View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Warwick Research Archives Portal Repository

Original citation:
Smith, Benjamin K. and Jensen, Eric. (2016) Critical review of the United Kingdoms "gold standard" survey of public attitudes to science. Public Understanding of Science, 25 (2). pp. 154-170.
Permanent WRAP URL:
http://wrap.warwick.ac.uk/82074

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
'The final, definitive version of this paper has been published in
Public Understanding of Science, 25 (2). pp. 154-170. 2016 by SAGE Publications Ltd, All rights reserved. © The Author(s)
Published version: http://dx.doi.org/10.1177/0963662515623248

A note on versions:
The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP URL' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

Critical Review of the UK's "Gold Standard" Survey of Public Attitudes to Science

Benjamin K. Smith (University of California, Santa Barbara)

Eric A. Jensen (University of Warwick)

Accepted for publication in the journal *Public Understanding of Science* in November 2015

Author Note

Benjamin K. Smith, Department of Communication, University of California, Santa Barbara; Eric Jensen, Department of Sociology, University of Warwick.

Correspondence concerning this article should be addressed to Eric Jensen, Department of Sociology, University of Warwick, Coventry CV4 7 AL, UK.

E-mail: e.jensen@warwick.ac.uk

We would like to thank representatives of the UK Department for Business, Innovation and Skills and Ipsos MORI for providing access to PAS 2014 data and answering questions we raised in a prompt and professional manner. We are also grateful to Professor Nick Allum (University of Essex) for his insightful feedback on an earlier draft of this manuscript.

Abstract

Since 2000, the UK government has funded surveys aimed at understanding the UK public's attitudes toward science, scientists and science policy. Known as the Public Attitudes to Science (PAS) series, these surveys and their predecessors have long been used in UK science communication policy, practice and scholarship as a source of authoritative knowledge about science-related attitudes and behaviors. Given their importance, and the significant public funding investment they represent, detailed academic scrutiny of the studies is needed. In this essay, we critically review the most recently published PAS survey (2014), assessing the robustness of its methods and claims. The review casts doubt on the quality of key elements of the PAS 2014 survey data and analysis, while highlighting the importance of robust quantitative social research methodology. Our analysis comparing the main sample and booster sample for young people demonstrates that quota sampling cannot be assumed equivalent to probability-based sampling techniques.

Keywords: public understanding of science, public opinion, survey methodology

Critical Review of the UK's "Gold Standard" Survey of Public Attitudes to Science

Since 2000, the UK government has funded a series of surveys – known as the Public Attitudes to Science (PAS) series - aimed at understanding the UK public's attitudes toward science, scientists and science policy. These surveys have long been used uncritically in UK science communication policy, practice and scholarship as a source of authoritative knowledge about science-related attitudes and behaviors. On the PAS 2014 blog, the survey is described by a science communication practitioner as, "rigorously conducted research that could inform… our own professional [science communication] practice" (Graphic Science, 2014, para. 14). Among a variety of uses of this research, the recent Wellcome Trust-commissioned review of UK informal science learning activity and impact (Falk et al., 2012) relied upon PAS 2011 data for key factual claims about public attendance levels at different informal science learning institutions.

The PAS surveys are intended to serve an important role in the evaluation and development of UK science policy. Minister of State for Universities and Science, Greg Clark, stated that the PAS studies are used to "make science and engineering policy decisions" (2014a, "Actions," para. 4), "to measure the success of government's science and society work and to identify areas that [the UK government] should work on in future" (Clark, 2014b, "Public Attitudes to Science survey," para. 2). Moreover, the 2014 PAS survey is framed as "one of the most robust studies of UK attitudes to science ever conducted…. Those who are interested in researching public attitudes can also draw lessons from how PAS 2014 was conducted" (Ipsos MORI, 2014a, p. 182). While two academic experts had input during the PAS 2014 research process as part of a larger 'steering group', the PAS studies have been published without undergoing careful academic review of their methodology or statistical inferences. Given the

importance of these surveys, and the significant investment of public funding they represent,[1] the lack of detailed, transparent academic scrutiny of its methods and claims to date is particularly problematic. In this essay, we review the most recently published PAS survey (2014), assessing its methodology and claims. Using the PAS 2014 data, we also test a key assumption underlying the study's temporal claims about change in public attitudes over time, namely that quota and probability sampling are comparable (e.g., Smith, 2008). Ultimately, this review is intended to identify which aspects of the influential PAS 2014 survey results should be accepted or rejected given established principles of good practice in quantitative social research. Our review begins with a brief history of science attitudes research. We then summarize the PAS 2014 survey's research objectives and methodology, before embarking on a critique of its reporting standards and methods.

Studying Public Attitudes toward Science

While the PAS surveys only began in 2000, the guiding influences underpinning their objectives and methods date back much further. According to Lewenstein (1992), the public's understanding of science became a focus of the scientific community around the start of the 19th century, when scientific institutions began constructing a public "vision of a rational world ruled by science" (1992, p. 46). These efforts developed further in the early 20th century, notably with the formation of the Science Service in 1920 and the National Association of Science Writers (NASW) in 1934. These organizations operated under the assumption that the public had a deficit of science knowledge: They aimed to foster appreciation for science and "rejection of superstitious beliefs" (Bauer, 2009, p. 223). In 1957, the NASW sponsored the first major national study of the public's understanding of science.

The National Science Foundation began to publish the biennial *Science Indicators* studies in 1972, based largely on the 1957 NASW survey. In 1979, "the *Science Indicators* studies significantly expanded the scope of the surveys and began to focus more attention on attitudes, knowledge measures, and expected participation measures for specific issues and controversies" (J. D. Miller, 1992, p. 24). These US-based surveys, along with the aims of their sponsoring organizations, greatly informed creation of the first UK-centered survey of the public's attitudes to science: *Public Understanding of Science, 1988*. This study of a random sample of 2,009 UK adults was heavily influenced by the 1957 NASW survey and the *Science Indicators* surveys, as well as the biases and assumptions that go with them (Stocklmayer & Bryant, 2012).

The Public Understanding of Science Paradigm and the Attitudes Deficit Model

A 1985 Royal Society report entitled *The Public Understanding of Science* symbolizes a paradigmatic re-orientation in the UK toward a public understanding (or attitudes) perspective (Bauer, 2009). S. Miller (2001) describes the shifting view of science in the UK media during the lead up to the 1985 report as a series of "mood swings" which caused "a tendency for scientists to retreat into their shells." In this context, the report, "reflected a concern amongst the scientific establishment that the retreat had reached such proportions that it made funding for scientific research politically vulnerable" (S. Miller, 2001, p. 115). If the scientific community was to maintain its social status (and funding), it needed to convince the public of its value. That is, they needed to erase the public attitudes deficit (cf. Jensen & Wagoner, 2009; Jensen & Holliman, 2015). The 'public understanding of science' model's "recurrent elements" included "a concern at the 'scientific ignorance' of the populace, a consequent desire to create a 'better informed' citizenry and an enthusiasm for making science 'more accessible'" (Irwin, 1995, p. 10).

Linked to these historical developments, the Public Attitudes to Science survey was designed explicitly to enable tracking of UK science policy's effectiveness. The PAS 2000 report states, "This research was designed to inform science communication policy and practice in Britain and to provide a rigorous baseline of public attitudes to science. Future changes in attitude can be tracked by repeating the study at regular intervals" (Office of Science and Technology & Wellcome Trust, 2000, p. 18). The research has now been repeated four times (in 2000, 2008, 2011 and 2014).

Many questions and methods in the PAS surveys can be traced back to the earlier US surveys. However, the PAS surveys did introduce a number of new measures aimed at evaluating trust in science and attitudes relating to science policy. These new dimensions are particularly evident in the last two iterations of the survey (2011 and 2014).

The Public Attitudes to Science 2014 Survey

The report for 2014's iteration of the Public Attitudes to Science survey details almost a dozen objectives, with results split into 13 chapters. These objectives link back to the historical roots of the 2014 survey (as outlined above), and heavily influence many of the research decisions at the heart of our critique. In order to simplify our discussion, we have identified four meta-objectives:

1. Detail the 'state of public attitudes' toward science, including the current state of scientific literacy, and public attitudes toward science, engineering and current scientific 'issues'.
2. Determine if and how attitudes toward science have evolved over time, by comparing the 2014 survey results to the other PAS surveys, and the 1988 and 1996 Public Understanding of Science surveys.

3. Describe the common interactions between the public and science, and how these interactions influence science attitudes and science literacy.
4. Analyze demographic differences in science attitudes, literacy and engagement, and re-examine the attitudinal segmentations from prior PAS surveys.

In order to address these main objectives, the report details the results of, "a representative survey of 1,749 UK adults aged 16+ and a booster survey of 315 16-24 year-olds" (Ipsos MORI, 2014a, p. 1), along with four qualitative research strands "intended to provide further context for the survey findings" (2014a, p. 20). The two surveys are the focus of this critique.

All PAS 2014 research was conducted by Ipsos MORI, a leading UK market research and public opinion polling company. Ipsos MORI used "random probability sampling methodology" (2014a, p. 2) to draw the main sample for the 2014 survey, in contrast to prior PAS studies that have historically used non-probability quota samples. Probability sampling is a technical term in statistics, which refers to drawing a sub-set of the population for your study in which each member of the overall population (in this case, UK adults) has a known, non-zero probability of inclusion. However, Ipsos MORI used a booster survey of 16-24 year olds for the 2014 study, which retained the non-probability quota sampling method from prior PAS surveys. At any point where young adults are discussed within the PAS 2014 report, the results are based on combining the 315 16-24 year old participants in the booster survey with the 195 in the main survey (Ipsos MORI, 2014a, p. 15).

The PAS 2014 report finds that the UK public is "overwhelmingly positive about the contribution science makes to the UK economy" and "to their own lives" (Ipsos MORI, 2014a, p. 1). For example, 81% agreed, "science will make people's lives easier" (p. 1). The report indicates that the UK public is both interested in understanding science better and in being better

understood by scientists: 55% indicated feeling uninformed about science, and 69% agreed that scientists "should listen more to what ordinary people think" (p. 3).

It is claimed that 'generic' public trust in scientists and engineers has increased significantly since the PAS 2011 survey, but "the proportion who feel they have no option but to trust those governing science has also increased from 60% to 67%, which suggests this increasing trust may also be an increasingly resigned trust, presenting a challenge for those looking to engage the public in decision-making" (Ipsos MORI, 2014a, p. 4). Social class is also reported to be a significant factor: the less affluent "tend to feel less well informed about science and are less likely to feel they know what scientists do" (Ipsos MORI, 2014a, p. 6).

It is reported that the UK public's attitude towards science has shifted in a positive direction over the long-term, with a 10% increase in agreement with the idea that the benefits of science outweigh the harmful effects (from 45% in 1988 to 55% agreement in 2014). The report also finds greater agreement with the positive view that it is important to know about science in people's daily lives (from 57% in 1988 to 72% agreement in 2014). The report also identifies much lower levels of agreement with the negative attitude that science makes people's lives change too fast (from 49% in 1988 to 34% agreement in 2014). Having summarized some of PAS 2014's major reported findings, we now critically evaluate whether these findings should be considered robust in light of the survey's sampling and data analysis methods.

Macro-Level Critical Review of the PAS 2014 Survey and Report

The PAS surveys have been highly influential in the UK policy context despite their history of employing low quality sampling methods. The use of non-probability quota sampling in past PAS studies means that they do not have a theoretical basis for making population-level claims about UK public attitudes (Baker et al., 2013). Prior to 2014, the high risk of sampling

bias inherent in the quota sampling method (Guignard, Wilquin, Richard, & Beck, 2013) was the most obvious limitation casting doubt on the validity of PAS-based statistics. It was then welcome news when it was announced that PAS 2014 would be "moving to a random sampling approach" (Silman, 2013).

In a blog post for the British Science Association, "PAS 2014: Main Report" co-author Silman wrote that the decision to move away from quota sampling was intended "to bring the PAS in line with other respected public opinion surveys on science" (Silman, 2013, para. 4). In fact, the report claims that the move to random sampling makes PAS 2014, "one of the most robust studies of UK attitudes to science ever conducted" (Ipsos MORI, 2014a, p. 182). While the move towards probability sampling did greatly increase the external validity and robustness of the research, this was undermined by a lingering reliance on quota sampling to reach youth populations and to make temporal claims. In addition, the study suffers from a number of unacknowledged methodological and conceptual limitations. We present a critical review of PAS 2014, with a focus on the survey portions of the study. The critique is split into three portions: (a) reporting standards, (b) sampling and temporal claims and (c) data handling / collection. While hardly exhaustive, our aim is to shed light on a number of the unstated / understated limitations of the survey and it's reported findings.

Limitations in Reporting Standards

Ipsos MORI reports that the 2014 PAS surveys were conducted in accordance with the international quality standard for Market Research, ISO 20252:2012. In addition, reporting for these surveys appears to meet the highest standards for disclosure set by the National Council on Public Polls. However, both of these standards are designed for market research, and fall well

short of what we, and a number of professional academic associations, consider the minimum standards for social scientific research.

As documented in our brief history of public attitudes to science studies, the main driving force for studies like PAS 2014 is to empirically assess whether science engagement interventions funded by the UK government and others are having a demonstrable effect on the general public. According to the UK government's website, the results of the PAS surveys "are used to measure the success of government's science and society work and to identify areas that we should work on in future" (Clark, 2014b, para. 6). The results are used to make 'evidence-based' decisions on best practices for science interventions and to determine funding. This trend toward evidence-based policy and practice is not unique to the study of science attitudes, with similar trends in medical research, education, psychology etc.

Evidence-based practices require that decision makers understand in full "the strengths and limitations of evidence obtained from different types of research" (APA Presidential Task Force on Evidence-Based Practice, 2006, p. 275). This requires a higher standard of reporting than was provided in the present case. Reports like PAS 2014 must walk a fine line between being accessible to a lay audience while also providing adequate detail to the reader in order for them to judge the quality of the data upon which the report is based. However, we contend that the PAS 2014 report (and many reports like it) errs too far on the side of accessibility, failing to provide critical information – even in the Technical Report and Appendices – needed to judge the robustness of the findings. To facilitate our critique, we compare the way results are reported for the 2014 survey against the standards set by the American Psychological Association (APA), as documented in the APA's most recent publication manual (2009). It is *not* our contention that the report produced by Ipsos MORI should have fully adhered to the APA standards (or any other

academic standard for that matter). Rather, the APA's publication manual offers useful and familiar guidelines for identifying possible areas for improvement.

On a number of marks, PAS 2014 meets or exceeds the guidelines proposed by the APA concerning reporting of methods and results. For example, the APA manual's suggestion of making raw data available on supplemental online archives has been amply addressed. In other instances, the PAS 2014 report does not meet APA standards, but for easily justifiable reasons. For instance, the APA manual emphasizes that information "…such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results" (2009, p. 33). Ipsos MORI comes as close to meeting this requirement as can be reasonably expected – by providing an appendix to the main report on statistical reliability that included example confidence intervals, and by providing rich explanations for most results. Had the report's authors provided APA style confidence intervals and traditional effect size measures for every result mentioned in the Main Report, the product would have been significantly longer and significantly less intelligible for a lay reader.

While there are many ways in which the standards used in reporting PAS 2014 survey results are satisfactory, in other areas meeting an academic standard would have provided additional clarity, helping to avoid leading readers towards specious conclusions. As it concerns the reporting of statistical and data analysis, the APA summarizes the guiding principles underlying all academic research, stating that: "accurate, unbiased, complete, and insightful reporting of the analytic treatment of data (be it quantitative or qualitative) must be a component of all research reports" (American Psychological Association, 2009, p. 33). In failing to fully uphold this standard, the overall quality of the PAS 2014 reports has been undermined, as we show below.

One of the 2014 survey's major objectives, as described earlier, was to analyze differences in science attitudes, literacy and engagement. To determine which differences to discuss in the PAS 2014 Main Report, the report's authors used inferential statistical tests, stating that: "Throughout this report, only differences that are statistically significant at the 95% level of confidence are commented on" (Ipsos MORI, 2014a, p. 16). They expanded on this in the Technical Report, describing how they determined whether something was statistically significant as follows: "The statistical test used for this is a two-tailed t-test" (Ipsos MORI, 2014b, p. 16). This thin level of detail may meet the market research industry standard, but it is inherently problematic from the perspective of social research methodology, as exemplified by the APA publication manual (2009).

At a "minimum," the APA expects researchers to completely report the results of *all* "tested hypotheses" (p. 33), both those that are 'significant' and those that are not. In reference to inferential statistical tests (e.g., t, F, and χ^2 tests), the APA also emphasizes the importance of including "the obtained magnitude or value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme as or more extreme than the one obtained (the exact p value),[2] and the size and direction of the effect" (p. 34). In addition, the APA argues it is "critical" to report the "frequency or percentages of missing data," provide "empirical evidence and/or theoretical arguments" for why data are missing, and "describe the methods for addressing missing data" (p. 33). Nowhere are these types of information, nor comparable information, provided. This information is not only missing from the PAS 2014 Main Report, but is also missing from the Technical Report; it is not provided in any appendix, nor is it provided separately on their website. Instead, the authors make a blanket statement that all their statistical findings should be trusted, without providing any supporting methodological evidence, and while

stating that they used a single (and in many instances inappropriate) method for conducting all null hypothesis statistical testing.

We are especially concerned with two implications stemming from the statement that "only differences that are statistically significant at the 95% level of confidence are commented on" (Ipsos MORI, 2014a, p. 16). First, where no other information is provided in the Main Report on statistical inference, and with little additional detail provided in the Technical Report, the reader is left to assume that the authors of the report relied upon appropriate statistical methods for determining what was and was not 'significant,' and that the results are accurate. Here we provide an example to show this assumption of good statistical practice is inaccurate. When discussing perceived risks and benefits of GM crops, the Main Report states:

> Those who feel well informed about GM crops are more likely to mention each of the perceived benefits in Figure 12.5[3] than those who do not feel informed about the technology. (Ipsos MORI, 2014a, p. 156)

There are at least three issues with just this one example statment presenting PAS findings. First, if the authors used *t*-tests to analyze these differences, as the Technical Report indicates, then they were in error. The nature of the data meant that the dependent variables could not have been normally distributed (as they are not continuous variables), and as such, the assumptions of the independent samples *t*-test would have been invalidated, leading to an increased risk of both Type I and Type II errors. Second, it is empirically incorrect to say that those who felt well informed were "more likely to mention each of the perceived benefits." As shown in Table 1, only 1 of the 7 reported benefits were statistically more likely to be mentioned by those who felt well informed.[4] Finally, this statment shows that the Main Report's blanket assertion that only statistically significant differences are commented on is not accurate: The reported difference in the likelihood of mentioning "destruction of natural crop species (18% versus 12%)" is *not*

significant at the 95% level of confidence, χ^2 (1, n = 432) = 2.91, exact p = .108, OR = 1.58, 95% CI [.93, 2.69].

Table 1

Tests of differences in benefits & risks mentioned by perceived level of knowledge about genetically modified crops.

Item	% who mentioned						*OR* 95% CI	
	Not Informed	Informed	χ^2 (1)	p^a	p^b	*OR*	Lower	Upper
Benefits								
Health benefits	10%	7%	1.32	.251	.297	.67	.33	1.34
Increased food production	46%	55%	3.02	.082	.084	1.40	.96	2.04
Make crops more consistent	16%	20%	1.02	.313	.322	1.29	.79	2.10
Make food tastier/better quality	6%	8%	0.64	.423	.460	1.35	.65	2.81
More disease resistant	16%	27%	7.64	.006	.007	1.92	1.20	3.06
More predictable harvests	10%	15%	3.11	.078	.082	1.68	.94	2.99
Allow crops in adverse climate	15%	17%	0.24	.626	.694	1.14	.68	1.90
Don't know	20%	11%	6.74	.009	.011	.49	.28	.84
Nothing/no benefits	8%	9%	0.06	.800	.866	1.09	.56	2.12
Risks								
non-GM crops cross pollination	8%	19%	10.82	.001	.001	2.60	1.45	4.65
Destroying natural crop species	12%	18%	2.91	.088	.108	1.58	.93	2.69
Disrupts ecosystem/wildlife	16%	17%	0.05	.816	.897	1.06	.64	1.77
Don't understand effects	23%	22%	0.02	.891	.909	.97	.62	1.52
Not natural	6%	10%	1.72	.190	.213	1.60	.79	3.26
Not properly tested	8%	8%	0.08	.774	.861	.90	.45	1.81
Negative impact on health	30%	22%	3.74	.053	.062	.65	.42	1.01
Don't know	23%	15%	3.61	.057	.067	.62	.38	1.02
Nothing/no risks	4%	7%	1.08	.299	.401	1.55	.67	3.57

Note: $n \approx 432$. Sampling weights, provided by Ipsos MORI in the main survey data file, were applied prior to running analysis, with cell counts rounded to the nearest interger prior to running χ^2 tests, calculating odds ratio, and determining percentage values. This method was chosen because it best corresponds with the proportions reported in the Main Report. No substantive differences were noted when using truncated cell counts, or when making no adjustments.
[a]Two-tailed asymptotic *p*-value.
[b]Two-tailed exact *p*-value.

Our other major concern is that, where no other information is provided either in the Main Report or elsewhere on how Ipsos MORI determined which 'differences' to comment on in their report, readers are left to assume that any unreported differences are *not* statistically significant because of the Report's blanket statement that only significant differences are included. Again, this does not seem to be the case. For example, despite an entire chapter of the

Main Report being dedicated to differences in public perceptions of engineering and science, and despite the numerous comparisons between men and women throughout the report, nowhere is it mentioned that women were significantly less likely to spontaneously mention 'Engineering' when asked "when I talk about 'science', what comes to mind?" then were men, χ^2 (1, n = 1749) = 15.91, $p < .001$, OR = 0.36, 95% CI [0.21, 0.61].

Both of the examples provided above highlight the risk of underreporting or misreporting of results, which is exacerbated by reliance on thin market research industry standards for statistical reporting. We are highlighting these examples because the misapplication and underreporting of statistics could have resulted in both false positives and false negatives, and because the failure to employ a more robust reporting procedure leaves readers without any straightforward way of determining whether the reported results are accurate, unbiased, and complete. If a more rigorous reporting standard was required for the PAS survey research, these methodological issues could be clarified much more straightforwardly, without requiring the kind of fresh data analysis we have presented above. Ultimately, we have shown here that there is a real risk that readers are being presented with inaccurate or incomplete conclusions.

Sampling and Temporal Claims

In this section we take issue with the common market research industry claim that quota-based sampling methodology is just as good at accurately assessing public attitudes as probability sampling (e.g., Smith, 2008). By implication, we are raising questions about the claims of PAS-style surveys that used quota sampling and comparisons to those surveys in the PAS 2014 results. While the PAS 2014 survey moved away from the prior surveys' use of problematic quota-based sampling methods, its goal of comparing current science attitudes to past attitudes necessitated comparisons between 2014 results and previous PAS surveys. This

means the PAS 2014 study's claims about change over time in science attitudes were based on data obtained using two different techniques. This should raise concern in light of the limitations of non-probability sampling methods. For example, in summarizing their latest report on non-probability sampling, the American Association for Public Opinion Research (AAPOR) states:

> AAPOR cautions that collecting data and producing estimates in the absence of a sound theoretical basis is inappropriate for making statistical inferences. Probability samples enjoy an underlying theory and set of assumptions that are widely known and accepted. This is generally not the case with non-probability methods. (Task Force on Non-Probability Sampling, 2013, para. 5)

As it is impossible to know the extent to which prior quota-based surveys accurately reflect the population's characteristics,[5] we contend that temporal claims about change over time in population attitudes based on these surveys are inherently problematic.

In addressing concerns about quota sampling, the PAS 2014 Main Report directs readers to a blog post by a report author, Tim Silman. Here, Silman explains the choice to move to the self-described "gold standard" of sampling, but also defends quota sampling: "Studies have [shown] that good quota samples can be comparable to random samples, so we can still make meaningful comparisons with previous waves done on quota" (Silman, 2013, para. 5). This statement in turn links to an essay written by Patten Smith, then Director of Survey Methods (now Director of Research Methods) for Ipsos MORI (2008).

The main thrust of Smith's argument in defense of quota sampling is that "data from quota and random probability samples are, in the main, comparable" (2008, p. 4). More precisely, Smith avers that there is empirical evidence based on years of successful research showing that quota samples are "reasonably accurate." To support this claim, he cites two studies, "suggesting that the numbers of significant differences between probability sample results are in-line with chance expectation" (Smith, 2008, p. 5). However, our search of the academic methodological literature revealed findings that emphatically conflict with Smith's

claims (e.g. Guignard et al., 2013), including one of the two articles cited by Smith, which says that quota samples carry "a greater risk of bias from the exclusion of people who are hard to find or interview" (Stephenson, 1979, pp. 494–495).

A key concern within social science methodology about using quota sampling is that the risk of error is entirely unknown. Thus, in any one study, a quota sample may resemble a probability sample, but it is impossible to know whether this is the case *a priori*. The question of whether a UK population survey using probability sampling is substantively equivalent to one using quota sampling is in itself an empirical question, which has been addressed by previous research (e.g., Guignard et al., 2013; Stephenson, 1979). This question can be answered here by conducting secondary analysis of the raw PAS 2014 data provided by Ipsos MORI. As mentioned earlier, the 2014 survey includes two parts: the main survey (which used probability sampling) and a booster survey of 16-24 year olds (which used quota sampling). This provides a perfect opportunity to test the assumptions put forth by Smith, Silman, and the PAS 2014 report.

Quota vs. Probability Sampling: Methodology

Our secondary analysis aims to determine whether the main sample and the booster sample were statistically equivalent, and therefore whether combining the two samples (as Ipsos MORI did) was appropriate. To test this, we used the Hou (2005) method for combining weighted *p*-values. A "*p*-value is the probability of obtaining a value of the test statistic at least as extreme as the one that was actually observed, given that the null hypothesis is true" (Fang & Wit, 2008, p. 435). For every variable collected in the PAS 2014 survey, we tested the null-hypothesis that the distribution of responses between the probability sampling-based main survey and the quota sampling-based booster survey were the same.[6] Doing this created a vector of *p*-values that all shared an underlying null-hypothesis. When a set of statistical tests share a similar

null-hypothesis, it is possible to combine those *p*-values into a single test statistic, which can then be used to determine the probability of the underlying null-hypothesis being true (Dai, Leeder, & Cui, 2014).

To determine the probability that the *p*-values were uniformly distributed, and thus that the two-samples were equivalent, we used a simplified version of Good's method (1955), as proposed by Hou (2005). This procedure involves creating a weighted Fisher's statistic, which is then evaluated using a scaled Chi-Square distribution. The probability result derived from this method can then be used to determine whether the following null-hypothesis should be rejected:

H_0 = People sampled using the quota method have the same distribution of responses to the questions in the PAS 2014 survey as people sampled by the probability method.

Under this null, the alternative hypothesis is:

H_1 = The null-hypothesis is false in the case of at least one variable, that is, people sampled using the quota method *do not* have the same distribution of responses to the questions in the PAS 2014 survey as people sampled by the probability method.

To keep the size of this section from being excessively large, we have moved the rest of the methodological details for this analysis into an Appendix, placed online as supplementary material. There, you will find a description of our data selection and preparation procedures, along with a detailed description of how we analyzed the dataset, and arrived at the final test results.

Results

We compared 16-24 year olds in the main sample with those in the booster sample across 488 different variables. We performed our initial comparisons twice, first without applying the sampling weights provided by Ipsos MORI, and then with those weights applied.

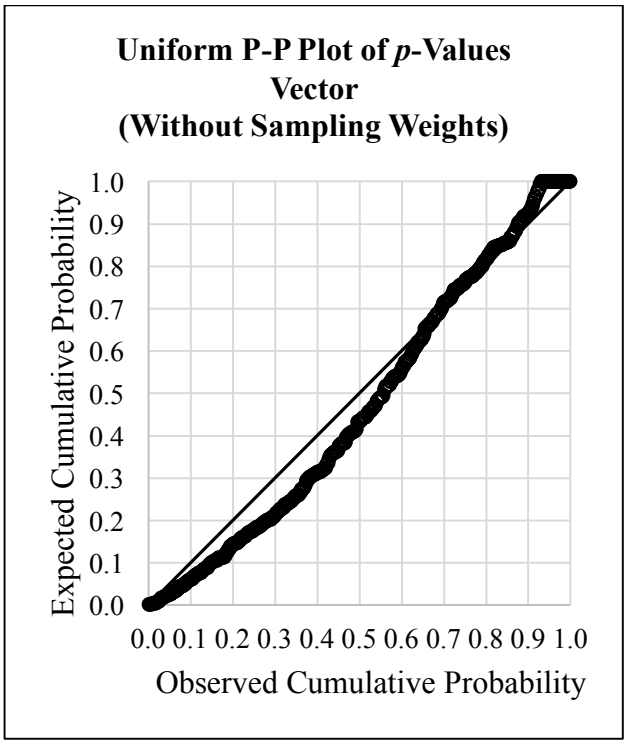

Figure 1. Uniform P-P Plot showing the cumulative proportions of the observed *p*-values vector (*without* sampling weights) compared against the cumulative proportions expected under the null hypothesis (i.e., a uniform distribution, see: Fisher, 1932; Stephenson, 1979). For this plot, the survey data were *not weighted* prior to running the appropriate statistics. The *p*-values for interval/ordinal level data were obtained using Mann-Whitney *U*-tests; *p*-values for categorical/nominal data were obtained using exact probability χ^2 tests, with cell counts rounded to the nearest integer.

Figure 1 shows the distribution of *p*-values for the first set of tests (i.e. without sampling weights); Figure 2 shows the same for the second set of test (i.e. applying sampling weights). In both cases, there appears to be a significant break from the expected uniform distribution, with the variation especially pronounced in Figure 2.

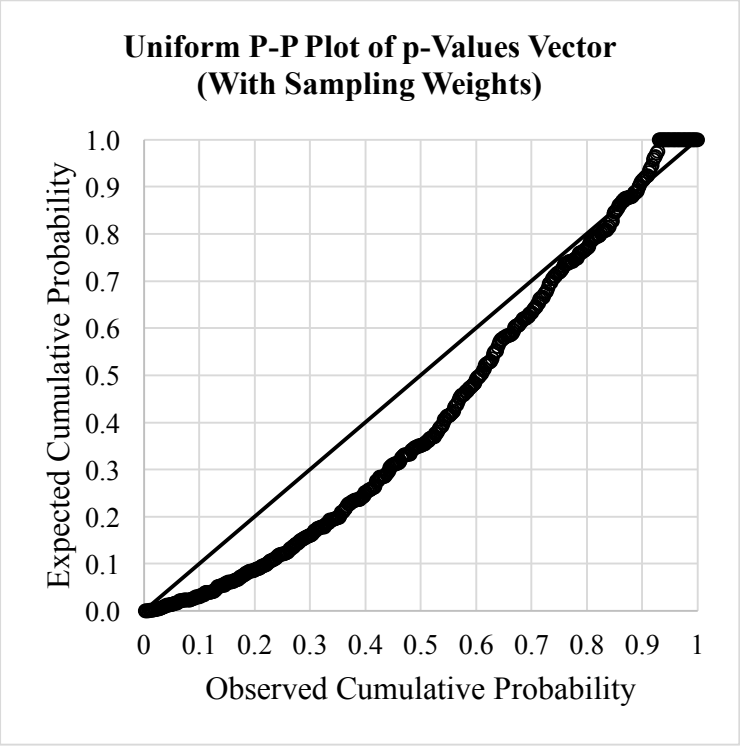

Figure 2. Uniform P-P Plot showing the cumulative proportions of the observed *p*-values vector (*with* sampling weights) compared against the cumulative proportions expected under the null hypothesis (i.e., a uniform distribution, see: Fisher, 1932; Stephenson, 1979). For this plot, the survey data *were weighted* (using the sampling weights provided by Ipsos MORI) prior to running the appropriate statistics. The *p*-values for interval/ordinal level data were obtained using Mann-Whitney *U*-tests; *p*-values for categorical/nominal data were obtained using exact probability χ^2 tests, with cell counts rounded to the nearest integer.

We then conducted a set of four combined probability tests. The differences in these four tests were defined by the use of weights, first whether the sampling weights provided by Ipsos MORI were applied when conducting the original analysis and then whether the test weights[7] were applied. The results of these four tests are found in Table 2.

Table 2
Combined probability tests for comparisons of sampling methodology

Tests	X	*c*	χ^2	*df*	*p*
Without Sampling Weights					
With Variant Test Weights[a]	2.279	.00291	783.907	688.006	.006
With Equal Test Weights[b]	2.387	.00205	1164.725	976	< .001
With Sampling Weights[c]					
With Variant Test Weights[a]	2.736	.00296	925.354	676.368	< .001
With Equal Test Weights[b]	2.911	.00205	1420.614	976	< .001

Note: X = weighted Fisher's combination statistic; *c* = sum of the squared weights; χ^2 = the evaluated chi-square statistic derived from X/c (see Appendix); *df* = degrees of freedom for the chi-square test statistic; *p* = approximate combined probability of the null-hypothesis being true, i.e. that there are no differences between the two samples.
[a]Weight for each test's *p*-value = SD^{-1}, i.e., the inverse of the squared error variance.
[b]Weight for each test's *p*-value = k^{-1}, i.e., the inverse of the number of *p*-values being combined.
[c]These *p*-values were obtained using the sampling weights provided by Ipsos MORI.

In all four instances, the combined *p* value is less than .01. As such, the null hypothesis is rejected, and we can infer that there is a significant difference between the booster sample (which used a quota-based sampling methodology) and the equivalent part of the main sample (which used a probability-based sampling methodology).

Post-hoc analysis. Following the steps of Stephenson (1979), we also explored the 'substantive' demographic differences between the booster sample and the main sample. In line with his analysis, these comparisons were conducted using the sampling weights provided by Ipsos MORI. As shown in Table 3, we found many of the same things as Stephenson, including substantive differences (i.e., $\varphi_c > .1$) – albeit not always 'significant' differences – in household composition, working status (especially among males), and geographic location. In addition, we

found a substantive difference in social grade, the probability that the respondent had "scientists / engineers among relatives / friends" and the probability that the respondent "works with scientists / engineers".

Table 3
Substantive demographic differences between the booster and main sample

Name	Description	*n*	χ^2	*df*	p^a	p^b	φ_c
Household Composition							
qh	# of Children aged 15 and under	505	4.18	4	.383	.351	.091
@qi1	(yes or no) 1 or more child aged 0-4	160	2.61	1	.106	.122	.128
@qi2	(yes or no) 1 or more child aged 5-7	160	3.51	1	.061	.086	.148
@qi3	(yes or no) 1 or more child aged 8-11	160	6.50	1	.011	.010	.201
@qi4	(yes or no) 1 or more child aged 11-15	160	0.37	1	.542	.621	.048
qc	Working Status	506	11.00	4	.027	.032	.147
qgcomb	Social Grade	510	13.80	4	.008	.009	.164
@qn13	"Scientists / engineers among relatives / friends"	510	11.88	1	.001	.001	.153
@qn15	"Works with scientists / engineers"	510	6.30	1	.012	.014	.111
Region	Respondents government region	510	17.67	11	.090	--[c]	.186

Note: φ_c = Cramer's phi (alternatively, Cramer's V). Sampling weights were applied prior to performing calculations; χ^2 calculated using rounded cell counts.
[a]Two-tailed asymptotic *p*-value.
[b]Two-tailed exact *p*-value.
[c]Exact *p*-values could not be calculated for this variable, due to the large number of cells.

Implications and Limitation

Before discussing the implications of this finding, a limitation must be addressed. This analysis does not account for dependence between the tests. A number of steps were taken to mitigate this problem, namely the use of a scaled χ^2 distribution, test weights, and conservative hypothesis tests, all of which have been shown to reduce error caused by dependency (e.g. Alves & Yu, 2014). Additionally, our method is comparable to other studies analyzing the differences between quota and probability samples (i.e., Stephenson, 1979), lending credibility to our results.

An additional limitation of this analysis is the inability to derive a meaningful effect size estimate across all variables. The goal of this analysis was simply to determine the probability that the distribution of the results obtained from the two sampling methods were different. While

the method we used did allow us to answer the question at hand, it did not answer the much larger and more detailed questions of why, how, and to what extent. We strongly encourage future researchers to do just that.

Despite the possible limitations in our analysis, the results clearly indicate that the quota-based sampling methods and the probability-based sampling methods used by Ipsos MORI did not provide equivalent results. Not only does this potentially undermine the reported results involving the 2014 booster survey of 16-24 year olds, but it also calls into question the numerous comparisons to non-probability based PAS surveys conducted in previous years. Given that the quota sampling-based results reported in previous years may be inaccurate to an extent that is impossible to determine (because no probability-based sampling was employed), there is no reason to believe that comparisons to the results of past quota-based sampling results are valid.

Limitations in Data Handling and Data Collection

We now shift the focus of our critique from the PAS 2014 results to its methodology. In general, we are impressed with the detailed reporting of methodology found in the Technical Report. We especially appreciated the detailed description of weighting, pre-testing and sampling. However, in a number of instances, there are serious questions regarding how data were collected and processed, which cast doubt on the quality of the PAS 2014 survey data.

We will not document here all of the potential issues with the data handling and data collection, as most of these issues are clearly stated in the technical report. We are far more concerned about the information that is *not* reported. For example, the lack of procedural transparency about data processing for the large number of open-ended questions is a concern because it could have directly affected the PAS 2014 results. Consider question 43, which asks: "And what would you say are the main risks, if any, of genetically modified crops" (Ipsos

MORI, 2014b, p. 52). For this question, field workers were instructed: “Do not prompt. Probe fully” (2014b, p. 52). This (and a number of similar questions) are clearly ‘open-ended questions’, yet the technical report unequivocally states: “There were no open-ended questions in the survey” (2014b, p. 17).

While this may seem like an inherent contradiction and simply a mistake, it is not. Instead, field workers were provided with a set of pre-specified possible responses, “based on what respondents gave as answers at the cognitive testing and pilot survey stages” (J. N. Shah, personal communication, June 30, 2014). If the field worker could find a way to place a response into one of the pre-identified categories, they were instructed to do so. The actual response was only recorded if there was no way to make the response fit. These responses were then coded by “members of Ipsos MORI’s coding department” (Ipsos MORI, 2014b, p. 17), using an unspecified methodology.

This procedure is in many ways methodologically questionable and makes it hard for an independent researcher to tell whether the results accurately represent what respondents said. The failure to check for inter-coder reliability among field workers, not recording most open-ended responses and failing to specify the coding decision rules that were used are all issues we see with this procedure for dealing with open-ended survey data. Most problematically, there is no way to verify the validity of the data presented by Ipsos MORI from any of these open-ended questions: Decision-making may have been completely idiosyncratic depending on subjective assessments by data collectors while administering surveys. In sum, the methodological procedure for processing qualitative survey data from all of the open-ended items in PAS 2014 makes it hard, if not impossible, to verify their reliability. We feel that this methodology falls far short of well-established minimum standards in social scientific research for quantifying such

data (e.g., Krippendorf, 2013; Neuendorf, 2002). Because key information was either underreported or left unreported, and because of issues with how open-ended data were handled and processed, it is hard to trust the results, making them potentially unsuitable for secondary analysis by social scientists that might wish to conduct a more sophisticated analysis of the PAS 2014 data.

Discussion

This essay has summarized and critically evaluated the 2014 Public Attitudes to Science survey. Clearly, changing sampling techniques to approximate a probability-based approach was a good decision by the PAS 2014 steering group. However, the importance of the PAS survey within UK science policy means that the results must hold up under a high level of scrutiny. Our review has revealed a number of methodological limitations that potentially undermine some of the headline claims about public attitudes towards science and scientists made in the PAS 2014 report. Specifically, we highlight that improvements are required in reporting practices to reduce the possibility of leading readers to infer false or overstated conclusions. We also demonstrate the limitations in the survey's continued reliance on quota sampling. We would recommend that the government focus its resource investment on ensuring systematic methods are used to reach young people, for example, oversampling using the same technique as the main survey. As long as the quantitative findings are important, it would be better to ask fewer questions to save cost, rather than sacrificing the quality of the sampling methods. Finally, we discuss limitations in the methods used by Ipsos MORI to conduct the 2014-survey data collection and analysis.

We believe there is a broader value in highlighting such technical issues because the methodological limitations we have found in the present study appear to be commonplace in UK government-funded public attitudes research, such as the Department for Climate Change's

'Public attitudes tracking survey' (https://www.gov.uk/government/collections/public-attitudes-tracking-survey). Thus, we would urge social scientists to check carefully (at least) the technical details flagged up in this essay prior to relying upon these and similar survey results in their own academic research.

Ultimately, some (if not most) of the PAS 2014 results emerge from our critique largely unscathed. The reported percentages of the UK population holding particular attitudes in 2013 (based on the closed-ended survey items) are among the findings most likely to be accurate. Likewise, comparisons between the closed-ended items in the PAS 2014 survey and the 1988 and 1996 Public Understanding of Science surveys (conducted with probability samples) are also likely to be valid. Therefore, we would encourage researchers considering using PAS data for secondary analysis to concentrate their efforts on these elements, while paying attention to the quality of survey question design when choosing which items to analyze.

At the same time, this review has shown that a number of key claims made based on PAS 2014 data are not adequately evidenced, due to potential methodological flaws or chronic under-reporting of essential details. For example, the failure to report *p*-values, test statistics, etc. in either the main report or the technical report forces the reader to simply trust that Ipsos MORI did everything right in their statistical analysis. However, this review suggests such blind trust is not fully warranted. Moreover, we argue that any temporal claims about changes in the prevalence of particular attitudes over time between PAS surveys cannot be considered robust given the failure to use probability sampling in prior iterations of the PAS survey. Likewise, we have demonstrated that PAS 2014-based claims about the prevalence of different attitudes amongst young people are not robust given the mixture of probability and non-probability sampling used to derive these figures. Our analysis comparing the main sample and booster

sample for young people reinforces the well-established social research principle that quota sampling cannot be assumed to be equivalent to probability-based sampling techniques. This also reinforces concern that the previous quota sampling-based PAS surveys in 2000, 2008 and 2011 may not provide accurate accounts of the prevalence of different science attitudes in the UK population.

While we have addressed a number of key issues in this review, there are further statistical tests and methodological reviews that would be beneficial for understanding the nature and extent of potential limitations in the PAS 2014 and previous PAS studies. It would be useful to conduct a thorough analysis to determine the extent to which the PAS 2014 survey results are suffering from false positive and false negative findings. We would also recommend that researchers considering using PAS 2014 data generated through the open-ended survey items undertake a more detailed assessment of the open-ended response options than we have been able to do here. Finally, a sensitivity analysis could be conducted to evaluate the degree to which the quota sampling used in previous iterations of the PAS studies may have affected the accuracy of their results.

In sum, higher quality research is still needed to fully address the worthwhile objectives of the Public Attitudes to Science survey. It is particularly concerning that this flagship government-commissioned survey by a leading market research and public opinion polling company with ample funding to use (what they describe as) ‘gold standard’ methods contains potentially problematic research practices. Policy, practice and academic communities should take account of the limitations we have identified when using results from this important survey. Finally, given the quality issues we highlight, government stakeholders should evaluate whether their tendering and procurement procedures are leading to the best use of government resources.

References

Alves, G., & Yu, Y.-K. (2014). Accuracy Evaluation of the Unified P-Value from Combining Correlated P-Values. *PLoS ONE*, *9*(3), e91225. http://doi.org/10.1371/journal.pone.0091225

American Psychological Association. (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285. http://doi.org/10.1037/003-066X.61.4.271

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., … Tourangeau, R. (2013). *Report of the AAPOR task force on non-probability sampling*. American Association for Public Opinion Research. Retrieved from http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Bauer, M. W. (2009). The Evolution of Public Understanding of Science—Discourse and Comparative Evidence. *Science Technology & Society*, *14*(2), 221–240. http://doi.org/10.1177/097172180901400202

Clark, G. (2014a, March 12). Engaging the public in science and engineering. Retrieved August 18, 2014, from https://www.gov.uk/government/policies/engaging-the-public-in-science-and-engineering--3

Clark, G. (2014b, March 12). Engaging the public in science and engineering - Making informed science policy decisions. Retrieved August 18, 2014, from

https://www.gov.uk/government/policies/engaging-the-public-in-science-and-engineering--3/supporting-pages/making-informed-science-policy-decisions

Dai, H. (Daisy), Leeder, J. S., & Cui, Y. (2014). A modified generalized Fisher method for combining probabilities from dependent tests. *Evolutionary and Population Genetics*, *5*, 32. http://doi.org/10.3389/fgene.2014.00032

Falk, J., Osborne, J., Dierking, L., Dawson, E., Wenger, M., & Wong, B. (2012). *Analysing the UK science education community: The contribution of informal providers*. London: Wellcome Trust. Retrieved from http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_peda/documents/web_document/wtp040860.pdf

Fang, Y., & Wit, E. (2008). Test the Overall Significance of p-values by Using Joint Tail Probability of Ordered p-values as Test Statistic. In C. Tang, C. X. Ling, X. Zhou, N. J. Cercone, & X. Li (Eds.), *Advanced Data Mining and Applications* (pp. 435–443). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-88192-6_41

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd.

Good, I. J. (1955). On the weighted combination of significance tests. *Journal of the Royal Statistical Society. Series B: Methodological*, *17*(2), 264–265.

Graphic Science. (2014, June 23). What can we learn from the Public Attitudes to Science 2014 survey? Retrieved from http://www.creststar.org/blog/what-can-we-learn-public-attitudes-science-2014-survey

Guignard, R., Wilquin, J.-L., Richard, J.-B., & Beck, F. (2013). Tobacco Smoking Surveillance: Is Quota Sampling an Efficient Tool for Monitoring National Trends? A Comparison with

a Random Cross-Sectional Survey. *PLoS ONE*, *8*(10), e78372. http://doi.org/10.1371/journal.pone.0078372

Hou, C.-D. (2005). A simple approximation for the distribution of the weighted combination of non-independent or independent probabilities. *Statistics & Probability Letters*, *73*(2), 179–187. http://doi.org/10.1016/j.spl.2004.11.028

Ipsos MORI. (2014a). *Public Attitudes to Science 2014: Main Report* (Version 2). Retrieved from https://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-main-report.pdf

Ipsos MORI. (2014b). *Public Attitudes to Science 2014: Technical Report* (Version 1.3). Retrieved from http://www.ipsos-mori.com/Assets/Docs/Polls/pas-2014-technical-report.pdf

Irwin, A. (1995). *Citizen Science: A Study of People, Expertise and Sustainable Development*. Psychology Press.

Jensen, E. & Holliman, R. (2015, published online before print). Norms and values in UK science engagement practice. *International Journal of Science Education – Part B: Communication and Public Engagement*.

Jensen, E. & Wagoner, B. (2009). 'A cyclical model of social change'. *Culture & Psychology, 15*(2), 217-228.

Krippendorf, K. (2013). *Content Analysis: An Introduction to Its Methodology* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Lewenstein, B. V. (1992). The meaning of `public understanding of science' in the United States after World War II. *Public Understanding of Science*, *1*(1), 45–68. http://doi.org/10.1088/0963-6625/1/1/009

Miller, J. D. (1992). Toward a scientific understanding of the public understanding of science and technology. *Public Understanding of Science*, *1*(1), 23–26. http://doi.org/10.1088/0963-6625/1/1/005

Miller, S. (2001). Public understanding of science at the crossroads. *Public Understanding of Science*, *10*(1), 115–120. http://doi.org/10.1088/0963-6625/10/1/308

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA: SAGE Publications Ltd.

Office of Science and Technology, & Wellcome Trust. (2000). *Science and the Public: A review of Science Communication and Public Attitudes to Science in Britain*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32583/wtd003419.pdf

Quigley, A., Pettigrew, N., & Shah, J. N. (2014, April 1). Public Attitudes to Science 2014. Retrieved October 1, 2014, from http://www.ipsos-mori.com/researchpublications/researcharchive/3357/Public-Attitudes-to-Science-2014.aspx

Silman, T. (2013, October 19). PAS 2014: adopting the gold standard. Retrieved from http://www.creststar.org/blog/pas-2014-adopting-gold-standard

Smith, P. (2008). *Is random probablity sampling really much better than quota sampling*. Unpublished Manuscript. Retrieved from http://ebookbrowsee.net/is-random-probability-sampling-really-much-better-than-quota-sampling-doc-d57750014

Stephenson, C. B. (1979). Probability Sampling with Quotas: An Experiment. *Public Opinion Quarterly*, *43*(4), 477–496.

Stocklmayer, S. M., & Bryant, C. (2012). Science and the Public—What should people know? *International Journal of Science Education, Part B*, *2*(1), 81–101. http://doi.org/10.1080/09500693.2010.543186

Task Force on Non-Probability Sampling. (2013). *Executive Summary*. American Association for Public Opinion Research. Retrieved from http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/FINALLayman_TaskforceonNonprobabilitySampling07-21-13_withLOGO.pdf

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, *18*(5), 1368–1373. http://doi.org/10.1111/j.1420-9101.2005.00917.x

Footnotes

[1] The PAS 2014 contract was for £500,000 / approximately $771,000 US dollars (as of September 2015). While this included qualitative research and other elements not covered in this article, this is nevertheless a substantial research budget for one project, which was intended to deliver a very high quality survey study.

[2] It is worth clarifying that the APA is referring here to the exact *value* of the *p*-value statistic (e.g., reporting $p = .034$ rather than $p < .05$), not the use of 'exact' *p*-value statistics (like those we use throughout this article).

[3] The values in the referenced figure are derived from responses to the question "…what would you say are the main benefits, if any, of genetically modified (GM) crops?"

[4] Even if ignoring statistical significance the statement is wrong, as those who did not feel well informed were nominally a third more likely to mention health benefits than those who felt well informed.

[5] Probability samples also suffer from limitations. For example, many social surveys get response rates below 30%, which makes these samples highly problematic in principle as well. However, it is much more feasible to account for the level of error in probability samples and ensure that error risk is fully disclosed.

[6] To be more precise (and as noted in the Appendix), we compare the distribution of responses from 16-24 year olds in the main survey with those in the booster survey, using a large series of Exact Pearson Chi-Square tests (for the categorical variables in the sample) and Mann-Whitney U tests (for the continuous, or nearly continuous, variables).

[7] Test weights were equal to 1 / *SD,* i.e. the inverse of the standard error for each variable. See the Appendix for more details.

Appendix

In order to determine whether the two samples were equivalent, we began by comparing the responses from 16-24 year olds in the main survey with those in the booster survey, using a large series of Exact Pearson Chi-Square tests (for the categorical variables in the sample) and Mann-Whitney *U* tests (for the continuous, or nearly continuous, variables). These statistical tests are both non-parametric, which means they are suitable for use on non-probability samples, and for the types of questions asked in the PAS survey. In addition, both of these tests tend to be highly conservative, implying that the combination of the *p*-values from these tests will be somewhat biased toward failing to reject the null-hypothesis. If the null-hypothesis is true and both the quota and probability samples yield similar results, then the distribution of *p*-values should follow a uniform distribution (Fisher, 1932; Stephenson, 1979).

Data Preparation

We derived all data for this analysis from the main and booster survey combined SPSS file, found on the Ipsos MORI website (Quigley, Pettigrew, & Shah, 2014). Given the intentionally raw nature of the data, it was necessary for us to take a number of steps to prepare the data for statistical analysis. The first step was to create a new variable delineating which of the two samples each participant was from, based upon his or her serial number.

The second step in our data preparation was to correct the Likert-type items in the data set, so they could be analyzed using traditional statistical approaches. For these variables, the responses include a “don’t know” option appended to each scale. This keeps the variables from being continuous and/or ordinal. In most cases, we simply excluded “don’t know” responses, which effectively corrected for the issue. However, we treated the two sets of questions differently, based on the unusually high proportions of respondents marking “don’t know.”

Question set 8 asked participants about the perceived risks or benefits for a set of controversial topics. For a number of these questions, respondents were unwilling to side either way, presumably because of their low level of information on the corresponding topic. This resulted in an unusually high number of "don't know" responses: between 3.9% and 19%. Rather than simply ignore these responses, we split "don't know" responses off into their own variables, following the same coding scheme as the other binary response variables in the data set (i.e. "don't know" = 1, "not, don't know" = 0).

Question 9 asked participants how confident they were that scientists in the UK consider risks. This question included response options for both "don't know" and "it depends on the area they work in". Together, these two response options accounted for over 6% of responses. Similar to what we did for question set 8, we split that pair of response options off into a new variable, which again followed the same coding scheme as the other binary response variables (i.e. "don't know and it depends" = 1, "no, don't know and it depends" = 0). While the two response options are qualitatively (and we would argue substantively) different from each other, our main goal was simply to capture and test the distribution of responses scross answers in the most fair and appropriate manner possible. We felt it was necessary to account for these responses, given the high frequency with which they were selected, but splitting the two non-ordinal response options into two response options (rather than one) could have biased the combined test in favor of rejecting the null hypothesis. As such, we determined it was better to err on the side of a conservative (rather than liberal) coding sheme.

Variable Selection

To determine which variables in the data set were eligible for comparison, we looked at three criterion. First, for all binary data, we looked at the number of respondents in the variable

who responded with the listed option. For example, question 1 asked respondents, “When I talk about ‘science’, what comes to mind?” Variable @q1@q7, or ‘question 1 option 7’, captures those who replied with some version of “Boring/dull.” Eight respondents gave that response. If the raw number of participants who responded with a certain listed option was less than 10 the variable was not included in our analysis (variables excluded = 337). 10 was chosen as our cutoff based on the common social science rule of thumb that there should be at least an expected count of 5 in each cell for a χ^2 test (10 responses split between the two conditions roughly equates to an expected count of 5, depending on the number of respondents in each condition). While the method we used was actually capable of detecting ‘real’ differences even with very small cell sizes, we found that including these variables drastically biased the combined test toward rejecting the null hypothesis. As such, we felt it would be more fair to use a conservative inclusion criteria.

Second, we looked at variables that were derivatives of other variables. For example, the data set included five variables that looked at types of newspapers read by each participant. These variables combined responses from ‘question q’ and ‘question p’ together into summary measures. In this instance, as is the case with most derivative measures, we chose to exclude the summary measures, and instead analyzed the original variables (variables excluded = 19). However, for education, social grade, religion, ethnicity and region we used the derivative measures instead, because the original measures were overly specific, including too many cells for accurate staticial testing. For example, on ethnicity, there were 19 different response options, 9 of which had 5 or fewer respondents. Finally, variables were not included if they simply captured meta-data (i.e. serial number, weights, module, and survey taken).

The data set began with 836 listed variables. As described above, we added 12 variables (10 for ‘question 8’ about risks/benefits of controversial science topics, 1 for ‘question 9’ about confidence scientists take account of risks and a variable for ‘survey taken’) to the data set, bringing the total to 848. Based on the criterion above, 360 variables were determined to be ineligible for inclusion as dependent variables in the data analysis. This brought our total number of variables analyzed to N = 488.

Data Analysis

The first step in our data analysis was to compare the booster survey responses and the main survey responses across the 488 variables. All of these tests were conducted using IBM SPSS 22 with the Exact Tests extension. The analysis was split into two groups: categorical data and ordinal data. Categorical data included true-false questions, demographic information, and binary variables capturing responses to open-ended questions (n = 315). Ordinal data included all Likert-type variables (e.g. responses ranging from 1 – ‘strongly agree’ to 7 – ‘strongly disagree’; n = 173). All tests were ran twice, once with the sampling weights provided by Ipsos MORI and once without.

For categorical data, the distributions of responses in the booster sample were compared to the distribution of responses in the main sample using the Pearson’s Chi-Square test. Using the Exact Tests extension for SPSS 22, we collected from each chi-square test the exact two-tailed probability of acquiring the distribution of responses across samples, given the null-hypothesis (that people sampled using the Quota method have the same distribution of responses as people sampled by the probability method). For ordinal data, we compared the distribution of responses in the booster sample to those in the main sample using the Mann-Whitney *U* test. We collected

from each test the two-tailed probability of acquiring the given distribution of responses across samples, given the null-hypothesis.

***p*-value transformation.** Having acquired the p-values associated with each of the 488 tests, we then compared the distribution of probabilities to the expected distribution under the null-hypothesis. This involved first assigning a weight to each *p*-value, and then transforming the *p*-value, to create a weighted Fisher's statistic (Good, 1955). This method was selected because it "produces more accurate *p*-values than other methods" (Alves & Yu, 2014), including the methods of Fisher, Lancaster, Strouffer *et al.* and Lipták. The function for transforming the variables is: $w_i[\ln p_i]$, where w_i serves as a weight function, adjusting the *p*-values by their relative weight. While there is no agreed upon method for how to weight *p*-values, there is significant evidence that using weights helps to "improve the accuracy of the combined *P*-value" (Alves & Yu, 2014, p. 10), especially when combining possibly dependent tests. Following the recommendation of Whitlock (2005), each test was weighted proportional to the inverse of the squared error variance (i.e. the standard deviation of the mean). The weights were normalized to sum to one. We also ran a set of analyses where each test was weighted equivalently. For these tests, the weights were equal to $1/k$, where k is the number of *p*-values being combined. All transformations were conducted in Excel 2013.

Summation and test-statistic. To arrive at our final test statistic, we summed the values as shown: $X = -2\sum_{i=1}^{k} w_i \ln(p_i)$. Hou (2005) showed that the null hypothesis associated with the combined weighted *p* value can be evaluated with the approximate rejection region $X > c\chi_f^2(\alpha)$. When the sum of the weights is 1, then $c = \sum_{i=1}^{k} w_i^2$, and $f = \frac{2}{\sum_{i=1}^{k} w_i^2}$. When $w_1 = w_2 = \cdots = w_k = 1/k$, these reduce to $c = \frac{1}{k}$, and $f = 2k$, or "exactly Fisher's original procedure"

(Hou, 2005, p. 183) We evaluated this function using Wolfram Mathematica 10.0. Keeping with social science tradition, we set the rejection zone for the null hypothesis at $p \leq .05$.

Note on Additional Analysis. It is worth noting that, in addition to the four sets of tests described in the results (with / without sampling weights & with / without variable weights), we also independently tested each type of variable, to ensure that any effects we noted were not due to the treatment of the variable, the type of test used to derive the *p*-value, etc. We separately analyzed three sets of tests: tests of ordinal closed-ended responses, tests of nominal closed-ended responses, and tests of nominal open-ended responses. In all cases, the results were not substantively different from the overall results.