wishing to replicate findings based on this dataset.

McCullough, McGeary and Harrison (2008) highlight the importance of this issue. A few journals have previously asked authors to share their data and code, including the journal *Federal Reserve Bank of St. Louis Review*. McCullough et al (2008:1416) attempted to reproduce results from 117 studies using this resource. They concluded that '*when all was said and done, we were able to replicate only 9 of the 117 articles*.' Similarly, McCullough, McGeary and Harrison (2006) found that only 62 out of 186 studies published in the *Journal of Money, Credit and Banking* shared their data and code (even though all authors were required to do so). Moreover, of these 62 studies, only 14 were actually replicable. The work of King (2003), Ray and Valeriano (2003), Boyer (2003) and Neuliep (1991) suggest that this is a serious issue across the quantitative social sciences.

The relevance for public policy can be illustrated with a recent example. A working paper published by two Professors (one at Harvard University) suggested a country's economic growth begins to suffer as national debt reaches 90 percent of GDP (Reinhart and Rogoff 2010). On the basis of this finding, a policymaker might decide to introduce public spending cuts to avoid this 'debt cliff'. Indeed, senior policymakers in the US and Europe (e.g. UK Chancellor of the Exchequer, George Osborne) cited this research to support austerity programmes (Arthur and Inman 2013). However, as with most QSS research, the workings and program code behind this analysis were not published alongside the paper

The danger of program code remaining unpublished is that mistakes can happen (they are a fact of life). And a mistake did happen in this case. After failing to replicate Reinhart and Rogoff's results, a PhD student contacted the Professors asking for their data and program code (which, to their credit, they provided – something they were under no obligation to do). The student found a typo meaning that 5 of the 20 countries had accidently been excluded from the analysis which, along with some other anomalies, made a substantial difference to the results[4].

One might argue this illustrates that the current system works. An error was made, but it was subsequently caught by another researcher, with the evidence base ultimately self-correcting. To this we argue it was only the openness of these researchers in sharing their calculations that allowed this error to become known. Had they been unwilling or (more likely) too busy to respond, this would have remained undetected, permanently distorting the evidence base. This is a significant issue given that authors may not share their raw calculations, even when this is required by journals (see our discussion of McCullough et al [2006] above).

In QSS, as in any other field, mistakes happen. The only way to limit their impact is to require QSS research to be independently verifiable. It is only the 'many eyes' of other researchers that will help prevent unreliable evidence causing social harm. The simplest solution is for data and code to be made openly accessible whenever this is legally possible. This requires action from a central organisation with leverage over researchers and publishers. For example, the ESRC could make publication of program code (where possible) a pre-condition of funding.

Failure to communicate uncertainty

Policymakers often employ QSS evidence as if it provides clear, unambiguous facts about the social world. For example, evidence from QSS has been used by UK politicians to assert with certainty that English children's numeracy skills have declined relative to other countries (Jerrim 2013); and that low ability rich children overtake high ability poor children in education by age 10 (Jerrim and Vignoles 2013). Unfortunately, QSS findings are usually subject to much more uncertainty than seems to be commonly understood by policymakers.

[4] See http://www.bbc.co.uk/news/magazine-22223190.

Of course, a certain level of uncertainty is inevitable in any research field. What is vital is that this be clearly communicated to readers. Whilst many researchers do highlight a range of uncertainties (e.g. unrepresentative samples, missing data) this is not always the case. Indeed, often one type of uncertainty (sampling variation) takes precedence over others (Gorard 2010)[5]. This form of uncertainty means researchers could observe a finding by 'chance', due to the fact that they (typically) only have data from a random sample of individuals, rather than the whole population. This is well understood in QSS, and quantified within the framework of statistical significance testing.

There is a strong focus on this uncertainty within QSS. Testing for statistical significance is taught for several weeks in most statistics courses, and is expected in most QSS publications (Gorard 2010). This can, however, obscure other important details of a result (Sterne and Davey-Smith 2001). Importantly, a result being 'statistically significant' does not mean it is policy relevant or important. For example, female maths students may answer 50% of test questions correctly, compared to 49.9% for men. With a large enough sample, this difference may be 'statistically significant', but this does not mean it warrants any kind of policy attention.

In no way does statistical significance rule out other possible uncertainties. Indeed, on many occasions unrepresentative samples, missing information, and poorly measured data pose a much greater threat to results (e.g. Jerrim 2013). Yet these challenges typically receive less attention than significance testing – both within academic papers and the training of students (Gorard 2010).

[5] An example is meta-analyses, where individual studies are typically weighted by the size of the standard error. Hence studies with lower sampling variation get more weight in results. Yet this ignores that these studies could be of lower quality (and have more uncertainty) in other dimensions (e.g. missing data and measurement error).

Take measurement error as an example. To produce informative results, the variables used in any analysis must be well-measured. Findings from poorly measured data can still be statistically significant, but this does not mean they are useful. This issue is hugely important in QSS, as researchers often rely on whatever data are available on a given issue. For example, investigating social mobility in Singapore, Ng (2009) measured the association between the incomes of survey respondents and those of their fathers when the respondents were children. However, fathers' income was reported by their children based on their own memories. Similarly, Brunello, Weber and Weiss (2012) investigated the link between the number of books there were in a person's home when they were age 10 and their salary in later life. The problem being that the information on books was based on 50 – 80 year olds recalling their bookshelves more than 40 years on. Strong, policy-relevant conclusions could be drawn from these results – that income mobility in Singapore is low, or that the number of books at home has a substantial effect on children's long-run outcomes. However, these conclusions would not recognise the great uncertainty arising from the use of weak measures.

Our argument is not that QSS studies should be perfect; researchers must commonly make-do with what data is available. Rather it is that the wide array of uncertainties (including weak measures, measurement error, and missing data) should be transparently communicated, beyond the focus on statistical significance. To our knowledge, there are no widely used guidelines for transparent reporting of QSS results. However, we offer the following as a first suggestion towards a 'gold-standard':

- Data and program code should be freely available
- The paper should contain a table showing the extent and selectivity of missing data
- There should be a discussion of possible measurement error for each of the key variables

- A range of sensitivity analyses should be shown in an appendix, including adjustments for missing data, investigations of measurement error, different operationalization and measurement of key variables, different regression model specifications and use of different statistical techniques
- Key findings should be shown to hold when using an alternative dataset

In reality, such a standard is likely to be difficult. Finding a second dataset to replicate results is often not possible, for example. Nevertheless, it is reasonable for policymakers to expect these issues to have been considered in QSS research, with evidence presented on their likely impact on results (where feasible). We suggest journals require that papers include a specific section providing straightforward answers to each of these issues.

Publication bias

Publication bias refers to certain types of result being more likely to be published in academic journals than others, for reasons other than scientific merit. This has been extensively studied in the medical literature, where it has been observed that 'positive' (i.e. statistically significant) results are more likely to be published (Easterbrooke et al 1991). There is no good scientific reason for this – a study showing no significant effect is often just as important as one showing a substantial effect.

Doctors are increasingly aware that this type of systematic publication bias harms patients (Dirnagl and Lauritzen, 2010). If studies showing that a drug is effective are more likely to be published than those that do not, then doctors may end up believing the drug to be more effective than it really is. They may then be more likely to prescribe the drug for patients, despite the totality of the evidence (i.e. from both published and unpublished studies) actually being rather mixed.

There is potentially a similar issue in public policy. If policymakers are exposed to a distorted picture of the scale of a social problem, or of the effectiveness of a particular policy solution, this restricts their ability to make informed decisions. While QSS may be less affected by commercial pressure than medical research, there are other forms of bias which may distort the messages policymakers receive.

Previous research has identified bias towards ‘positive’ findings as particularly pernicious (e.g. Fanelli 2012). This bias can be divided into two components. First, journal editors and peer reviewers seem to prefer clear, significant findings to negative or null results (Ioannidis and Trikalinos 2005; Young et al 2008). This is clearly demonstrated by Emerson et al (2010), who sent two versions of a fabricated manuscript to 238 reviewers. The only difference between the versions was that one showed positive results and the other null results. Reviewers rated the ‘positive’ manuscript more highly, and were more likely to recommend it for publication.

Franco et al (2014) highlight the second component of publication bias, which intrudes before journal submission. They found researchers were 60 percentage points more likely to submit strong, significant findings for publication than null findings. This means that tests of many hypotheses never enter the academic evidence base. Although there are methods to detect publication bias in meta-analyses and systematic reviews (e.g. funnel plots), these have limitations. For instance, Lau et al (2006) note that funnel plots can mistake genuine differences in effect sizes between small and large studies (due to differences in target populations) for publication bias[6].

[6] Moreover, funnel plots can only identify publication bias due to the chase for ‘statistically significant’ results. It cannot detect journals’ preferences for ‘eye-catching’ findings, or researchers’ reduced incentives to write up findings with small effect sizes (even if a large sample means it can be labelled statistically significant).

This problem may be particularly acute in high-profile journals, which gain their status partly from the number of people citing articles they publish. They therefore have a strong incentive to carry eye-catching findings. Indeed, Barto and Rillig (2012) find that high profile journals are more likely to publish research showing large effects, with null results more likely to be published in low-tier journals (Littner et al 2005). This is partly why studies showing positive results are more often cited (Barto and Rillig 2012; Jannot et al 2013).

The lure of an influential article in a 'top' journal could also encourage researchers to exaggerate or selectively report results. Whilst most researchers are committed to conducting and reporting honest research, questionable research practices remain an issue. Concrete academic malpractice (e.g. altering or inventing results) is thankfully rare, though a recent meta-analysis found that two percent of academics admitted falsifying data (rising to 14 percent when asked about research practices of colleagues - Fanelli 2009)[7]. These figures are much higher for less extreme forms of poor practice, such as dropping data points based on a 'gut feeling' or selectively reporting results. Fanelli (2009) found 34 percent of researchers admitting to such practices, and 70 percent said their colleagues engaged in them. Similarly, a meta-analysis by Dwan et al (2008) found selective reporting of positive results even when outcomes were pre-specified in a registered protocol.

The majority of the above studies are drawn from medicine. However, similar processes are likely to operate in QSS. In fact, these factors may be even more prevalent, due to the reliance of QSS on 'observational' data, compared with the greater use of experimental methods in medicine (randomised controlled trials – RCTs; Haynes et al (2012).

RCTs are expensive and time-consuming to conduct (Ho et al 2008). Research teams enter the field to recruit participants, administer treatments, define control/intervention groups and

[7] This study included academics from any scientific discipline.

measure outcomes. These trials are usually pre-registered on a central database, making it hard (though not impossible) for there to be no record of an RCT (even if findings don't appear in an academic journal). Moreover, academics running RCTs face significant upfront costs –a lengthy ethics process, recruitment of participants, administration of the intervention, and management of the trial. Thus, by the time data analysis can begin, academics have already invested a lot of time and effort they cannot get back whether results are published or not. There are hence still reasonable incentives for academics who have conducted an RCT to write up their findings, even when results are small or statistically insignificant[8].

The same is not true for QSS, which makes extensive use of pre-existing data resources, normally funded and collected by a central organisation and freely shared for research purposes. Consequently, it is easy to test new hypotheses quickly, and to get provisional results within a matter of days. This means little is lost if findings are never written up. This is a huge potential source of publication bias, which is almost impossible to detect, and whose consequences are rarely recognised.

Added together, these biases have uncomfortable implications for public policy. If insignificant, small, or uninteresting findings are rarely written up, the evidence base becomes dominated by exciting, surprising, positive findings. Fanelli (2012) found that this bias has grown over time, particularly within social science disciplines. Consequently the QSS literature is likely to exaggerate (i) the extent and severity of social problems and (ii) the extent to which policy interventions can effect change. This has important implications, potentially leaving policymakers too active in public policy – investing in ineffective solutions to social problems that might not even exist.

[8] There are other biases which create difficulties for the medical literature, such as the influence of pharmaceutical companies, which do not strongly affect the QSS evidence base

Unfortunately, another source of bias amplifies this problem yet further – the mainstream media. This is the medium through which QSS findings are often made accessible to non-academic audiences. Unlike academic journals, the media need not even try to fairly represent the evidence base within a given field. They instead concentrate on the most striking, controversial, and politically sensitive findings – often focusing on 'bad news' or 'scare' stories (Goldacre 2008). As well as highlighting only the most eye-catching QSS findings, the mainstream media also often exaggerate results – e.g. claiming causality from purely correlational research (Sumner et al 2014), or generalising findings from a small, unrepresentative sample to an entire population (Pellechia 1997). Worryingly, these exaggerations can often be traced back to press-releases issued by universities and academic journals to promote a particular paper (Sumner et al 2014)[9].

Of course policymakers are not only exposed to QSS findings through the media. Policy development is often collaborative, involving extensive reviews of the evidence, and expert advice from independent academics. However, individual, highly publicized QSS studies can bypass this process to have an outsized influence on policy discussions. One such example is the Reinhart and Rogoff (2010) study described above, and we discuss a further example in our case-study below.

Given the strong effect of the media, it is unrealistic to expect policymakers to have a completely unbiased picture of the quantitative evidence on a given topic. However, some practical steps could be taken by the QSS community, and by academic publishers, to reduce the biases outlined above:

1. Extending open publishing models, such as that adopted by PLOS One, across all QSS journals. Rather than allowing editors and peer-reviewers to make judgments

[9] This research was conducted in health sciences – however it is likely to also apply to QSS research.

based on interest and relevance, PLOS One publishes all studies deemed methodologically sound. At a stroke, this reduces journals' incentives to publish only 'exciting' results, and the incentive for researchers to fail to publish small, uninteresting, or insignificant findings.

2. A requirement that research projects be pre-registered in a central location before data access is provided, along with publication of research protocols before projects begin (as per best practice in the RCT literature). Journals such as the Journal of Work, Ageing, and Retirement (which publishes QSS research), are moving in this direction through 'Registered Reports'. This means studies are pre-registered for publication (based on peer-review of their protocols) prior to data analysis taking place – preventing them from going 'missing in action' based on their results.

3. Increase the accountability and transparency associated with academic press-releases. For instance, Goldacre (2014) recommends that all releases have named authors, including both the press-officers and the academics, and should contain direct links to the academic paper.

Peer review does not guarantee research quality

At the heart of quality assurance in academia is the publication of articles in peer-reviewed journals. It is through such publications that academic QSS receives its 'quality' stamp, illustrating the work has been scrutinized and accepted by other experts in the field.

There seems to be a strong belief amongst policymakers and the public that academic research has a strong quality assurance process, with papers published in peer-reviewed journals representing a high standard of work. A search of UK parliamentary debates yields a large number of appeals to the fact that research has been 'peer-reviewed' as an indicator of its quality. For instance, Caroline Lucas MP implicitly used the fact a particular study had been 'peer-reviewed' to try to dismiss a key government claim:

> '*Will the Minister explain why the Government's leaflet on "TTIP myths" claims that a family of four would benefit by £400 a year yet makes no mention of the peer-reviewed paper from Tufts university that predicts that over 10 years the average working Briton will be more than £3,000 worse off as a result of the lower wages that TTIP will fuel'* (House of Commons Debate, 15 January 2015, c)

Indeed, even in the midst of the panic surrounding the H1N1 pandemic in Canada, senior academics emphasised the need for vaccine research to have the 'imprimatur of a high impact peer review journal' before it could be considered in public policy (PLoS Medicine editors 2010).

In reality, peer-review is far from a fool-proof system. The typical process is for journal editors to send a manuscript out to (usually) two academic experts, who provide comments and either recommend it for publication, reject the paper, or suggest resubmission after revisions. If you have your paper rejected, you move on to another journal, where this process begins again. In theory, every new peer review phase should improve the rigour of the paper, until it is both methodologically sound and relevant to the readership of the journal.

Appreciating the limitations of this process has important implications for the use of QSS in public policy. First, acceptance for publication is heavily dependent on the opinion of reviewers. Ideally, this should be based on objective criteria, resulting in a high level of consistency. However, this has been described as a 'fantasy' by a former editor of the British Medical Journal (Smith, 2006), who noted that 'inevitably, people will take different views on [a paper's] strengths, weaknesses, and importance'. This means that publication at any particular journal depends on who judges the paper (and thus involves a certain amount of luck). Indeed, meta-analyses of peer review have shown extremely low levels of agreement between reviewers (Bornmann et al 2002), supporting Smith's (2006) assertion that much of peer review was 'little better than tossing a coin'. Moreover, while Rothwell and Martyn (2000) suggest that standardised review forms may improve consistency, our experience is that few QSS journals apply this practice consistently. This can mean that the label 'peer-reviewed' does not indicate that a paper has met an absolute standard of quality. Instead it may simply mean that, among the many editors and reviewers who have seen the paper, at least someone deemed it worthy of publication. The limitations of peer review are strongly highlighted by Gans and Shepherd (1994) who contacted a number of leading economists about their experience of publication. They found that several Nobel Prize winning papers were initially rejected by journals due to negative reviews - illustrating how even very good ideas can be dismissed.

A further wrinkle is added by the fact that, due to the glacial pace of the journal review process (it often takes a year for papers to be accepted after initial submission), many academics now produce 'working papers' (e.g. Ammermueller 2006). These are preliminary versions that authors release into the public domain before the academic review process has begun. As such, most working papers have been through a weaker form of peer review (an internal department review) at best. Nevertheless, they are often cited by policymakers (e.g. the Reinhart and Rogoff paper described above).

Of course, there are other potential ways to identify a high quality study, such as publication in a prestigious journal. However, as previously discussed, high status journals may be particularly prone to publication bias (Franco et al 2014). Indeed, there are many examples of high-profile, but low-quality studies published in highly respected journals (e.g. the widely discredited study of the MMR vaccine by Wakefield et al [1998] was published in the

Lancet). Put simply, it is very difficult to know what represents a ‘good journal’ where one could expect to find high quality research. Many academics may have some notion of what the ‘good’ journals are within their own field. However, this hierarchy is probably unknown to individuals working outside academia (or, indeed, outside the specific academic discipline).

To summarise, the quality assurance procedure employed in academia is considerably looser than seems to be assumed by policymakers and the public. Indeed, perhaps the most damning indictment comes from Gans and Shepherd (1994:179) who, after sharing correspondence with 15 Nobel Prize winning economists, stated:

> *‘the outpouring of irritation and anger at the publication process that our project provoked—by the famous economists whom the process has benefitted most—creates concern about whether the process functions adequately.’*

As with publication bias, there may be no satisfactory resolution to all the limitations of peer-review. Any system which could effectively weed out all low quality research would likely be punitively expensive and time-consuming. Consequently, some have suggested a stronger role for quality assurance activities *after* publication – ‘post-publication peer review’ (Hunter 2012). The rise of open access electronic publication allows for a type of crowd-sourced peer review – the ‘many eyes’ of other academics evaluating research and judging its quality. Currently, only very limited metrics are collected on a given article (e.g. number of citations) and these are, at best, weak measures of research quality (Seglen 1997; PLoS Medicine Editors 2006). However, internet technology allows for much more than this, from reader comments and direct evaluations, to deeper analysis of the nature of citations. The value of these systems would be enhanced by the publication of program code (as recommended above) alongside academic articles – allowing readers to easily replicate studies, and feed results back into the evidence base. There would undoubtedly be difficulties with this framework. However, unlike the present system, it would allow readers to have a ready gauge of the academic response to a particular study – whether it has been widely replicated and accepted, or whether other academics have concerns.

Policymakers' use of QSS evidence

The primary purpose of this paper is to highlight the most pressing challenges faced by QSS as a policy-making resource. However, these challenges are often compounded by how policymakers deploy this evidence. Here we describe a case-study involving a problematic use of QSS evidence. We also suggest ways research practice could be improved to address the issues we discuss.

Our case study is based upon Feinstein (2003). A single graph from this paper, suggesting poor children who performed well on developmental tasks at 22 months were 'overtaken' by age 10 by rich children who did poorly on the same tasks, has had significant policy impact. Former UK Deputy Prime Minister Nick Clegg used this to suggest:

'*By the age of five, bright children from poorer backgrounds have been overtaken by less bright children from richer ones—and from this point on, the gaps tend to widen still further'*

While former Secretary of State for Education Michael Gove stated:

'*rich thick kids do better than poor clever children when they arrive at school*'

David Halpern (former chief analyst at the Prime Minister's Strategy Unit) even more tangibly illustrates how a single graph from this study managed to bypass the ideal system of evidence review:

'*one of the Ministers present tore out one of the Strategy Unit's slides and – leaning forward to put it in front of the Prime Minister declared – '...but what are we going to do about this?'*

The slide ... showed how the cognitive ability of bright children from poor backgrounds appeared to be overtaken by that of much less able children from affluent backgrounds ... Within a year more than £500m was assigned to build a programme of pre-school provision for the UK.' [Institute of Education 2010].

These statements were made despite Feinstein placing important caveats on the results; the sample selected was unusual (and not nationally representative), there were challenges with missing data, and the number of observations was small (just 36 children in the high scoring poor group). The data also came from the 1970s, and so was of limited relevance to contemporary public policy.

Moreover, a replication study was conducted by Jerrim and Vignoles (2013) (available as a working paper in 2011). They discussed a further limitation of the evidence, focusing on measurement error in children's test scores and 'regression to the mean.' The authors highlight how such statistical issues could be driving the graph that became so highly cited by public policymakers. The author of the original study himself recently stated that his paper[10]:

'*never says anything about bright or dim kids. Nor there being a specific age of some type of formal crossover. I certainly have never said it*.'

This marks a clear example where the transfer of QSS evidence to policy has not worked as one might hope. Drawing on the themes outlined in previous sections, what can we learn from the experience of this work?

First, transparency and independent verification are vital. The study by Jerrim and Vignoles, which corrected policymakers' interpretation of the evidence, was only possible due to open and free data access, and the methodological details provided in the original study. (Open access to the original program code would have made this easier. But, on this occasion,

[10] See http://blogs.lse.ac.uk/impactofsocialsciences/2015/01/21/misunderstanding-data-feinstein/

replication was possible due to the simplicity of the methods used, and the rigor of their description in the original paper). It hence demonstrates that anything which helps improve the replicability, transparency and independent verification of academic studies (e.g. open access data and code) can only benefit the evidence-policy interface.

Second, it illustrates the importance of recognising all forms of uncertainty, and doing everything possible to ensure policymakers cannot misinterpret results. The difficulties with this evidence could not be revealed by statistical significance testing alone, which is why Feinstein (2003) placed a number of other important caveats on his results. This, however, did not stop his work being misused, possibly due to findings being presented as a graph. Graphs are very effective communication tools – but they often do not capture many of the uncertainties present in QSS research. Not only do all forms of uncertainty need to be recognised in a study, but they also must be clearly articulated in the presentation of results.

Third, it illustrates how the peer-review process can operate sub-optimally, and does not rule out certain limitations with the evidence going unnoticed. Indeed, we believe this to be a good example of where ‘post-publication’ peer review could have led to an earlier warning of potential difficulties.

Finally, despite growing use of systematic reviews and meta-analyses, this case study highlights how findings from single, small-scale studies can still have outsized influence on public policy. This makes the limitations outlined above even more pressing. We cannot rely on policymakers to employ a systematic approach to the evidence on any given topic. Therefore problems with any individual paper may not necessarily be ‘washed-out’ by subsequent research.

The folly of taking such a one-dimensional approach to QSS evidence was recognised by David Willetts (former Minister of State for Universities and Science) who stated[11]:

'Sometimes over-reliance on one specific piece of evidence can leave you vulnerable. I remember being influenced by Leon Feinstein's very interesting paper.....I served on Nick Clegg's social mobility group and recommended this powerful evidence to him and he too was impressed and cited it. But Leon's work was challenged by other academics because it was affected by reversion to the mean. The result was that the Guardian ran a piece that the Coalition's social mobility strategy was undermined because the research on which it rested had been disproved. That is not, of course, a reason for giving up on evidence-based policy: but it is a reminder of how careful we have to be in using it.'

This point is key. Despite the limitations of QSS, we should not abandon an evidence-based approach to policy-making. Rather, both academics and policymakers need to engage more critically with the information that QSS provides, and to work out how the quality of evidence used in policy-making can be improved.

Conclusions

QSS has the potential to make a significant contribution to public policy. Ideally, QSS research should:

- Be independently verifiable
- Clearly communicate the uncertainty associated with any given result
- Be free from publication and media bias
- Be subject to a clear, consistent, and transparent quality assurance process

[11] See http://blogs.lse.ac.uk/impactofsocialsciences/2015/01/12/the-messiness-inherent-to-policymaking/

In this ideal world, the QSS evidence base would be one of the most powerful tools available to policymakers. Unfortunately, as we have shown in this paper, this is not the world in which we currently live. It is perhaps unrealistic to expect all the above ambitions to be achieved in the near future. However, two practical steps in particular could easily be taken by QSS researchers and funders to make a dramatic positive difference.

First, the ESRC and other research funders should insist that QSS research be independently verifiable – including mandatory sharing (where possible) of all data and program code. This is a quick, cheap and easy policy to implement – yet one that has the potential to increase the quality and transparency of academic QSS research immensely.

Second, QSS research reports should be required to go much further in explaining the uncertainty surrounding their results – beyond the current fixation with statistical significance. In the RCT literature, standardised reporting (CONSORT – Schulz et al 2010) and methodological quality (the Jadad scale – Jadad et al 1996) scales have been developed for such purposes, with other academic disciplines following suit (e.g. the ongoing development of the Newcastle-Ottawa scale in epidemiology – see Wells et al 2013). Such standardised frameworks are a powerful tool for communicating the strengths and weaknesses of quantitative research. Funding should therefore be made available to enable the development of similar scales for QSS. Similarly, clear quality appraisal criteria should also be developed to assist the QSS peer-review process, offering standardised guidelines to be applied across journals.

These recommendations cannot overcome all the limitations discussed in this paper. Indeed, some of the challenges highlighted, such as media bias and arbitrary peer review, will likely be difficult to solve. However, this does not mean that these challenges should be ignored. Indeed, for QSS research to progress, they must be openly acknowledged and discussed. QSS can make an important contribution to public policy. But more consideration needs to be given to the limitations of what can be achieved, what expectations of QSS are realistic, and how the QSS community can do their utmost to ensure these expectations are met. This paper has tried to illuminate these challenges for non-specialist audiences, and has hopefully offered some practical solutions. Nevertheless, further work is needed to ensure that academic QSS produces the maximum possible benefit for the general public who, after all, pay for the vast majority of our research.

References

ALSPAC, 2015, Access Policy, Accessed 28/05/2015 from http://www.bristol.ac.uk/media-library/sites/alspac/documents/research/Access%20Policy_v6.0.pdf

Allen, M, and Preiss, R, 1997, Comparing the persuasiveness of narrative and statistical evidence using meta-analysis, *Communication Research Reports*, 14:2, 125-131

Ammermueller, A, 2006, Educational opportunities and the role of institutions, *ZEW Discussion Paper Number 05-44*, ftp://ftp.zew.de/pub/zew-docs/dp/dp0544.pdf

Anderson, R, Greene, W, McCullough, B and Vinod, H, 2008, The role of data/code archives in the future of economic research, *Journal of Economic Methodology*, 15,1, 99–119

Arthur, C and Inman, P, 2013, The error that could subvert George Osbourne's austerity programme, *The Guardian 18th April 2013,* www.theguardian.com/politics/2013/apr/18/uncovered-error-george-osborne-austerity

Barto, E, K, and Rillig, M, C, 2012, Dissemination biases in ecology: effect sizes matter more than quality, *Oikos*, 121, 2, 228-235

Bornmann, L, Mutz, R, Daniel, H, D, 2010, A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rate reliability and its determinants, *PLoS One*, 5, 12, e14331

Boyer, M, 2003, Symposium on replication in international studies research, *International Studies Perspectives*, 4, 1, 72-107

Brownson, R, Chriqui, J, and Stamatakis, K, 2009, Understanding evidence-based public health policy, *American Journal of Public Health, 99,9, 1576-1583*

Brunello, G, Weber, G, and Weiss, C, 2012, Books are forever: early life conditions, education and lifetime income, *IZA Discussion Paper 6386* http://ftp.iza.org/dp6386.pdf.

Dirnagl, U, and Lauritzen, M, 2010, Fighting publication bias: introducing the negative results section, *Journal of Cerebral Blood Flow and Metabolism*, 30, 1263–1264

Dwan, K, Alrman, D, G, Arnaiz, J, A, Bloom, J, Chan, A, W, Cronin E, Williamson, P, R, 2008, Systematic review of the empirical evidence of study publication bias and outcome reporting bias, *PLoS One*, 3, 8, e3081

Easterbrooke, P, J, Berlin, J, A, Gopalan, R, 1991, Publication bias in clinical research, *The Lancet,* 337, 8746, 867-72

Emerson, G, B, Warme, W, J, Wolf, F, M, Heckman, J, D, Brand, R, A, Leopold, S, S, 2010, Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial, *Archives of internal medicine,* 170, 21, 1934-1939

Fanelli, D, 2009, How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data, *PLOS-ONE*, doi: 10.1371/journal.pone.000573

Fanelli, D, 2012, Negative results are disappearing from most disciplines and countries, *Scientometrics*, 90, 3, 891-904

Feinstein, L, 2003, Inequality in the Early Cognitive Development of British Children in the 1970 Cohort, *Economica*, 70, 73-97.

Franco, A, Malhotra, N, and Simonovits, G, 2014, Publication bias in the social sciences: Unlocking the file drawer, *Science*, 345, 6203, 1502-1505

Giles, C, 2014, Piketty findings undercut by errors, *Financial Times 23rd May 2014*, http://www.ft.com/intl/cms/s/2/e1f343ca-e281-11e3-89fd-00144feabdc0.html#axzz3bVqJs3vB

Gans, J, and Shepherd, G, 1994, How are the mighty fallen: Reject classic articles by leading economists, *Journal of Economic Perspective*, 8, 1, 165-79

Goldacre, B, 2008, *Bad Science: Quacks, Hacks, and Big Pharma Flacks*, McClelland and Stewart

Goldacre, B, 2013, *Bad Pharma: How Medicine is Broken and How We Can Fix It*, Fourth Estate: London.

Goldacre, B, 2014, Preventing bad reporting on health research, *British Medical Journal*, doi: http://dx.doi.org/10.1136/bmj.g7465

Gorard, S, 2010, All evidence is equal: the flaw in statistical reason, *Oxford Review of Education*, 36, 1, 63-67.

Haynes, L, Service, O, Goldacre, B, Torgerson, D, 2012, Test, learn, adapt: Developing public policy with randomised controlled trials, *Cabinet Office Report, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf*

Hunter, J, 2012, Post-publication peer review: opening up scientific conversation, *Frontiers in Computational* Neuroscience, doi: 10.3389/fncom.2012.00063

Ho, M, Peterson, P, and Masoudi, F, 2008, Evaluating the evidence: is there is rigid hierarchy?, *Circulation*, 118, 1675-1684

Institute of Education, 2010, Case study on the impact of IOE research. The British Birth Cohort studies, http://www.ioe.ac.uk/Research_Expertise/IOE_RD_A4_BCS_1.3_d.pdf

Ioannidis, J, and Trikalinos, T, A, 2005, Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials, *Journal of Clinical Epidemiology*, 58, 543-549

Ioannidis, J, 2006, Evolution and translation of research findings: from bench to where?, *PLoS Clinical Trials*, doi: 10.1371/journal.pctr.0010036

Jadad, A, Moore, R, A, Carroll, D, Jenkinson, C, Reynolds, D, J, M, Gavaghan, D, J, and McQuay, H, J, 1996, Assessing the quality of reports of randomized clinical trials: is blinding necessary?, *Controlled Clinical Trials*, 17,1, 1–12

Jannot, A, Agoritsas, T, Gayet-Ageron, A, Perneger, T,V, 2013, Citation bias favoring statistically significant studies was present in medical research, *Journal of Clinical Epidemiology*, 66, 3, 296-301

Jerrim, J, and Vignoles, A, 2013, Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes, *Journal of the Royal Statistical Society series A*, 176, 4, 887 – 906

Johnson, S, and Antill, S, 2011, Impact evaluation of the Millennium Cohort Study, http://www.esrc.ac.uk/_images/MCS_Impact_Evaluation_September_11_tcm8-17258.pdf

Kelley, J, and Simmons, B. 2015, Politics by number: indicators as social pressure in international relations, *American Journal of Political Science,* 59, 1, 55-70

King, G, 1995, A revised proposal, proposal, *Political Science and Politics*, 28, 3, 494-99

Lau, J, Ioannidis, J, Terrin, N, Schmid, C, and Olkin, I, 2006, The case of the misleading funnel plot, *British Medical Journal*, 333, 7568, 597–600

Littner, Y, Mimouni, FB, Dolberg S, Mandel, D, 2005, Negative results and impact factor: a lesson from neonatology, *Archives of Pediatric and Adolescent Medicine*, 159, 1036-1103

McCullough, B, McGeary, K, and Harrison, T, 2006, Lessons from the JMCB archive, *Journal of Money, Credit and Banking*, 38, 4, 1093-1108

McCullough, B, McGeary, K, and Harrison, T, 2008, Do economics journals promote replicable research?, *Canadian Journal of Economics*, 41, 4, 1406-20

Merry, S, 2011, Measuring the world: indicators, human rights, and global governance, *Current Anthropology*, 52, s3, s83 – s95

Ng, I, Shen, X, and Ho, K, W, 2009, Intergenerational earnings mobility in Singapore and the United States, *Journal of Asian Economics*, 20, 2, 110-19

Neuliep, J, 1991, *Replication research in the social sciences,* New York: Sage Publications

Parsons, D, Thomas, R, Strange, I, and Walsh, K, 2014, Evaluating the impact of ESRC economics centres: Final report of the evaluation, http://www.esrc.ac.uk/_images/Evaluating_the_impact_of_ESRC_economics_centres_tcm8-31189.pdf

Pellechia, M, G, 1997, Trends in science coverage: a content analysis of three US newspapers, *Public Understanding of Science*, 6, 1, 49-68

PLoS Medicine Editors, 2006, The impact factor game, *PLoS Medicine*, 3, e291

PLos Medicine Editors, 2010, Journals, Academics, and Pandemics. *PLoS medicine,* 7, 5, e1000282

Ray, J, L, and Valeriano, B, 2003, Barriers to replication in systematic empirical research on world politics, *International Studies Perspectives*, 4, 1, 79-85

Reinhart, C, and Rogoff, K, 2010, Growth in a time of debt, *American Economic Review: Papers and Proceedings*, 100, 2, 573–78

Rothwell, PM, and Martyn, CN, 2000, Reproducibility of peer review in clinical neuroscience: Is agreement between reviewers any greater than would be expected by chance alone?, *Brain*, 123, 9, 1964-1969

Schmierbach, M, 2005, The influence of methodology on journalists' assessments of social science research, *Science Communication,* 26, 3, 269-287

Schulz, K, Altman, D, Moher, D, 2010, CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials, *British Medical Journal,* 340, doi: http://dx.doi.org/10.1136/bmj.c332

Seglen, P, 1997, Why the impact factor of journals should not be used for evaluating research, *British Medical Journal,* 314, 498–502.

Smith, R. 2006, Peer review: a flawed process at the heart of science and journals, *Journal of the Royal Society of Medicine*, 99, 7, 178-182

Solt, F, 2009, Standardizing the World Income Inequality Database, *Social Science Quarterly*, 90, 2, 231-242

Sterne, J, and Davey-Smith, G, 2001, Shifting the evidence – what's wrong with significance tests, *British Medical Journal*, 322, 7280, 226-231.

Sumner, P, Vivian-Griffiths, S, Boivin, J, Williams, A, Venetis, C, Davies, A, Ogden, J, Whelan, L, Hughes, B, Dalton, B, Boy, F, Chambers, C, 2014, The association between exaggeration in health related science news and academic press releases: retrospective observational study, *British Medical Journal,* 349, http://dx.doi.org/10.1136/bmj.g7015

Tewksbury, R, 2009, Qualitative versus quantitative methods: understanding why qualitative methods are superior for criminology and criminal justice, *Journal of Theoretical and Philosophical Criminology*, 1, 1, 38 -58

Wakefield, A, J, Murch, S, H, Anthony, A, Linnell, J, Casson, D,M, Malik, M, Berelowitz, M, Dhillon, A, P, Thomson, M, A, Harvey, P, Valentine, A, Davies, S, E, Walker-Smith, J, A 1998, Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children, *Lancet*, 351, 9103, 637–41

Wells, G, A, Shea, B, O'Connell, D, Peterson, J, Welch, V, and Losos, M, 2013, The Newcastle-Ottawa scale for assessing the quality of nonrandomized studies in meta-analyses, http://www.ohri.ca/programmes/clinical_epidemiology/oxford.htm

Young, NS, Ioannidis, J, Al-Ubaydli, O, 2008, Why current publication practices may distort science, *PLoS Medicine*, 5, 10, e201