Social Research Association

Issue 1
Winter 2015

Social Research Practice

The SRA journal for methods in applied social research

Social Research Practice

Issue 1 Winter 2015

The Social Research Association journal for methods in applied social research

Contents

Editorial

Richard Bartholomew

Editor

Welcome to the first issue of **Social Research Practice**, *the SRA's new journal for methods in applied social research.*

The journal will provide a twice yearly forum for developing and discussing quantitative and qualitative methods in designing, conducting and commissioning research on social policy and practice.

The journal aims to encourage methodological development by helping practitioners share their knowledge. Many social researchers have limited opportunities and too few incentives to reflect on their methodological experiences and insights. There are often competing pressures of reporting results, submitting new tenders, applying for grants or commissioning new work. By establishing this new journal, the SRA aims to give researchers both the space and the incentive to reflect and share their knowledge but to do so in a format which is more practical and, we hope, less daunting than the classic academic journal.

We want to encourage exploration and discussion of cutting edge approaches as well as a greater willingness, on the part of research commissioners, to take the plunge and fund new methods. SRA members are in a good position to relate, not just the theory behind new approaches, but also how they can be made to work (or adapted) in practical research settings.

Many SRA members are also using and applying the best research to try to influence the thinking of policy makers, whether in government or the voluntary sector. The journal will, therefore, discuss how messages from research can have a practical effect – the so-called 'impact agenda'. What have we learned about the more and less successful methods for achieving this?

The four articles in this first issue are good examples of researchers experimenting with new approaches and reflecting on what can be learned from practical experience.

Declining response rates are a major challenge for maintaining the quality of social research. **Gerry Nicolaas** and colleagues present the results of a large experiment using the GP Patient Survey to test different approaches to maximising response rates and minimising non-response bias in a large national postal survey. The scale of the survey has allowed the research team to directly compare the effectiveness of different strategies using randomly-assigned treatment and control groups.

Likert agreement scales have been widely used in survey questionnaires for many years. But they have limitations – potential acquiescence bias and the complexity which agreement scales add to the response process, including the burden on respondents. **Tim Hanson** describes an experiment comparing responses to agreement-scale questions with those to equivalent questions using item-specific scales which directly capture the dimension of interest. The results, and other related research, suggest that using item-specific scales may, in some cases, provide more reliable measures.

As many parents will attest, gaining and retaining teenagers' attention can be difficult; the potential distractions are so pervasive. This is the challenge which **Lisa Calderwood** and colleagues face in mounting the current wave of the Millennium Cohort Study of 14-year-olds. Just how do you persuade thousands of teenagers that it's worthwhile taking part in your survey, and then keep them interested for over an hour? This article provides helpful insights for those conducting or commissioning research with young people, not least on asking about sensitive topics.

In the final article, **Ruth Levitt** and **William Solesbury** dissect the challenges involved in ensuring that social research findings inform policy. Their case study of 'policy tsars' illustrates the limitations of essentially linear concepts of research impact. It also demonstrates the need for a more sophisticated understanding of the other forces (interests, ideology and institutions) which influence the extent to which research affects policy.

I would like to thank all the members of the editorial board for their support in launching the journal, and particularly those who have refereed articles for this first issue.

If you are interested in offering an article for a future edition of Social Research Practice, details are on our website at **www.the-sra.org.uk** along with guidelines for authors and a template for articles. If you have an idea for an article but are not sure if it is suitable, just drop me a line: **rabartholomew@btinternet.com**

Increasing response rates in postal surveys while controlling costs: an experimental investigation

Gerry Nicolaas, Patten Smith and Kevin Pickering, Ipsos MORI
Chris Branson, NHS England

Abstract

Although much is known about maximising postal survey response rates, survey clients and practitioners would benefit from a better understanding of how to maximise response and minimise non-response bias while controlling costs. We present the results of a large experiment carried out in England which tests the impact of four design features: a pre-notice letter; a postcard reminder; cover letter design; and length of questionnaire. The large size of the experiment allows us to examine the impact of each separate feature as well as combinations. We examine the impact on response rates; the socio-demographic profile of the achieved sample; and key survey estimates. We also discuss the implications for survey costs.

Acknowledgement

This study was carried out on the GP Patient Survey which is funded by NHS England.

We would also like to thank Alex Kong and Will Scott, Ipsos MORI, for their contribution to this study.

Introduction

In this paper we report the results of an experiment designed to identify interventions which could be used to increase response rates on a large English postal survey. In the remainder of this section, we describe the background to the experiment and the rationale for the work.

Background

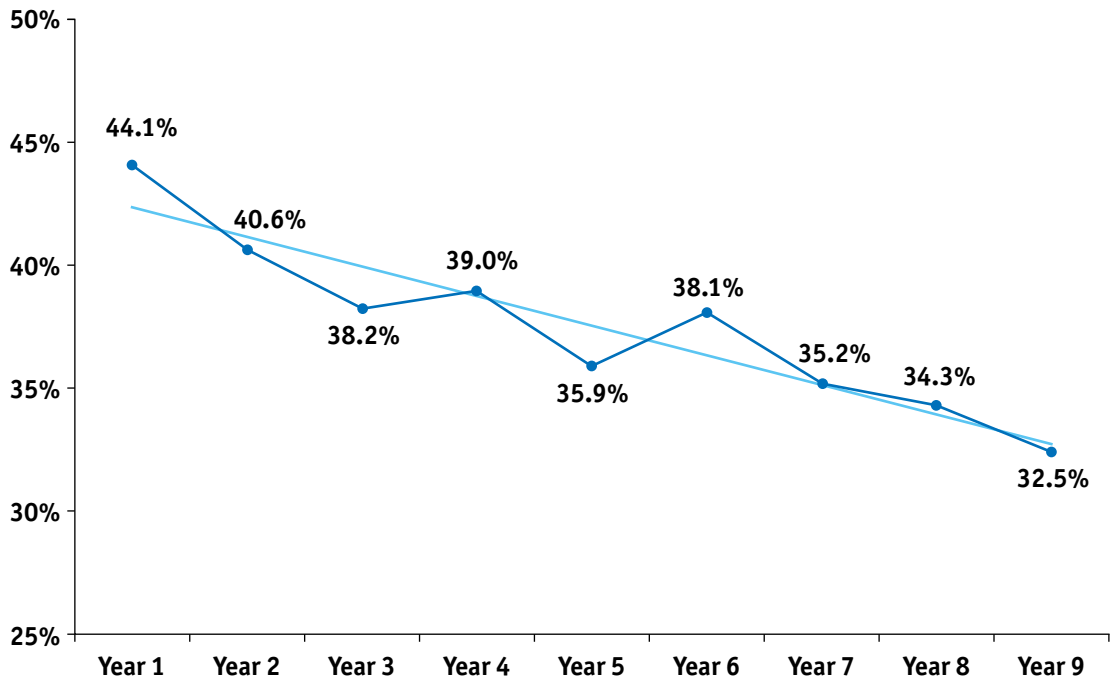
The GP Patient Survey (GPPS)[1], which is funded by NHS England, collects data on patient experiences, attitudes and characteristics from patients who have been registered with a GP practice in England continuously for at least six months and are 18 years of age or over (HSCIC, 2014). The survey has just completed its ninth year.

The current design of the survey involves sending around 2.6 million postal questionnaires across two waves of fieldwork each year (July to September, and again from January to March). Two reminder packs, each comprising a letter and another copy of the questionnaire, are sent, at monthly intervals, to those not yet replying. Patients are able to complete the survey on paper, online or by phone when calling the helpline. There is also an option to complete the survey in 13 languages other than English, plus British Sign Language.

Survey response rates are often used as a proxy indicator of data quality. Although GPPS response rates have fluctuated over time, in common with response rates for other patient surveys, they have exhibited a clear downward trend over time which could reduce the perceived credibility of the survey data (figure 1).

[1] GP is an abbreviation of general practitioner which roughly equates to a family doctor.

Figure 1: Response rates to the GP Patient Survey



Source: GP Patient Survey, 2006-2007 (Year 1) through to 2014 (Year 9 Wave 1)

Rationale for the work

In the light of this decline, Ipsos MORI and NHS England undertook a review of possible interventions which might be used to increase response rates. Our review was primarily guided by the Tailored Design Method (TDM) approach to postal survey design developed by Professor Don Dillman and his colleagues.

The TDM is an approach based on social exchange theory principles, which predict that survey participation is best encouraged by minimising its perceived costs and maximising its perceived benefits. It was developed over 40 years ago by Professor Dillman, and has been regularly revised since, in the light of the widespread use of TDM, changes in technology (Dillman et al, 2009) and empirical findings on methods for increasing response rates in postal surveys (for a comprehensive review see Edwards et al, 2009).

The TDM guidelines for postal survey implementation were used to evaluate the design of the GPPS in order to identify design features that could be changed to potentially increase the survey response rate without substantially increasing costs. Although several of its principles had already been applied to the GPPS design (for example, personalisation, multiple contacts, replacement questionnaires, and so on), we identified four main areas where we felt that TDM-based design enhancements would have a good chance of improving response rates. Of these four areas, one, the provision of respondent incentives, was ruled out on grounds of public acceptability. The three remaining possible enhancements were (i) varying the look and feel of each mailing by redesigning the cover letters, (ii) adding a pre-notice letter and (iii) using a postcard reminder one week after the first questionnaire mailing.

In the experiment reported here, we tested the impact of each of these interventions. We discuss the rationale for each of them below.

Five principles guided the redesign of the GPPS cover letters. These were based on the TDM (Dillman et al, 2009), supplemented with behavioural science insights (for example Cialdini, 2007) and evidence from the UK on the design of advance letters in face-to-face surveys (Finch, 1981; Clarke et al, 1987; Brook et al, 1992; Lynn, 1998; Wedeman and Farnell, 2014; Moore, forthcoming) and were as follows:

1. Letters should be short and simple
2. Letters should focus on a limited number of key messages (for example the survey's importance, motivation for taking part in the survey, confidentiality)
3. Importance should be conveyed, in particular, by giving prominence to the NHS England logo, by using a high-status signatory, and by using a professional letter format
4. Motivational statements should vary across the three letters, thereby increasing the likelihood of converting different types of non-respondents
5. Graphics should be used to break up text and increase readability

There is ample evidence that use of pre-notice letters can increase response rates (Edwards et al, 2009). Until recently, Dillman has recommended using them to provide a positive and timely notice that sampled individuals will be receiving a request to help with an important study (Dillman et al, 2009). The TDM guiding principles used for the cover letters were also used to design the pre-notice letter (Dillman et al, 2009).

Many of those who do not return completed questionnaires after a first mailing in a postal survey are not definitive refusals: they may have intended to complete the questionnaire but forgot to do so; they may not have seen the first invitation; or they may not have engaged with the request at the time. Sending a postcard reminder one week after the first questionnaire mailing could nudge these non-respondents to complete and return the questionnaire. The design of the postcard was based on the TDM approach (Dillman et al, 2009) and followed the principle that each individual reminder is more likely to encourage response if it stands out as being different in form and/or content from previous communications. The main intention of the postcard was to jog memories and rearrange priorities, rather than to overcome resistance.

The TDM also recommends reducing the cost of participation to the respondent by making the questionnaire short and easy to complete. There is ample evidence in the literature showing that questionnaire length is positively related to postal survey response rates (Edwards et al, 2009). For this reason, it was decided also to investigate the impact of questionnaire length on the GPPS response rate.

Although survey response rates are often deemed important in their own right, the ultimate purpose of increasing them is to reduce non-response bias. The relationship between non-response levels and non-response bias has often been found to be surprisingly weak (Groves, 2006; Groves and Peytcheva, 2008). Given this, we considered it important in this experiment to investigate the impact of the interventions on both response rates **and** non-response bias.

Finally, some of the interventions which increase response rates also incur additional costs (for example postcard reminders) whereas others are cost neutral (for example redesign of letters). Decisions about which enhancements to adopt, if any, will need to consider the trade-offs between survey cost, sample size and response rate. We therefore give brief consideration to the costs of possible survey enhancements in our analysis.

Method

Two experiments were carried out to investigate the impact of using a pre-notice letter, a postcard reminder, redesigned cover letters and a shorter questionnaire.

Experiment 1

Experiment 1 was embedded within the main GPPS year 9, wave 2 survey, and tested the effect of the pre-notice letter, the postcard reminder and redesigned cover letters. It was carried out on a small sub-sample (1.6%) of the cases issued for the main survey. This sub-sample was selected in proportion across all the GP practices in the sample.

Sampled cases were allocated at random to each of the following treatment groups:

- Treatment A: pre-notice letter (sub-sample of 3,000)
- Treatment B: postcard (sub-sample of 3,000)
- Treatment C: redesigned cover letters (sub-sample of 3,000)
- Treatment AB: pre-notice letter and postcard (sub-sample of 3,000)
- Treatment AC: pre-notice letter and redesigned cover letters (sub-sample of 3,000)
- Treatment BC: postcard and redesigned cover letters (sub-sample of 3,000)
- Treatment ABC: pre-notice letter and postcard and redesigned cover letters (sub-sample of 3,000)
- Control group: no enhancements (the control group comprised the remaining GPPS sample of 1,299,972)

For pre-notice letter treatments (A, AB, AC and ABC), a pre-notice letter was sent about one week before the first questionnaire mailing.

For postcard reminder treatments (B, AB, BC and ABC), a postcard reminder was sent to all sample members one week after the initial questionnaire mailing (except for those explicitly refusing to participate). As well as reminding those who had not yet replied, the postcard thanked those who already had.

For the letter treatments (C, AC, BC, and ABC), initial and reminder letters were simplified and redesigned following the principles outlined in the introduction. As mentioned, motivational messages were varied across the three letters. They differed as follows:

1st letter: NHS England needs your help; improve health care and dental services in your area

2nd letter: we haven't received your questionnaire; we need to hear from as many people as possible

3rd letter: last opportunity to complete the questionnaire; improve healthcare and dental services that you and your family may need

All three letters included a short statement on confidentiality.

Copies of the standard cover letters, redesigned cover letters, pre-notice letter and postcard are at: **http://the-sra.org.uk/smith-nicolaas-2015**

The experiment was designed to allow analysis of the impact of each of the three suggested changes comparing each against the current approach, and of all interactions between them. The sample sizes are sufficiently large to allow us to identify which combination of changes would be maximally effective.

Experiment 2

Experiment 2 examined the impact of shortening the length of the questionnaire from eight to four pages, with and without design enhancements. Unlike experiment 1, experiment 2 could not be included as part of the main survey because important GPPS questions had to be dropped from the short questionnaire. The shorter version of the questionnaire was, therefore, issued to an additional 6,000 cases sampled in an identical manner to the main sample.

These extra cases were allocated to two treatment groups:

- Treatment D: shorter version of the questionnaire (additional sample of 3,000)
- Treatment ABCD: shorter version of the questionnaire **and all three** changes tested in experiment 1 (pre-notice letter, postcard reminder and redesigned covering letter) (additional sample of 3,000)

The first treatment group allowed us to test the impact of shortening the questionnaire against the current design. The second allowed us to test the additional impact of shortening the questionnaire over and above making the three enhancements tested in experiment 1. Its inclusion was intended to generate evidence on whether shortening the questionnaire would give additional meaningful benefit even if the three other changes were adopted.

Allocation of sample to treatment groups for both experiments

An initial sample was selected of 1.327m cases (1.321m cases for experiment 1 plus 6,000 cases for experiment 2). From this, 27,000 cases were systematically selected (using the method of random start and fixed interval) to take part in the experiment, and allocated to each of the nine treatment groups.

Results

Response rates: experiment 1

Table 1 shows final response rates for each experimental treatment. None of the treatment groups had a lower final response rate than the control group. The highest final response rate was achieved among those who received both the postcard and the redesigned letters (treatment group BC); 8.7 percentage points higher than the final response rate for the control group.

Table 1: Experiment 1 response rates

	Final response rate (%)
Control	32.7
A: Pre-notice letter	32.9
B: Postcard reminder	37.8
C: Redesigned letters	34.4
AB	38.6
AC	34.9
BC	41.5
ABC	40.3

The significance of main effects and interactions for the three TDM features were tested using multiple linear regression (table 2). These tests confirm the patterns which we can see in table 1:

- The pre-notice letter had no significant impact on the response rate
- The postcard had a highly significant positive effect on response rate, increasing response rate by an average of 5.7 percentage points
- The redesigned letters also had a significant positive effect on response rate, increasing response rate by an average of 2.1 percentage points

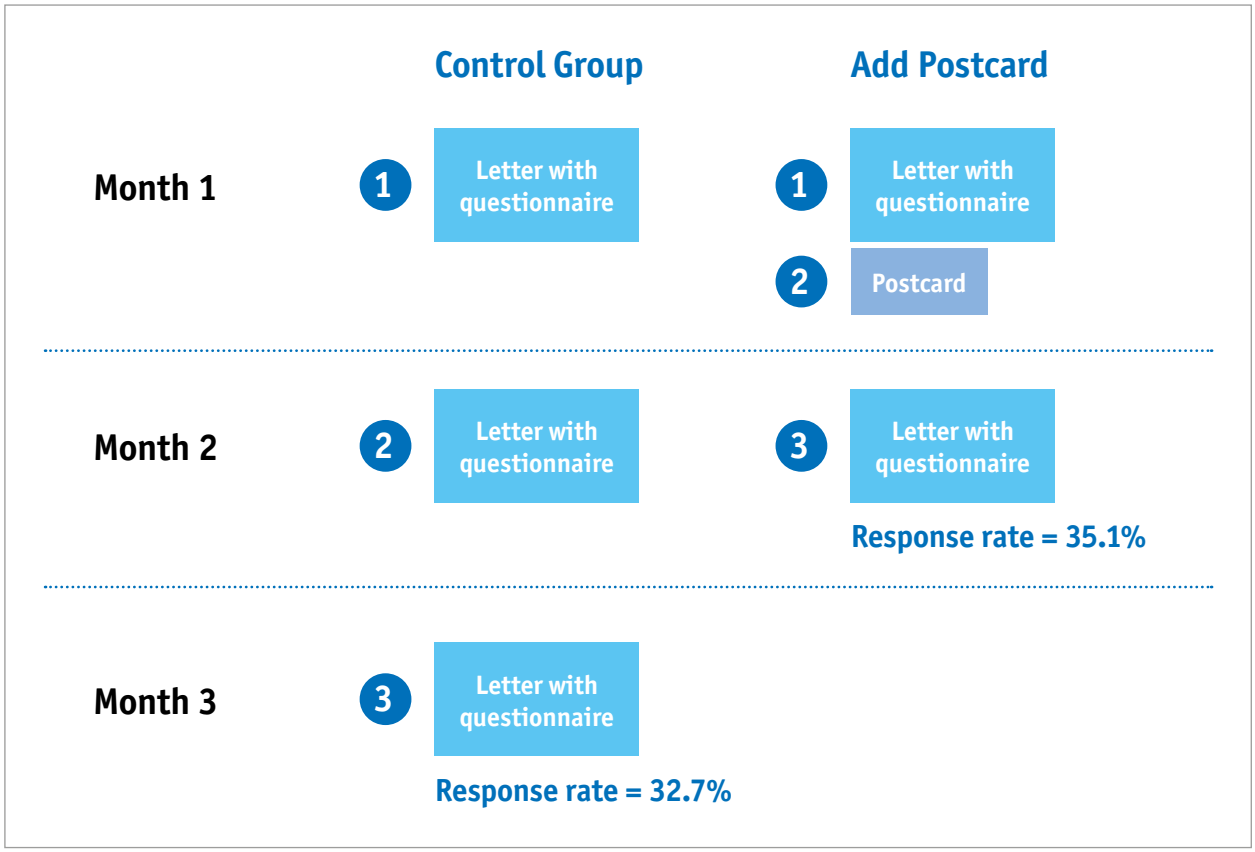
We note that there were no significant interactions between the three features, and for this reason they are not shown in table 2. In other words, the effects of the postcard and the redesigned letters on the final response rate were additive.

Table 2: Linear regression of final response rate on experimental treatments

	Coefficient	Standard error	t value	Significance
Constant	0.327			
A. Pre-notice letter	-0.000	0.005	-0.05	0.964
B. Postcard	0.057	0.005	10.93	0.000
C. Redesigned letters	0.021	0.005	4.08	0.000

It is generally accepted that the response to postal surveys will increase with the number of contacts being made (Dillman et al, 2009). It is, therefore, possible that the postcard's impact on the final response rate was simply due to the fact that it involved an extra contact attempt, and that the same impact would have been achieved if the postcard had been substituted with another questionnaire mailing. If this were so, we would expect the response rate for treatment group B, after the first questionnaire reminder, to be similar to the response rate for the control group after two questionnaire reminders (when both groups had received three mailings – see Figure 2). However, results from the experiment show that the response rate for treatment group B, after the first questionnaire reminder, is 2.3 percentage points higher than the final response rate for the control group ($t=2.71$, $p< 0.01$). This demonstrates that, for a fixed number of mailings, a higher response rate would be achieved by replacing the second standard questionnaire reminder with a postcard sent one week after the first questionnaire mailing.

Figure 2: Diagram showing mail-out strategy for control group and treatment group BC, limiting the maximum number of mailings to three



Response rates: experiment 2

Table 3 shows that shortening the questionnaire increased the response rate by 1.9 percentage points ($t=2.21$; $p<0.03$) when none of the other three design enhancements were implemented and by 3.3 percentage points ($t= 2.59$; $p<0.01$) when all such enhancements were implemented.

Table 3: Experiment 2 response rates

	Final response rate (%)
Control	32.7
D: Short questionnaire	34.6
ABC	40.3
ABCD	43.6

Sample profile and non-response bias

As stated in the introduction, the main reason survey researchers aim to maximise response rates is to reduce non-response bias. In GPPS (as in most surveys), for most variables, we have no population benchmark distributions against which to compare survey distributions, and for these variables cannot measure non-response bias directly. We can, however, assess whether treatments which increase response rates also change levels of non-response bias by comparing distributions across sub-samples allocated different treatments. Of course, to identify changes in levels of non-response bias as response rates increase, is not to identify reductions in such bias (although in the absence of anything else to go on, we often do make the assumption – implicitly or explicitly – that this is the case).

In order to assess whether our experimental treatments affected levels of non-response bias, we compared sample profiles and a range of outcome measures for the treatment groups against the control group. In order to give an indication of where there might be differences we used a t-test – at the 0.05 significance level – to identify significant differences between the estimates for every category for each measure against the corresponding estimates in the control group. Given the number of comparisons made, some significant differences were to be expected by chance. Therefore, these tests were used cautiously and purely to identify patterns in the profiles and outcome estimates.

Table 4: Demographic and patient experience questions used in comparisons

Demographic variables	Patient experience variables
Age (8 categories)	Generally, how easy is it to get through to someone at your GP surgery on the phone? (5 categories)
Gender (2 categories)	How helpful do you find the receptionists at your GP surgery? (5 categories)
Ethnicity (18 categories)	How convenient was the appointment you were able to get? (4 categories)
Religion (9 categories)	Overall, how would you describe your experience of making an appointment? (5 categories)
Working status (8 categories)	Did you have confidence and trust in the GP you saw or spoke to? (4 categories)
Whether a parent or legal guardian (2 categories)	Did you have confidence and trust in the nurse you saw or spoke to? (4 categories)
Whether a carer (6 categories)	How satisfied are you with the hours that your GP surgery is open? (6 categories)
Sexual orientation (5 categories)	Overall, how would you describe your experience of your GP surgery? (5 categories)
Whether has a long-standing health condition? (3 categories)	Would you recommend your GP surgery to someone who has just moved to your local area? (6 categories)
EQ-5D score	Overall, how would you describe your experience of NHS dental services? (5 categories)

When comparing the treatment groups with the control group across these measures, several significant differences across categories would be expected by chance alone. For all but two variables, the number of statistically significant differences found was in line with the number estimated to appear by chance, and where there were significant differences, discernible patterns were not apparent. The two exceptions to this were age and working status.

Table 5 shows the age distribution for the control sample and for the five experimental treatments which led to a response rate increase. (Given that the pre-notification treatment had no impact on final response rate, pre-notification treatments have been combined with appropriate non-pre-notification treatments.)

Inspection of the table reveals that shortening the questionnaire, sending a postcard reminder on its own or using redesigned letters on their own made little difference to observed age distributions (although using redesigned letters slightly decreased the proportion of 65- to 74-year-olds). On the other hand, when postcards were combined with redesigned letters and the full-length questionnaire was used, the proportion of the sample aged 25 to 54 increased, and the proportion aged 65 and over decreased. Similarly when postcards and redesigned letters were used with the shorter questionnaire, the proportion of 25 to 54 year olds increased (although not significantly for 25- to 34-year-olds), and the proportion aged 65 and over decreased (although only significantly for those aged 65 to 74).

Unexpectedly, for this treatment group, the proportion aged under-25 **decreased** significantly. Whether or not this finding proves to be robust, it is clear that improved response rates led to no improvements in the representation of under-25s in the sample.

Of course, we are unable to assess directly from table 5 whether or not these observed changes in age distribution represent reductions in non-response bias unless we compare them against criterion data. The distributions shown in table 5 are based on unweighted data, and are not strictly comparable with published population figures[2]. In table 6, we therefore show corresponding figures for the control condition and for the two combined treatments (postcards + redesigned letters; and postcards + redesigned letters + short questionnaire) **after applying inverse probability weights** alongside the corresponding Office for National Statistics population statistics.

Nearly all adults in England are registered with a GP, and hence eligible for inclusion in the GPPS[3]. We can, therefore, draw two important conclusions from this table. First, the control sample substantially over-represented those aged 55 and over, and under-represented those aged under-55. And second, the experimental interventions slightly reduced age bias in the sample by virtue of increasing the proportion of 25-44 year olds and decreasing the proportion of those aged 65 and older. However, despite this, the sample remained substantially age biased after weighting by inverse probability weights.

It is important to note that such age bias can be (and is when GPPS results are reported) largely eliminated through post-stratification weighting.

[2] GPPS does not use an equal probability sample design because it sets a minimum sample size for each practice; inverse selection weights therefore need to be applied when making national estimates.

[3] A recent study carried out by Ipsos MORI found that 96% of people aged 15 or over were registered with a GP practice (Ipsos MORI, 2015).

Table 5: Age distribution by experimental treatment (unweighted)

	Control	Postcard reminder (B+AB)	Redesigned letters (C+AC)	Postcard reminder+ Redesigned letters (BC+ABC)	Short q'naire (D)	Short q'naire+ Postcard reminder+ Redesigned letters (ABCD)
Response rate	32.7%	38.2%	34.7%	40.9%	34.6%	43.6%
Age	Col. %	Col. %	Col. %	Col. %	Col. %	Col. %
18-24	3.6%	3.2%	4.0%	3.5%	3.7%	2.6%*
25-34	8.5%	9.3%	7.9%	9.8%*	8.2%	9.8%
35-44	12.0%	13.2%	13.3%	13.7%*	12.0%	14.2%*
45-54	16.8%	17.3%	17.7%	18.6%*	16.2%	19.9%*
55-64	20.1%	19.7%	21.1%	21.2%	20.7%	19.7%
65-74	21.7%	21.4%	19.7%*	19.5%*	20.3%	18.2%*
75-84	13.1%	12.1%	12.6%	10.4%*	14.3%	12.3%
85+	4.2%	3.9%	3.6%	3.4%*	4.5%	3.3%
Base:	419,163	2,269	2,053	2,422	1,032	1,300

*** Significantly different from control (P<0.05).**

Table 6: Population and survey age distributions (weighted)

Age	Population*	Control	BC+ABC	ABCD
18-24	11.5%	3.3%	3.5%	2.8%
25-34	17.4%	7.7%	8.9%	8.6%
35-44	16.6%	11.5%	13.0%	14.0%
45-54	17.9%	16.7%	19.2%	18.9%
55-64	14.3%	20.2%	21.1%	19.5%
65-74	12.1%	22.5%	19.8%	19.6%
75-84	7.3%	13.6%	11.0%	13.1%
85+	3.0%	4.5%	3.5%	3.5%

***ONS 2014 mid-year estimates for England**

The analyses of working status (table 7) were in line with the results on age. The higher response rate treatments tended to increase the proportion of people in full-time work and reduce the proportion retired (although not always by a statistically significant margin). The short questionnaire treatment was an exception to this, however. Despite leading to a comparable increase in response rate to the redesigned letter treatment (B + AB), providing a short questionnaire (D) did not increase the proportion of full-time workers or decrease the proportion in retirement. This is in line with the findings, reported above, showing that providing a short questionnaire had no significant impact on the sample age distribution (table 5).

Finally, given that no other similar finding is to be found in the table, we suspect that the increase in the proportion of sick and disabled in treatment ABCD is purely down to chance.

Unfortunately, no strictly comparable population data are available for comparison (working status figures are critically dependent on how data are collected and which definitions are used). But a comparison with the Annual Population Survey employment rates[4] indicated that, even with experimentally-induced response rate increases, the GPPS sample[5] substantially underestimated the proportion of the population in employment. In practice, when reporting GPPS results, this bias is addressed by including controls for age in post-stratification weighting.

Table 7: Working status by experimental treatment

	Control	Postcard reminder (B+AB)	Redesigned letters (C+AC)	Postcard reminder+ Redesigned letters (BC+ABC)	Short q'naire (D)	Short q'naire+ Postcard reminder+ Redesigned letters (ABCD)
Response rate	32.7%	38.2%	34.7%	40.9%	34.6%	43.6%
Full-time paid work (30 hours or more each week)	32.9%	35.3%*	35.5%*	35.7%*	32.7%	34.7%
Part-time paid work (under 30 hours each week)	13.3%	13.9%	13.4%	14.3%	13.1%	14.2%
Full-time education at school, college or university	1.4%	1.1%	1.6%	1.4%	1.4%	0.9%
Unemployed	3.7%	3.6%	3.4%	4.3%	3.1%	2.9%
Permanently sick or disabled	4.5%	4.2%	4.7%	4.4%	4.8%	6.9%*
Fully retired from work	36.4%	34.7%	34.0%*	32.4%*	36.6%	33.2%*
Looking after the home	5.4%	5.6%	5.3%	4.9%	6.0%	4.7%
Doing something else	2.4%	2.1%	2.2%	2.8%	2.3%	2.5%
Base	405,899	2,202	1,973	2,343	939	1,179

*** Significantly different from control (P<0.05).**

Costs

We estimated the approximate costs of printing, despatch, postage and scanning for the control group and each treatment, and used these to calculate the marginal costs of obtaining a completed questionnaire for each treatment. We then created a cost index, set to 100 for the control group, showing the relative costs of obtaining a set achieved sample size with each treatment.

[4] For the 12 months starting in April 2014, the Annual Population Survey estimated the employment rate for those aged 16+ in England as 59.6%. The highest observed employment rate, after weighting by inverse probability weights, in the experiment was 52.6% for the ABC treatment.

[5] When weighted by inverse probability weights only.

Table 8 shows the cost index and what can be achieved within a fixed budget for each of the effective (response rate increasing) experimental conditions apart from the short questionnaire + postcard + enhanced letters treatment for which separate costs could not be estimated[6]. Table 8 shows:

- The shorter questionnaire on its own had the lowest cost per completed questionnaire, the largest achieved sample size for a fixed budget and was associated with a modest increase in the final response rate
- Relative to the control group, redesigned letters reduced the cost per completed questionnaire, slightly increased the fixed-cost sample size and modestly increased the response rate
- The postcard reminder increased the cost of each completed questionnaire despite its positive impact on response rate: the reduction in the total number of reminders required was not large enough to offset the additional costs of printing and posting the postcards
- The combined use of postcard reminders and redesigned letters reduced the cost of each completed questionnaire, increased the fixed budget sample size and was associated with a substantial increase in response rate

Table 8: Response rates, cost per completed questionnaire and achieved sample sizes for a set budget

	Final response rate	Cost per completed q'naire (control indexed at 100)	Achieved sample size for set budget (control indexed at 1,000)
Control group	32.7%	100	1,000
D. Shorter questionnaire	34.6%	92	1,088
C. Redesigned letters	34.4%	96	1,038
B. Postcard reminder	37.8%	106	940
B + C. Postcard reminder + Redesigned letters	41.5%	97	1,026

Discussion and conclusions

In this study we tested the impact of four TDM strategies for increasing the GPPS response rate: a pre-notice letter, a postcard reminder, redesigned cover letters, and reduced respondent burden by using a shorter questionnaire. All of these, apart from the pre-notice letter, had a positive impact on the final response rate. The lack of impact of a pre-notice letter may be seen as surprising given previous research indicating that these letters can increase response rates by around three to six percentage points (Dillman et al, 2009). However, more recently Dillman and his colleagues (Dillman et al, 2014) have suggested that, as people become more inundated with messages and requests, the efficacy of pre-notice letters may be in decline and this, perhaps, renders our finding less surprising.

The three other treatments had positive effects on the final response rate with the postcard reminder having the greatest effect: using regression modelling, we estimated that the effect of using a postcard reminder was to increase the response rate by an average of 5.7 percentage points. This finding is very much in line with Dillman's expectation that a postcard reminder will increase response rates by between five and eight percentage points (Dillman et al, 2009). The redesigned letters and shorter questionnaire were each estimated to increase the response rate by an average of about two percentage points. These findings were in line with our general expectations from the literature: although we anticipated finding modest response rate increases, the literature offered little guidance as to their likely magnitude (Edwards et al, 2009).

[6] Because the ineffective but cost increasing pre-notification letters were always used with this treatment.

The impact of using a postcard reminder and redesigned letters was additive when used with the full-length questionnaire. If we assume that the impact of reducing questionnaire length does not interact with the impact of postcard reminders or of the redesigned questionnaire[7], the results suggest that it would be possible to increase the GPPS response rate from 32.7% to around 42-43% by combining all three treatments. However, the use of a shorter questionnaire would be undesirable because it would substantially reduce the amount of data that can be collected. Even without shortening the questionnaire, we estimate that using postcard reminders and redesigning the letters would increase response rate to around 40-41%.

As previously noted, the main reason for maximising response rates is to reduce non-response bias. An increase in the response rate will only reduce non-response bias if the postcard reminder, redesigned letters and shorter questionnaire are disproportionately attractive to sample members who would otherwise be under-represented in the survey. Of 20 variables compared (10 socio-demographic and 10 patient experience variables), none of the patient experience and only two of the socio-demographic variables (age and activity status) exhibited any systematic changes in distributions under the different experimental conditions. The postcard reminder and redesigned letters marginally improved the sample profile by slightly increasing the proportions of 25-44 year olds and people in work and slightly decreasing the proportions of those aged 65 and over and in retirement. However, even with this slight improvement, the sample remained substantially age biased[8].

Cost comparisons indicated that shortening the questionnaire, redesigning the letters and simultaneously redesigning the letters and sending postcard reminders, all reduced the cost of each completed questionnaire.

Despite the very limited positive impact on sample profile, we conclude that it is still worthwhile to increase the GPPS response rates because this will increase trust in the GPPS estimates among its data users and other stakeholders. Further, we conclude that this is best achieved through the use of postcard reminders and redesigned cover letters because this delivers a substantial increase in response rate; increases the completed sample size for a fixed budget; and requires no reduction in questionnaire length.

Postcard reminder and redesigned letters have been introduced on the 10th wave of the GPPS. Further research will explore the actual cost implications of these changes to the design of the survey.

It is often hazardous to generalise methodological findings from the particular surveys which generated them to surveys more broadly. However, in this case we think there is some justification for doing so for two reasons. First, unlike many methodological studies which cover narrowly defined populations (for example students or particular professional groups), this one is based on a general population sample, and as such, it is reasonable to expect its findings to be generalisable to the many surveys which cover broadly-drawn population groups. Second, as discussed above, our findings are very much in line with those from previous methodological research. Indeed, the very reason we chose to test the treatments we did was that the previous literature had suggested they may prove to be fruitful.

With the rapid growth of online surveys, the traditional postal survey might be regarded as being old fashioned and of little relevance to research today. Such views are, we believe, misguided. Postal surveys remain effective ways of collecting data from genuinely random samples of many populations which cannot be effectively sampled for online surveys. As this paper demonstrates, they can deliver reasonable response rates if good-practice guidelines are scrupulously followed. In saying this, we do not wish to underplay the fact that the estimates they produce can be subject to significant levels of non-response bias such as we observed here. We fully acknowledge that this can be a problem. However, we have no reason to suppose that it is a greater problem for postal surveys than for alternative data collection modes.

[7] Our design did not allow us to test for this interaction.

[8] This age bias is largely corrected by post-stratification weighting when GPPS results are reported.

References

Brook, L., Prior, G. and Taylor, B. (1992) British Social Attitudes 1991 Survey: Technical Report. London: Social and Cultural Planning Research.

Cialdini, R. B. (2007) Influence: The Psychology of Persuasion. New York: Collins.

Clarke, L., Phibbs, M., Klepacz, A. and Griffiths, D. (1987) 'General Household Survey advance letter experiment.' Survey Methodology Bulletin 21: 39-42.

Dillman, D. A., Smyth, J. D. and Christian, L. M. (2009) Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method. 3rd Ed. Hoboken, New Jersey: Wiley.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014) Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. 4th Ed. Hoboken, New Jersey: Wiley.

Edwards, P. J., Roberts, I., Clarke, M. J., Diguiseppi, C., Wentz, R., Kwan, I., Cooper, R., Felix, L. M. and Pratap, S. (2009) 'Methods to increase response to postal and electronic questionnaires (Review).' Cochrane Database Systematic Review 3: MR000008.

Finch, C. (1981) 'General Household Survey letter experiments.' Survey Methodology Bulletin 13: 30-37.

Groves, R. M. (2006) 'Nonresponse Rates and Nonresponse Bias in Household Surveys.' Public Opinion Quarterly 70(5): 646-675.

Groves, R.M. and Peytcheva, E. (2008). 'The impact of nonresponse rates on nonresponse bias: a meta-analysis.' Public Opinion Quarterly 72(2): 167-189.

HSCIC (2014). Numbers of Patients Registered at a GP Practice – Oct 2014. [online] Available at **www.hscic.gov.uk/searchcatalogue?productid=16172&topics=1%** [Last accessed 09/09/2015].

Ipsos MORI (2015). Exploring patient choice in GP services – 2014 survey. A report prepared for Monitor. [pdf] Available at **http://www.bgs.org.uk/pdfs/2015coice_in_GP_services_MORI_survey.pdf** [Last accessed 09/09/2015].

Lynn, P., Turner, R. and Smith, P. (1998) 'Assessing the effects of an advance letter for a personal interview survey.' Journal of the Market Research Society 40: 265-272.

Moore, H. (forthcoming) 'Opinions and Lifestyle Survey Advance Letter Split Sample Trial.' Survey Methodology Bulletin.

Wedeman, L. and Farnell, J. (2014) "Communicating with Respondents: Using Qualitative Research to Improve ONS's Advance Letter for Social Surveys". Survey Methodology Bulletin 72: 99-133.

Comparing agreement and item-specific response scales: results from an experiment

Tim Hanson, TNS BMRB

Abstract

First developed by Likert, agreement scales have been widely used in survey questionnaires for many years. However, there are potential shortcomings associated with the format, including the potential for acquiescence bias and the complexity that agreement scales add to the response process, placing an additional, and often unnecessary, burden on respondents. We conducted an experiment to compare responses to agreement-scale questions with those to equivalent questions using item-specific scales which directly capture the dimension of interest. The results of the experiment, and other related research, suggest that using item-specific scales may produce more reliable measures. This should be borne in mind by researchers when designing questionnaires, particularly when they are not constrained by maintaining a time-series, and have the freedom to develop new questions.

Background

Since the development of the Likert Scale in 1932, agreement scales have been widely used in questionnaire design. Likert used his scale to identify the extent of a person's beliefs, attitudes or feelings towards objects by asking the extent to which they agreed with a statement using a five-point scale, ranging from 'strongly agree' at one end to 'strongly disagree' at the other, and a 'neither agree nor disagree' code included in the middle.

The agreement scale measure has maintained its popularity ever since, whether in the form of Likert's original five-point scale; a four-point version with the midpoint removed; or alternative versions with a greater number of points included. The format offers the opportunity to present a range of measures in a uniform, and therefore efficient, manner. There is only a need to present the agreement scale once, at the start of a battery of questions. An alternative would be an 'item-specific' approach, using response categories which relate directly to what each question is seeking to measure. For example, when seeking to measure the perceived importance of an issue or concept, respondents would be shown an importance scale (from 'very important' to 'not at all important'), rather than presenting them with a statement about the importance of a concept and asking them how strongly they agree with it. If item-specific scales were used for each question, however, this would mean introducing the respondent to each scale, and therefore add time and potential complexity to the process. Partly as a result, the agreement scale format is widely used and has mostly been seen as a 'tried and tested' mechanism for asking survey questions.

In recent years, however, a number of concerns have been raised about agreement scales which may lead to researchers questioning their future use – or at least considering the feasibility of using alternative formats:

- Agreement scales may encourage acquiescence, a category of response bias in which survey respondents have a tendency to agree with statements, regardless of their content. Saris el al (2010) report that 'more than one hundred studies using a wide variety of methods have demonstrated that some respondents are inclined to agree with just about any assertion', with studies showing that 10-20%

of respondents tend to agree with both a statement and its opposite, when the direction is reversed (for example Schuman and Presser, 1981). The risk of acquiescence bias will partly depend on the circumstances that questions are asked in and the type of respondent; for example, the risk is likely to be greater for questions asked towards the end of lengthy interviews or if respondents are less engaged in the subject matter.

- Tourangeau el al (2000) identified four components in the process of answering questions: 'comprehension of the item, retrieval of relevant information, use of that information to make required judgments, and selection and reporting of an answer'. The use of agreement scales can add complexity to this process, asking respondents to first provide their opinion about an item or concept and then to translate this to an agreement response scale. For example, The Citizenship Survey asked respondents the question: 'How much do you agree or disagree that people whose housing needs are more urgent should receive priority over those who have been waiting longer but whose needs are less urgent?'. Determining their opinion over which side of the argument to take (that is, which group should receive greater priority) does not in itself represent a straightforward task. However, once this is done, the respondent then needs to decide whether or not this opinion represents agreement with the statement. The consequence of this additional complexity in the cognitive process is likely to be a greater degree of measurement error.
- It is not always clear what disagreeing with a statement actually means. Saris et al (2010) note that if presented with a statement such as 'I am generally a happy person' and asked whether they agree or disagree with it, a respondent may disagree because they are never or rarely happy or because they are always happy. If the question instead asked people how often they were happy and used a frequency scale (for example, always, mostly, sometimes, rarely, never), we would have a far better idea of what each point on the scale represented.
- The theory of 'balanced batteries' may be flawed. When considering the agreement scale format, it is often argued that acquiescence bias is overcome by including an equal number of positive and negative statements in a battery. If acquiescence bias does exist, then it is thought that it will impact across all statements and so the battery as a whole will be 'balanced'. However, this approach only makes sense if people who do not acquiesce respond equally reliably to both positively- and negatively-phrased items. In reality, research suggests that this is not always true when people process negative statements compared with positive ones (for example Eifermann, 1961). In particular, confusion can be caused introducing negation into a statement (for example 'I am not a happy person') as this can result in a double negative when combined with the 'disagree' end of the response scale.

While agreement-scale questions continue to be widely used, there have been recent signs of movement to alternative approaches. Saris et al (2010) reported on four studies conducted in different countries, including the European Social Survey, and noted that 'the evidence from all these studies is consistent with the conclusion that data quality is indeed much higher for questions offering item-specific response options'.

Despite this ongoing debate, there has been relatively little research conducted comparing agreement scales with item-specific scales (Saris et al, 2010). To investigate this issue further, we conducted an experiment by including, on the TNS UK face-to-face omnibus, a set of agreement-scale items currently asked on a leading UK Government-commissioned survey.

The experiment

The Crime Survey for England and Wales (CSEW) includes four statements about attitudes towards the police, and asks respondents to answer using a four-point agreement scale. For our experiment, we took all four questions and developed alternative (item-specific) four-point response scales based on the nature of each question. The original questions, and the alternatives we developed, are shown in table 1. In addition to the response options shown in table 1, each question also included a 'don't know' and 'refused' code.

Table 1: The experimental questions

	CSEW agreement-scale question	Item-specific alternative
1	The police in this area can be trusted to make decisions that are right for people in this neighbourhood. (Strongly agree/Tend to agree/ Tend to disagree/Strongly disagree)	To what extent do you think that the police in this area can be trusted to make decisions that are right for people in this neighbourhood? (Always/Mostly/Sometimes/Never)
2	The police in this area abuse their power. (Strongly agree/Tend to agree/Tend to disagree/Strongly disagree)	How often, if at all, do you think the police in this area abuse their power? (Always/Often/ Sometimes/Never)
3	The police in this area reflect the mix of people in your community. (Strongly agree/Tend to agree/Tend to disagree/Strongly disagree)	How well do you think the police in this area reflect the mix of people in your community? (Very well/Fairly well/Not very well/Not at all well)
4	The police in this area understand the issues that affect this community. (Strongly agree/ Tend to agree/Tend to disagree/Strongly disagree)	How well do you think the police in this area understand the issues that affect this community? (Very well/Fairly well/Not very well/ Not at all well)

Half of respondents on the TNS omnibus were randomly allocated to be asked the agreement-scale questions while the other half were asked the item-specific questions. The questions were asked in April 2012. Interviews were conducted by Computer Assisted Personal Interviewing (CAPI), with interviewers reading the questions to respondents. Interviewers were instructed to show the screen to respondents, to allow them to see the response options. A total of 568 respondents were asked the agreement-scale questions and 525 were asked the item-specific questions. Responses were weighted at the analysis stage by sex, age, social grade and region to reflect the UK adult population aged 16+.

Each respondent was asked the same version of the questions twice: first near the start of the interview and then again towards the end of the interview. This provides a measure of the reliability of each question – the extent to which a survey would achieve the same results if repeated under identical conditions. This resulted in a 'gap' of around 20 minutes between the first and second times the questions were asked (the average interview length was 25 minutes). While spacing the questions further apart would have been desirable, previous studies have found that if the administration of two questions is separated 20 minutes or more, memory of the earlier answer is minimal (van Meurs and Saris, 1990) and so this was thought to represent an adequate time lag.

The order of the questions was fixed, so they were presented in the same order to all respondents, and in the same order when asked at the start and end of the interview. Before repeating the questions at the end of the interview, interviewers were prompted to read the following statement out to respondents: 'To help us improve our questions in the future, here are some final questions which are similar to previous ones. Please don't try to remember what you answered before but treat them as if they were completely new questions.'

The nature of an omnibus survey means that a wide range of topics can be included in a single questionnaire. The middle part of this interview (that is, between the questions included for our experiment) included questions on travel, voting intentions, stamp purchasing, smoking, gift purchasing, gambling, telephone pay phone usage and food and cooking.

Based on other experiments and theories put forward in this area, three hypotheses were set-out prior to the experiment taking place:

1. Due to acquiescence bias, a greater proportion of respondents will agree with the statements in the agreement-scale questions than provide the equivalent responses (that is top two response categories) in the item-specific questions
2. The proportion agreeing to the agreement scale statements will be greater the second time they are asked; by this time respondents are likely to be less engaged in the interview and levels of acquiescence may increase
3. There will be greater variation between responses at the start and end of the questionnaire with the agreement-scale format compared with the item-specific format. This is because the agreement-scale questions place greater cognitive burden on respondents, which may increase the proportion who answer 'randomly'

Tables 2, 3, 4 and 5 provide breakdowns of responses to each question[1].

Table 2: Q1: Trust in police to make decisions that are right for people in neighbourhood

Agreement scale versions			Item-specific versions		
Response	**% start**	**% end**	**Response**	**% start**	**% end**
Strongly agree	18	16	Always	27	27
Tend to agree	55	61	Mostly	47	51
Tend to disagree	7	8	Sometimes	13	10
Strongly disagree	4	3	Never	3	2
Agree (NET)	74	77	Top two responses (NET)	74	78
Disagree (NET)	12	11	Bottom two responses (NET)	16	13
Don't know	14	12	Don't know	10	10
Refused	*	*	Refused	-	-
Base	*568*	*568*	*Base*	*525*	*525*

Table 3: Q2: Whether police abuse their power

Agreement scale versions			Item-specific versions		
Response	**% start**	**% end**	**Response**	**% start**	**% end**
Strongly agree	3	2	Always	2	4
Tend to agree	6	6	Often	4	4
Tend to disagree	40	39	Sometimes	24	29
Strongly disagree	39	39	Never	48	44
Agree (NET)	9	9	Top two responses (NET)	6	8
Disagree (NET)	79	78	Bottom two responses (NET)	72	73
Don't know	12	13	Don't know	22	19
Refused	-	1	Refused	-	*
Base	*568*	*568*	*Base*	*525*	*525*

[1] * = less than 0.5%; – (dash) = 0

Table 4: Q3: Whether police reflect mix of people in community

Agreement scale versions			Item-specific versions		
Response	**% start**	**% end**	**Response**	**% start**	**% end**
Strongly agree	14	12	Very well	20	16
Tend to agree	47	52	Fairly well	49	57
Tend to disagree	9	10	Not very well	10	10
Strongly disagree	4	4	Not at all well	1	1
Agree (NET)	61	64	Top two responses (NET)	70	73
Disagree (NET)	13	14	Bottom two responses (NET)	11	10
Don't know	26	20	Don't know	19	16
Refused	*	1	Refused	-	*
Base	*568*	*568*	*Base*	*525*	*525*

Table 5: Q4: Whether police understand issues that affect community

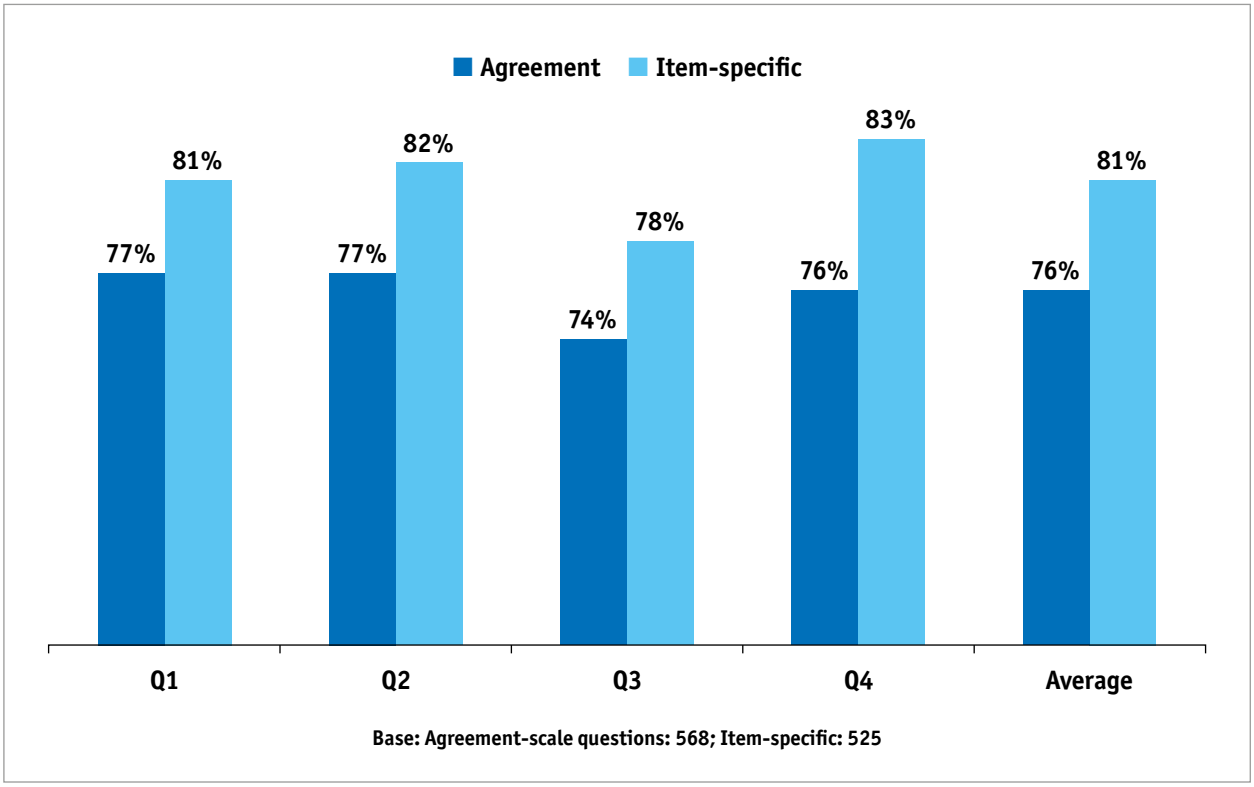
Agreement scale versions			Item-specific versions		
Response	**% start**	**% end**	**Response**	**% start**	**% end**
Strongly agree	18	18	Very well	22	19
Tend to agree	55	56	Fairly well	53	57
Tend to disagree	10	10	Not very well	10	9
Strongly disagree	4	3	Not at all well	1	1
Agree (NET)	74	74	Top two responses (NET)	75	75
Disagree (NET)	13	14	Bottom two responses (NET)	11	10
Don't know	13	12	Don't know	15	14
Refused	-	*	Refused	-	*
Base	*568*	*568*	*Base*	*525*	*525*

There was no consistent evidence to support the first hypothesis. Analysis here has been limited to Q3 and Q4, since the top two and bottom two responses were less comparable for Q1 and Q2 (for example, at Q1 and Q2 'sometimes' is included as a 'bottom 2' response but this response may more closely reflect agreement with the statement in the agreement scale version than disagreement). For Q3, the proportion in the top two boxes was greater for the item-specific version, while for Q4 there was no difference between the two versions. However, the two sets of questions are not entirely 'like-for-like' and so, what the top two response categories represent in one, is not directly comparable with the other.

There was also no clear evidence to fully support the second hypothesis. For two of the agreement-scale questions (Q2 and Q4), the proportion agreeing was unchanged between the start and end of the interview. For the other two questions (Q1 and Q3), the proportion agreeing was higher the second time the questions were asked. However, the differences here were small (three percentage points in both cases), and do not provide strong evidence to support our hypothesis. However, the total interview length was fairly short (around 25 minutes). For future experiments, it would be interesting to repeat questions at the end of a longer interview, to investigate whether this results in a higher level of agreement.

There was, however, evidence to support our third hypothesis. Chart 1 shows the proportion of respondents who provided the identical answer when asked each question at the start and the end of the interview – and the average proportion across the four questions.

Chart 1: Consistency of response for each question

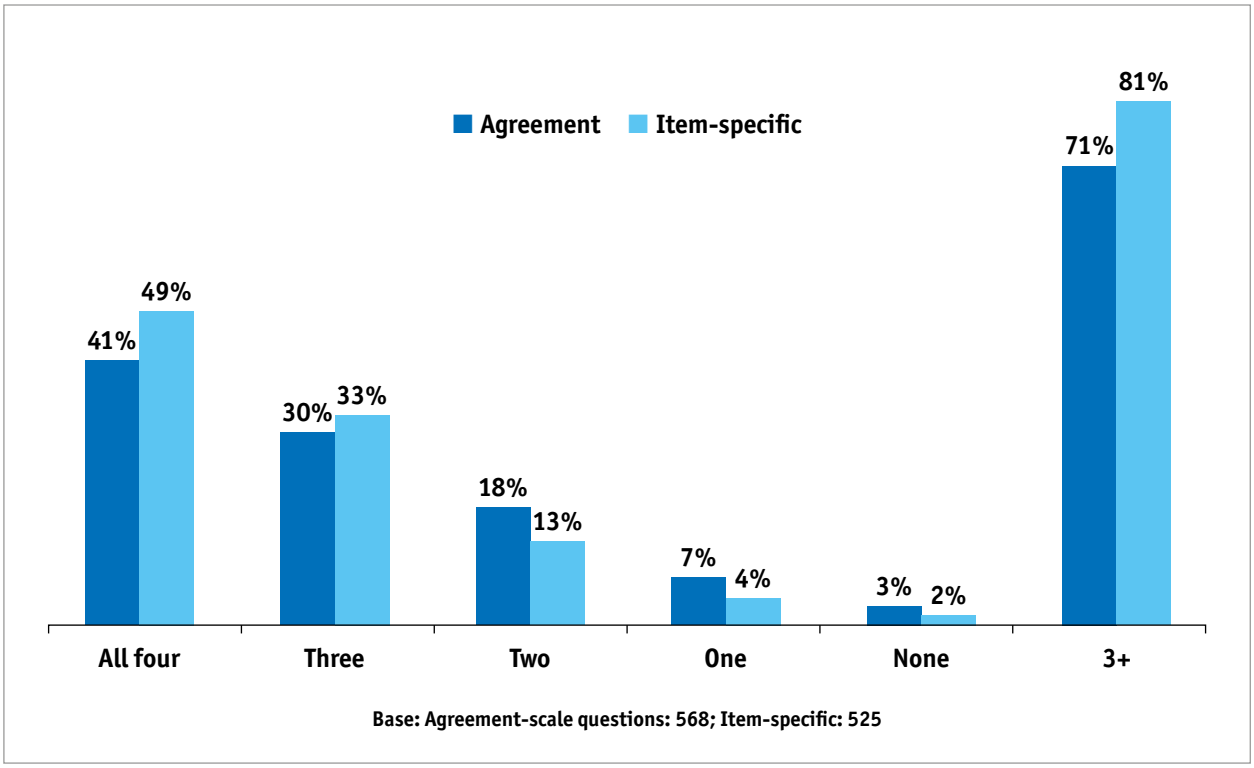


Base: Agreement-scale questions: 568; Item-specific: 525

While there was not a substantial difference in the proportion responding in a consistent way across the two modes, the difference was in the same direction across all four questions, with a greater proportion of respondents providing an identical response to the item-specific questions compared with the agreement-scale questions.

We also looked at the number of questions for which the same response was recorded at the start and end of the interview (chart 2).

Chart 2: Number of questions where response matched

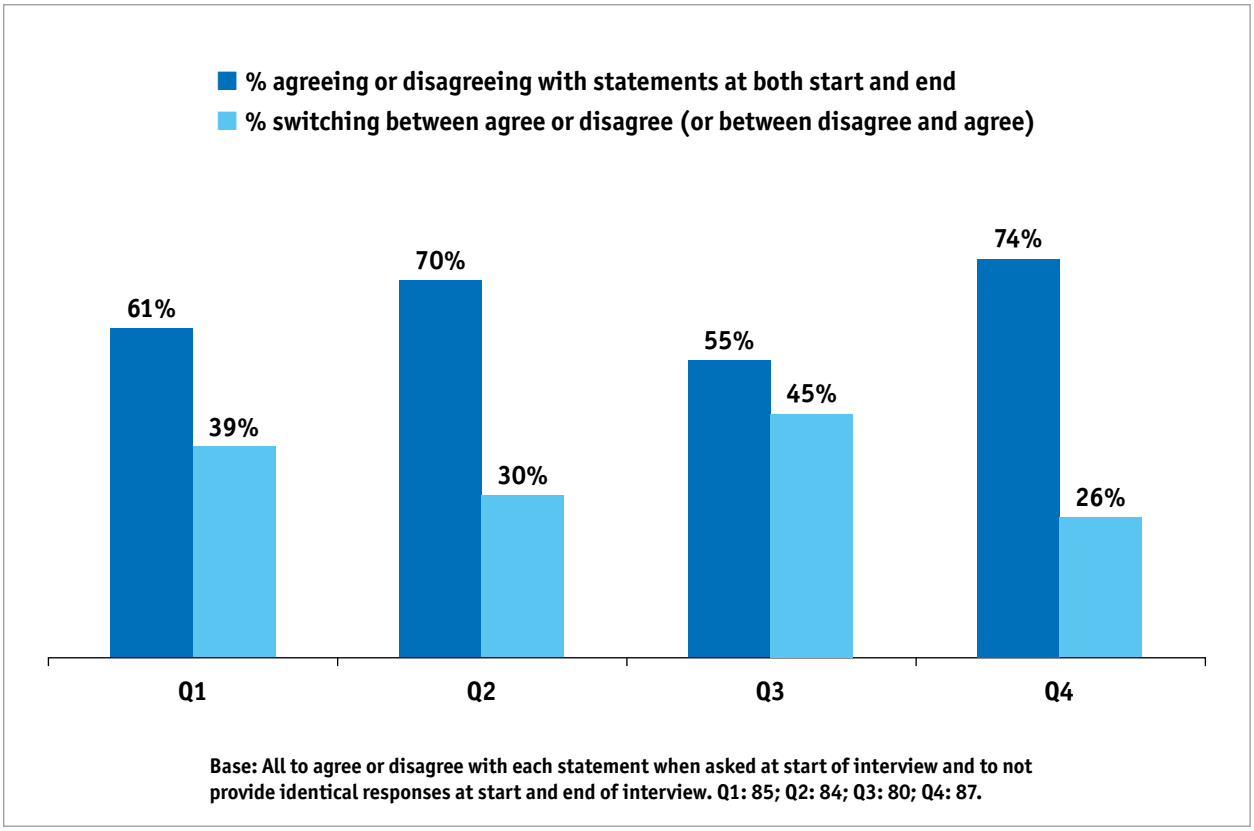


Base: Agreement-scale questions: 568; Item-specific: 525

Overall, 49% answered all four questions consistently where the item-specific approach was used compared with 41% in the case of the agreement-scale approach. Likewise, 81% provided an identical response to at least three of the four item-specific questions compared with 71% for the agreement-scale questions.

As well as looking at whether responses change between the start and end of the interview, it is important to look at **how** responses have changed. This provides greater insight into the extent of the issue. For example, if respondents have moved from 'strongly agree' to 'tend to agree', this may be seen as less of a concern than if they have moved from agreeing with a statement to disagreeing with it, since 'strongly agree' and tend to agree' will often be combined in analysis. Chart 3 shows how respondents who were asked the agreement-scale questions changed response between the first and second statement. This divides respondents into two groups: those who changed their response but still agreed or disagreed with a statement each time they were asked (that is, switched between 'strongly agree' and 'tend to agree' or between 'strongly disagree' and 'tend to disagree'), and those who switched between agreeing and disagreeing with a statement (or vice-versa).

Chart 3: Movement between agreement and disagreement with statements



Base: All to agree or disagree with each statement when asked at start of interview and to not provide identical responses at start and end of interview. Q1: 85; Q2: 84; Q3: 80; Q4: 87.

As chart 3 shows, for each statement, the majority of those who changed their response did not switch from agreeing or disagreeing with the statements when asked at both the start and end of the interview; in most cases they moved between two neighbouring points on the response scale. Nevertheless, a sizeable minority moved from agreeing with the statements to disagreeing with them. This was particularly apparent for statement Q3 ('The police in this area reflect the mix of people in your community.'), for which almost half (45%) of those who changed their answer moved from agreement to disagreement (or vice-versa).

For the item-specific version of Q3 the proportion of respondents who made an equivalent switch (that is from 'very well' or 'fairly well' to 'not very well' or 'not at all well', or vice-versa) was smaller (22%) compared with the agreement-scale version. The proportion switching between the top two codes and bottom two codes (that is between 'agree' and 'disagree' or 'well' and 'not well') at Q4 was similar for both formats (26% for the agreement-scale version and 28% for the item-specific version).

Conclusions and implications

Two broad conclusions can be drawn from these results.

FIRST, in both approaches, a substantial minority of respondents provided different answers to an identical question when asked around 20 minutes apart. The levels were fairly consistent between questions, with around two in ten providing a different answer at each point they were asked each question. Even for the more 'reliable' item-specific approach, only half of respondents provided an identical answer at all four questions each time they were asked. Furthermore, a sizeable minority of those who switched response moved from one side of the scale to the other (for example, from agreeing with a statement to disagreeing with it). This suggests that some respondents may not always consider questions and provide an informed response, and instead may answer 'randomly'. This, in turn, further reinforces the importance of designing questions which are clear and unambiguous and attempting to maintain respondent engagement throughout the interview.

SECOND, the item-specific questions showed greater consistency in response when repeated compared with the agreement-scale questions, suggesting that the item-specific questions provided the more reliable measure, and may illustrate some of the shortcomings of agreement-scale questions highlighted earlier. This supports the findings of the experiments reported by Saris et al (2010), which found that agreement-rating-scale questions had much lower quality than responses to comparable questions offering item-specific response options.

While our experiment was conducted as part of a face-to-face interview, this debate is particularly relevant in a context of increasing movement to online research, with pressure on survey designers to reduce questionnaire length and maintain engagement among respondents. It may be hypothesised that lengthy batteries of agreement-scale statements have a negative effect on respondent engagement and, consequently, the quality of responses when conducting research online. Future studies, therefore, may wish to vary the design of our experiment to compare item-specific and agreement scales in an online context, while also varying the length of batteries, number of response options and question topics, to provide further important evidence in this area. We also recommend including the experiment in a longer survey with a greater time lag between repeat questions, or repeating the questions on separate waves of a survey (days or weeks apart), to assess what effect this may have on the results. Running such experiments online is much easier and cheaper compared with a face-to-face survey, and so this may represent a fruitful area for further investigation.

As already noted, agreement-question scales have been around for a very long time, and it is not expected they will disappear any time soon. Many surveys monitor long-term trends and it may be impractical (and unwise) to change existing questions without a loss of important time-series data. Furthermore, agreement scales are efficient when asking large batteries of statements. Given the substantial information requirements that are often present in fairly short interviews, this cannot be completely ignored. However, this should be balanced against the extra cognitive burden placed upon respondents when using agreement scales and the potential for measurement error that this format may produce.

Therefore, when designing questionnaires in future – and particularly in the rare cases where there is opportunity to design questions from scratch – consideration should be given to using item-specific response scales that directly capture the dimension of interest rather than standard agreement scales. If feasible, researchers may also wish to take advantage of opportunities for improving existing studies by running item-specific questions alongside agreement scales, to both maintain the existing time series and start a new time series based on the improved, item-specific questions.

References

Eifermann, R. R. (1961) 'Negation: A Linguistic Variable.' Acta Psychologica 18: 258-273.

Saris, W. E., Revilla, M., Krosnick, J. A. and Shaeffer, E. M. (2010) 'Comparing Questions with Agree/ Disagree Response Options to Questions with Item-Specific Response Options.' Survey Research Methods 4(1): 61-79.

Schuman, H. and Presser, S. (1981) Questions and answers in attitude surveys: Experiments on question form, wording and context. New York: Academic Press.

Tourangeau, R, Rips, L. J. and Rasinski, K. (2000) The Psychology of Survey Response. Cambridge: Cambridge University Press.

Saris, W. E. and van Meurs, L. (1990). 'Memory effects in MTMM studies.' In W. E. Saris and L. van Meurs (eds.) (1990) Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies. Amsterdam: North Holland: 89-103.

Securing participation and getting accurate answers from teenage children in surveys: lessons from the UK Millennium Cohort Study

Lisa Calderwood, Kate Smith Emily Gilbert and Meghan Rainsberry, Centre for Longitudinal Studies, UCL Institute of Education

Sarah Knibbs and Kirsty Burston, Ipsos MORI

Abstract

This article describes the approach taken to surveying teenagers in the Age 14 Survey of the Millennium Cohort Study. Our key challenges were persuading teenagers to take part, and to give honest and accurate answers, particularly to sensitive questions. Our approach was informed by research with teenagers and their parents. A main motivation for young people taking part in research was that they wanted a 'voice'; to be listened to. Young people and their parents were happy to answer questions on sensitive topics, and understood why these were important to include. Ensuring privacy, and giving teenagers control over their participation was important for encouraging honest answers. We developed several innovations in best practice in research with teenagers, which we anticipate will be of interest to others conducting research with young people.

Introduction

Surveying teenagers in a home setting can be challenging (Levine 1981; Levine 2008). As well as securing their agreement to taking part, it is necessary to ensure that the design and implementation of the survey questionnaire is conducive to obtaining honest and accurate answers. Although this is a not a new challenge, the development of the Age 14 Survey of the UK Millennium Cohort Study (MCS) offered an opportunity to review and develop survey practice in this area. This paper describes the different elements of the strategy adopted to engage the teenage participants in the study, and explains how the questionnaire for the 14-year-old study members was designed and implemented. We anticipate that this will be of interest to other survey practitioners carrying out research with teenage children, particularly in a home setting.

The MCS is one of the British birth cohort studies, following the lives of over 19,000 children born at the turn of the century. Study members have been surveyed five times in the past at key development ages and stages of life: at ages 9 months, 3, 5, 7 and 11 years. The sixth wave, the Age 14 Survey, is taking place in 2015. The MCS is funded by the Economic and Social Research Council and a consortium of UK government departments, and managed by the Centre for Longitudinal Studies (CLS). The fieldwork for the sixth wave is being carried out by Ipsos MORI.

As part of the survey development process, we carried out focus groups, depth interviews and a short survey with 14-year-olds and their parents, including some families who are members of the study. These explored young people's understanding of research; their preferences in modes of data collection and mode for survey

communications; their views about the acceptability and relevance of particular topics for their age group; and their understanding of the consent process. Parents' views were also included, particularly relating to topic relevance and acceptability, and consent. This research informed the development of both the participant engagement strategy and questionnaire design approach on the Age 14 Survey. We refer to some of the key findings in this article. Most of the research reports are available on the CLS website **www.cls.ioe.ac.uk**

Participant engagement

Age 14 presented a particular set of engagement challenges. Fourteen is a transitional age between childhood and adulthood; teenagers are making important decisions independently, particularly around education, which will affect their future. They are also often being given more independence by parents outside school. However, parents remain very important in their lives, and have ongoing responsibilities for them as minors. Additionally, teenagers today are growing up in a fast-paced, digital age, and often lead very busy lives. Evidence from other cohort studies shows that adolescent respondents are likely to have high levels of attrition, and that the transition from parents to young people as primary respondents also often leads to attrition (Boys et al., 2003; Mostafa and Wiggins, 2014). It was, therefore, crucial to ensure that the participant engagement strategy was designed to be appealing and engaging to the study members at age 14, and importantly in the context of a longitudinal survey, also aimed at securing longer-term participation.

Our research found that one of the main motivations for young people in general taking part in research was that they wanted a 'voice' – they wanted to be listened to.

The research with MCS study members found that they were not necessarily aware of the aims of the study, particularly the fact that the study wanted to continue to follow them throughout their adult lives. They knew that the study was important, but did not really understand why. They were particularly interested in the findings from the study, and what difference they make by taking part. They also appreciated the gifts (or 'swag') they received for participating in each survey, and liked their contribution being recognised in this way.

For communication, our research found that the young people liked post, reporting that it was exciting to receive mail addressed to them. Age-relevance was highlighted, with young people saying anything sent directly to them should be written and designed specifically for them rather than for adults. However, many 14-year-olds had a sophisticated view of the content and style of materials, meaning they disliked materials that had been over-designed, or tried too hard to be 'cool'.

As a result of this research, it was decided to 'relaunch' the study to the cohort members themselves, prior to the Age 14 Survey. The aim was to give them important background information about the study to provide the context for their decision to participate at age 14. This focused on some of the key messages to get across to cohort members. These messages highlighted to cohort members the aims of MCS – 'building a picture of your generation' – and the fact that each member of the study was irreplaceable. The leaflet also highlighted the importance of MCS through key findings and policy impact. It made clear that the study is an ongoing longitudinal project covering their 'life story'.

Additionally, the study was rebranded in order to ensure that the materials looked professional, attractive and engaging to 14-year-olds. This included a new logo and visual identity. The rebranding was carried out by a professional design agency, which also conducted focus groups with young people in this age group during the branch development. The branding was applied consistently to all materials families received and on the new study website.

Prior to the Age 14 Survey, all study members were sent a relaunch mailing, in the form of a 'participant pack'. The participant pack was intended to gain buy-in to the study from cohort members themselves. The pack included a letter with a personalised membership card stating that they were a valued member of the study; a booklet with information about the study, including how findings have made a difference and pictures; biographical information about the team working on MCS; and some small, branded gifts – a keyring, a travel-card holder and a notebook. The materials were enclosed in a branded plastic wallet.

As the participant pack did not include any materials for parents, the mailing was addressed solely to the young people. This approach was used because we wanted to engage young people themselves directly, rather than parents. Our research with young people suggested that receiving post addressed directly to them was exciting, and something they liked. This, in part, was because it happened so rarely, so when they received post, they were sure to open it. Additionally, as this mailing was sent purely for information purposes, that is, it did not contain a request for the study members to provide any information, there were no ethical concerns about sending it directly to the young people. The longitudinal study context was also an important factor in this decision – families have been in the study for a long period of time and have a high level of familiarity and trust in it. Overall, the participant pack mailing received a positive response, with several messages received from study members expressing their gratitude for the information and gifts. No concerns were raised about the mailing being sent directly to study members.

As part of the study relaunch, the online presence of MCS was also revamped. The website for study members (**www.childnc.net**) was redeveloped, and included more information about findings, publications and media coverage of MCS. As well as this, a short animated video was commissioned and produced, with the aim of informing study members about the value of the study and their contribution. This was supplemented by social media accounts – a Facebook page and Twitter account – to allow study members to keep up to date with news from the study. The social media accounts are used for information purposes only, and privacy settings have been set to help maintain anonymity of study members. For example, comment functions on Facebook have been disabled where possible, and the Twitter account is protected. The website includes guidance for study members about following the study on social media and, more generally, about staying safe online.

As well as rebranding and relaunching the study prior to the Age 14 Survey, we also wanted to ensure that the survey approach was appropriate and engaging. The main challenge here was to develop a set of survey materials which was attractive, clear and easily understandable to 14-year-olds, but at the same time provided them with sufficient information about what they were being asked to do for them to make an informed decision about participating. This was particularly challenging given the scope and complexity of MCS; young people were asked to complete a 45-minute questionnaire; 20 minutes of cognitive assessments; 10 minutes of physical measurements; and to give a saliva sample for DNA extraction during the home visit. They were also asked to complete a time-use diary and wear an activity monitor for two days following the visit.

The survey advance mailing comprised a letter and information booklet for young people, and the same for parents. All survey materials used the new study brand. As the survey approach included a request to participate, it was important to provide full information to parents and young people about what was being asked of study members so that both parties could make an informed decision about whether or not to take part. As the study members were minors, that is under 16, it was necessary to secure the consent of parents as well as the young people. Parents were also asked to take part in an interview themselves, so their leaflet included information about this. The survey advance mailing was a joint mailing sent to parents and young people at the same time. We wanted to ensure that they had 'equal status' within the mailing; not prioritising parents over young people or the other way around. To achieve this, we put the personalised letter and leaflet for the young person in one envelope, and the personalised letter and leaflet for their parents in another. Both envelopes were then put into a larger envelope addressed jointly to young people and parents.

The survey is carried out face-to-face, so interviewers were also crucial to securing agreement to take part. The interviewer training for the project included specific sessions on engaging young people in the different survey elements and additional guidance was given about carrying out research with children and young people. A set of interviewer FAQs was also developed. No financial incentives were offered to either parents or young people. The study members were given a study-branded USB stick for taking part.

All the materials used in the participant pack mailing and the Age 14 Survey are available on the participant website: **www.childnc.net**

Developing a questionnaire for 14-year-olds

At 14, teenagers are often best placed to tell us about their lives, rather than relying on reports from parents (Jaccard et al, 1998; Fisher et al, 2006). The Age 14 Survey aims to collect a lot of information from study members themselves, relying much less on parental reporting than in previous waves.

Asking teenagers about their lives can necessitate the inclusion of sensitive and personal questions in order to cover everything relevant. In a home setting, it can prove challenging to get honest answers to questions of this nature. Previous research has shown that this age group, more than any others, are less likely to give accurate and honest answers to personal questions if they think someone else may see their answers (De Leeuw, 2011). One of the main challenges for the Age 14 Survey was how to collect information about sensitive and risk-taking activities and get accurate and honest answers in a home environment. This section briefly describes how we designed the young person questionnaire to overcome this challenge.

An important task was to ensure that the content of the questionnaire was appropriate. Fourteen is a significant age in young people's education; teenagers at this age are in transition from childhood to adulthood. It is an age by which some will be experimenting with risk-taking activities such as smoking, drinking and drug use, as well as antisocial behaviour, romantic relationships and sexual experiences. Parents may not be aware that their children are engaging in these activities, so it is important to ask these questions to the young people directly rather than their parents. In addition to these sensitive topics, we also wanted young people to answer a wide range of other questions about their lives. This meant that the young person questionnaire at age 14 was longer than in previous waves: 45 minutes compared with 30 minutes at age 11, and 10 minutes at age 7.

As the MCS is run as a research resource for the academic and policy community, the development of the questionnaire content involved extensive consultation. Key decisions about topic inclusion were driven by scientific and policy needs. However, we also wanted to ensure that the questionnaire was engaging and interesting to young people, and covered the issues that are important and relevant to them at this age. Additionally, we were keen to test the acceptability of the questionnaire topics with parents, particularly for the sensitive topics. Our research found that, although many young people at 14 had not embarked on any of the sensitive activities we wanted to ask about, most of them and their parents thought that it was acceptable to ask questions about these topics as they recognised that some 14-year-olds may be doing these things. Many parents said while their child would not have direct experience of some of the activities, particularly drug taking, smoking and alcohol consumption, they assumed that some other young people of this age would have had these experiences, and, therefore, appreciated that it was important to collect this information from young people. Some topics were found to be less relevant, such as some types of illegal drugs, and were therefore not included. In addition to this research, two pilot studies were carried out to test the length and acceptability of the proposed questionnaire in the spring and summer of 2014. Some questions were cut as a result of this piloting work, for example detailed questions about gambling, which very few young people had experienced.

In addition to determining the content, another key issue was the design and implementation of the questionnaire itself. It was important to carefully consider how to encourage young people to give honest and accurate answers recognising that it was being completed at home.

Given the content and length of the questionnaire, an early decision was taken to use computer-assisted self-interviewing (CASI) for the young person questionnaire. Although this is a commonly used approach in surveys of this age group, a number of the specific design features of the instrument are new and innovative.

As the questionnaire was being implemented in the context of a home visit survey, the young person completed it on the interviewer's tablet in touch-screen mode. A significant benefit of computer-assisted interviewing (CASI) was that it enabled the questionnaire to be filtered depending on an individual study member's experience of an activity, so they were automatically routed around questions about activities they had little or no experience of. This was particularly important for sensitive topics. For example, the

questions on romantic relationships and sexual experiences filter more advanced experiences depending on answers to earlier questions. For example, if a young person has not held hands or hugged another young person, they are not asked if they have kissed another young person.

Using self-administration was a crucial part of encouraging young people to provide honest and accurate answers. As this was implemented using a tablet, the young people could take the questionnaire away with them and complete it in private in another room if they wished. They are reminded throughout the questionnaire that their answers are confidential, and they can skip any questions that they do not want to answer. We included a 'hide screen' button to aid privacy if someone else came into the room, or if they needed to take a break. At the start of questions which are potentially sensitive or about risk-taking behaviour, additional introductory text was included to forewarn them that the next questions might be sensitive, and to encourage them to answer honestly. This text also reminds them that no one else can see their answers, that they can hide the screen at any time and skip any question they do not want to answer. It also flags that that not all young people their age will have done the things that we are asking them about, and it is important that all young people answer honestly. At the end of each section of the questionnaire the young people 'lock' their answers to that section, which means they cannot be accessed – even by them, or by the interviewers. Given the length of the questionnaire, we also included a topic-based progress bar on the side of the screen to give them visual feedback about where they were in the questionnaire and to help motivate them to complete the whole questionnaire.

We also needed to consider the protocol for implementation in a home setting. In particular, as the young people are still children, parents are still gatekeepers, and we needed parents' informed consent for the interviewer to approach the young people for their consent. This meant balancing the needs of giving parents enough information to gain their consent, but not giving them too many details about the actual questions we were asking. As described earlier, both parents and young people are given information in advance to explain what we want them to do and how long the survey will take, to ensure that their consent is informed. The information booklets for young people and parents briefly outlined the topics included in the young person questionnaire. If parents wanted more information about what the questions would cover, the interviewers provided a show card with a fuller list of topics. However, there was no provision for allowing parents to see the exact questions. Parents gave written consent for the interviewer to approach their child. Young people gave verbal consent using a structured consent form, which was read out and signed by the interviewer. Consent from parents and young people was obtained and recorded separately for each element of the study. Although this general protocol is a commonly-used one in surveys of this age group, specific aspects of it, such as the use of a formalised process and structured form to gain consent from the young people, and the method used for providing information to parents about the content of the questionnaire, represent innovations in best practice.

In addition to gaining informed consent, another important ethical consideration was protecting the wellbeing of the participants. This was particularly important as the questionnaire included topics which could be upsetting to the study members. As part of our duty of care, we wanted to provide study members with advice on where to get support if they were upset or distressed by anything, or if the questions raised concerns for them. We developed an additional 'further information' leaflet which was given to the young person by the interviewer at the end of the visit which signposted who they should speak to if they were worried about anything. It advised them to speak to their parents, teachers or other adults, and gave contact details for age-appropriate sources of support: ChildLine, Get Connected and Talk to Frank. Again, this is not a new idea; rather an example of the application of good practice in the MCS context.

In summary, the design and implementation of questionnaires for young people need to give careful consideration to mode of implementation. In determining the content, survey practitioners should consider the length, relevance of topics, range of questions, sensitivity of topics and acceptability. Ethical considerations are important: balancing informed consent and the right of young people to privacy over their answers, as well as the wellbeing of participants.

Conclusions

This paper has described the approach taken on the MCS to securing participation and getting accurate answers from 14-year-old study members. It was informed by extensive research with teenagers and their parents. We feel that our approach is a strong example of best practice in surveying teenagers, and that several of the specific innovations developed on the MCS Age 14 Survey are significant improvements to good practice. We anticipate that this will be of interest to other survey practitioners carrying out research with teenage children, particularly in a home setting.

References

Boys, A. Marsden, Stillwell, J. G., Hatchings, K. , Griffiths, P. and Farrell, M. (2003) 'Minimizing respondent attrition in longitudinal research: Practical implications from a cohort of adolescent drinking.' Journal of Adolescence 26(3): 363-373.

De Leeuw, E. (2011, May) 'Improving Data Quality when Surveying Children and Adolescents: Cognitive and Social Development and its Role in Questionnaire Construction and Pretesting.' Report prepared for the Academy of Finland: Research programs public health challenges and health and welfare of children and young people.

Fisher, S. L, Bucholz, K. K., Reich, W., Fox, L., Kuperman, S., Kramer, J., Hesselbrock, V., Dick, D. M., Nurnberger, J. I., Edenberg, H. J. and Bierut, L. J. (2006) 'Teenagers are right – parents do not know much: An analysis of adolescent-parent agreement on reports of adolescent substance use, abuse, and dependence.' Alcoholism: Clinical and Experimental Research 30(10): 1699-1710.

Jaccard, J, Dittus, P. J. and Gordon, V. V. (1998) 'Parent-adolescent congruency in reports of adolescent sexual behavior and in communications about sexual behaviour.' Child Development 69(1): 247-261.

Levine, C. (1981) 'Commentary: teenagers, research and family involvement' IRB Ethics and Human Research 3(9): 8.

Levine, R. J. (2008) 'Research involving adolescents as subjects.' Annals of the New York Academy of Sciences 1135(1): 280-286.

Mostafa, T. and Wiggins, R. D. (2014) Handling attrition and non-response in the 1970 British Cohort Study: CLS Working Paper. London: Institute of Education, University of London.

Seeking impact for research on policy tsars

Dr Ruth Levitt and William Solesbury, visiting senior research fellows, King's College London

Abstract

This case study concerns research on identifying and characterising the work of policy tsars which was undertaken with a clear ambition to bring improvements to their role in Whitehall policy development. To that end, a programme of communication and contact with interested parties was undertaken both during and subsequent to completion of the research; in the second phase of work a draft code of practice for tsars was prepared, similar to those for other kinds of policy advisers. Yet the impact of the research was limited. It is concluded, from this case study, that major impact requires interested and powerful policy actors to use the evidence from researchers to develop their argument for desired change.

Background

In the SRA's guidelines about what constitutes high-quality social research, two of the criteria concern research being 'useful' and 'useable.' To be useful 'good research should have some practical relevance'; to be useable 'research outputs should be readily actionable without too much further interpretation and translation' (SRA, 2015). We believe, therefore, that researchers' methodological concerns should be as much with the communication as with the conduct of their research. This paper is a case study reflecting upon our experience of seeking to maximise the impact of research which we undertook on the work of policy tsars.

Over the last two decades, the media have dubbed as 'tsars' a succession of outsiders appointed by UK Government ministers to advise them on policy matters. It is an odd term, with implications of executive authority, which these appointees certainly do not have. In practice, on appointment, these people were given one of many other titles including reviewer, champion, representative or advocate. Some attracted much publicity. One such was Helen Newlove, whose husband had been murdered in 2007 when confronting drunken youths; she subsequently campaigned against binge drinking; was given a peerage; and was then appointed by Home Secretary, Theresa May, in 2010 as 'champion for active, safer communities'. Another was Lord Browne, former chief executive of BP, who was appointed by Peter Mandelson, trade and industry secretary at the time, to review higher education funding, and whose recommendations for higher student fees were adopted by the UK Coalition Government in 2010. And there was Mary Portas, TV shopping guru, commissioned by David Cameron and Nick Clegg in 2011 to advise on 'the future of the high street'.

Arising from our previous research on the interplay of evidence and policy[1], we decided in 2010 to investigate this type of appointment. There had been no previous research on this specific source of advice to ministers. We defined a 'tsar' as 'an individual from outside government (though not necessarily from outside politics) who is publicly appointed by a government minister to advise on policy development or delivery on the basis of their expertise'. We decided to focus on appointments which ministers in Whitehall departments had made between 1997, when the new Labour administration took office, and 2012. We adopted as our research question: 'How does the development of policy and practice by government benefit from Whitehall's use of tsars?'

[1] Earlier work had focused on the contribution of outsiders coming to work in Whitehall (Levitt and Solesbury, 2006) and on the use of evidence in audit, inspection and scrutiny work (Levitt et al, 2010).

The research

We conducted the research partly with assistance from six students who were studying for a Masters in Public Policy at King's College London. It comprised several strands. Drawing on various online sources such as press releases, Hansard, media reports and on our own preliminary interviews, discussions and on freedom of information (FOI) requests we:

- Identified over 260 such appointments between 1997 and mid-2012
- Created a profile for each appointment including the remit, their background, client minister, work content and timetable, payment, outputs and outcomes

We then conducted interviews with 16 tsars and 24 of the colleagues, ministers and officials with whom they worked, to explore their experience of the role.

We also discussed our work in progress with academics, researchers, commentators, individuals and organisations interested in Whitehall policymaking, including a seminar at the Institute for Government.

A full account of our research and findings is at: **www.kcl.ac.uk/sspp/departments/politicaleconomy/research/Current-Research-Projects/tsars.aspx**

Our ambition from the outset was to seek impact for the research. This was strengthened by our initial discovery that the tsar phenomenon had not hitherto been documented in government, let alone researched, and even more so by what our research progressively revealed about the scope and significance of this source of advice. The debate over the period since 1997 on evidence-based policy (to which the New Labour mantra 'what matters is what works' gave expression) provided one context. Another was the UK Government's commitment to 'open policy making' in its June 2012 Civil Service Reform Plan, which stated:

> *'Whitehall has a virtual monopoly on policy development, which means that policy is often drawn up on the basis of too narrow a range of inputs and is not subject to external challenge prior to announcement … the need to maintain a safe space for policy advice should not be used to prevent the maximum possible openness to new thinking or in the gathering of evidence and insight from external experts' (HM Government, 2012: 14).*

This apparent commitment to a bigger role for external expertise in policy making seemed to provide a favourable practical context for our research, even if the use of tsars as one kind of external expert, alongside academics, think tanks and consultants, had not hitherto been systematically recognised in Whitehall departments.

What emerged from our research was revealing. Our key findings:

- Since 1997, the rate of such appointments had accelerated: from three to 11 to 26 a year in Labour's three terms, to 43 a year in the coalition's first two years; these appointments have continued to be made since then
- Some ministers seemed keener than others: Gordon Brown as chancellor holds the record with 23 appointments, although Ed Balls, Alastair Darling and Michael Gove with 11 appointments each were also enthusiasts. The last two Prime Ministers have been particularly busy: only five tsar appointments by Blair, but 23 by Brown and 21 by Cameron
- Tsars address very diverse policy issues: strategic (for example Andrew Dilnot on social care funding), or operational (Tom Winsor on police pay and conditions), perennial (Sir Alan Steer and others on school behaviour) or topical (Richard Brown on rail franchising), a government priority (Alan Milburn on social mobility) or a minister's enthusiasm (Tony Hall on dance education)
- Ministers appoint tsars quite informally; there is no standard practice. A name is identified usually because the minister knows or knows about them; an official or special adviser or the minister phones; then they meet informally and agree the broad terms of the remit. No advertisement, open competition or tendering. Nor acceptance that these are public appointments and subject to the rules and procedures that govern them

- Tsars' career backgrounds vary: private sector business is most common (40% of appointments); public service and civil service (often retirees) are close (37%); researchers (mostly academics, few from think tanks or consultancies) next (23%); before politicians (18%), serving and retired, including several ex-ministers (some tsars have dual characteristics so figures do not add to 100%)
- Their expertise varies: some are specialists in the field in which they advise; some are generalists relying on their experience and knowledge to bring an 'open mind' to the topic; some are already known advocates for a particular course of action
- Tsars are strikingly un-diverse: predominantly male (85%), white (98%), aged over 50 (71%), and 38% have titles (lords, baronesses, sirs and dames)
- Some are paid fees and/or expenses, others not – in our interviews the latter were often surprised to learn of the former
- Usually tsars are given administrative and analytical support from civil servants. Research methods may include reviews of past work, stakeholder consultations, visits, private discussions with experts, statistical analysis. Some have advisory groups
- The typical timescale for their work is 6 to 12 months
- Most tsars produce final reports, some of which are publicly acknowledged by their client minister and published. But for a sixth of appointments, we could find no such report; there might have been an oral briefing. And for 5% of tsars, there was no evidence whatsoever in the public domain of what they had done, and our FOI requests for such information were either rejected or produced uninformative responses; usually after long delays
- Our impression is that there has been a positive outcome to the majority of tsars' work: in about 40% of cases a policy change can be traced to it; in 40% of cases a practice change; and in 20% of cases an organisational change – in some cases more than one of these categories of outcome. But there is a residue of cases with no discernible output, which we speculate may be for various reasons: little work was done by the tsar; the advice was rejected; agendas and political priorities had changed; the commissioning minister or administration had gone by the time work was complete
- There are no departmental or central records of these appointments. And there has been no formal – or even informal – evaluation of what tsars achieve individually or collectively

Our overall conclusion was that the appointment of tsars as a source of external advice to ministers (alongside such others as research and consultancy, lobbying, advisory committees) has grown in importance; is not documented; is operated at the discretion of ministers without any guidance on good practice or rules to ensure propriety; costs time and money; and has produced variable results. Above all, in comparison with other external sources of advice to ministers (such as advisory committees, consultancies and inquiries), there are no clear rules and procedures for the use of tsars as external advisers.

To promote the use of our research, we made four practical recommendations with actions in pursuit of each:

1. Ensure that a tsar appointment is the most appropriate source of expert advice for the particular subject and be clear what personal attributes are required of an appointee to achieve success
2. Make a 'contract' between the client minister or department and the tsar to agree the remit, timetable, civil service support, payment and reporting
3. Ensure transparency regarding the appointment of the tsar, the output of the tsar's work and the minister's response; for example departmental annual reports could record activity related to the appointment, processes and outcomes of tsars' work
4. Identify and promulgate good practice in the recruitment, conduct and management of tsars; identify a senior official in each department given clear overall responsibility for overseeing and guiding such appointments and assessing their value; and the Cabinet Office should prepare a code of practice for tsar appointments

Publicity and promotion[2]

Right at the start of our research, we publicised the project informally through contacts in the practice world, and formally through an article in Public Money and Management. We continued our informal contacts as work progressed.

In late 2012, we actively publicised and promoted the findings, conclusions and recommendations in our final report. We used only one academic occasion: a paper to the September 2012 Policy and Politics conference. Otherwise, we intentionally sought out print and broadcast media which reached politicians and civil servants. We:

- Secured coverage (sometimes writing ourselves, sometimes briefing journalists) in Civil Service World, the parliamentarians' magazine, The House, in Prospect and in Public Money and Management
- Briefed political correspondents and had broadcast coverage on BBC Radio 4's The Westminster Hour and BBC1's The One Show, being interviewed in both cases
- Issued a press release and had news coverage in both the Daily Telegraph and The Independent, including a supportive editorial in The Independent headlined 'Our governance has yet to enter the 21st century'[3]
- Made a podcast for the King's College London website **www.kcl.ac.uk**
- Guest-blogged on the LSE's British policy and politics blog, on NESTA's blog, and the research was covered on the Institute for Government blog
- Gave talks about the research and its findings at the Constitution Unit, to staff in the House of Commons and to the Politics Society at KCL

We also promoted our research results direct to key players in Whitehall and Westminster. We obtained meetings to discuss our work with:

- The Cabinet Office teams dealing with 'open policy making' and with public appointments
- The Commissioner for Public Appointments
- The chair and staff of the House of Commons Public Administration Select Committee (PASC) to whom we subsequently submitted a memorandum for their inquiry on The Future of the Civil Service
- The secretary of the Civil Service Commission
- The newly-appointed chair of the Committee on Standards in Public Life

We contacted others seeking a response but without success: cabinet secretary Sir Jeremy Heywood, the civil service head of policy profession, the shadow minister for the Cabinet Office, the general secretary of the First Division Association (the trades union for senior civil servants), the Whitehall and Industry Group, the Fawcett Society (which campaigns for gender equality) and Sir Christopher Foster's 'Better Government Initiative'. All in all, though we felt that we had successfully found an audience and secured attention for our research.

While most people we contacted expressed astonishment at what we had uncovered about the scale, scope and practices of tsar appointments, it became clear that none of them was sufficiently concerned about the weaknesses in ministerial use of tsars as sources of external advice to embrace our analysis and adopt our recommendations. The Cabinet Office claimed its existing arrangements for such appointments (which are limited to potential conflicts of interest in the appointments) were adequate. The Commissioner for Public Appointments argued that bringing the large number of tsar appointments within his remit would

[2] Details of our publicity and promotion activities are at: **www.kcl.ac.uk/sspp/departments/politicaleconomy/research/Current-Research-Projects/tsars.aspx**

[3] 18 September 2012 (**www.independent.co.uk/voices/editorials/editorial-our-governance-has-yet-to-enter-the-21st-century-8145118.html**)

overwhelm his work programme. Some civil service sources claimed that, because tsar appointments tend to be relatively short, they do not count as public appointments at all. Select committees indicated that the issues we raised were not a high priority for investigation by them, given their already busy work agendas. Above all, we sensed a widespread view that the informality of tsar appointments was an important part of their attraction to ministers and to some appointees, and the alternatives might serve ministers and appointees much less well. This view disregards many of the principles which supposedly underpin modern governance, such as making public appointments on merit, through processes which are open and fair, reflecting diversity, ensuring value for money, transparency in policy making and the use of evaluation to determine good practice.

So, we decided to take our work a stage further, to make it not just useful but also useable in the SRA's distinction quoted above. If the Cabinet Office or others declined to introduce a code of practice for the work of tsars along the lines that we had recommended, then we would prepare one for them. We obtained a small grant from King's College London to undertake this next impact-generating phase of the work. To this end we:

- Examined a number of existing codes of practice for external advisers of other kinds, such as special political advisers, scientific advisory committees, appointments to public bodies, to see what format and scope they had
- Converted our research recommendations into a code format with an accompanying commentary
- Consulted our former contacts and interviewees on how appropriate and useful such a code would have been to their work as tsars
- Held an event to launch our draft code and issued the code and a press statement to a targeted list

The event was held at King's College in October 2013. Invitations went to all our previous contacts in the media, government and academe. About 30 people attended. We gave a brief presentation about the code's format and scope, and then three former tsars spoke about their experiences. They were Otto Thoresen, director-general of the Association of British Insurers, who in 2007 had reviewed generic financial advice which led to the creation of the Money Advice Service; Dame Stephanie Shirley, who had been appointed by PM Gordon Brown in 2008 as an ambassador for corporate philanthropy; and Professor John Hills of LSE who had had three tsar appointments reviewing equality and social housing for Labour ministers and fuel poverty for the UK Coalition Government. They all expressed a belief that a formalised code would have helped them in their work. Bernard Jenkin MP, chair of Public Administration Select Committee, attended and spoke supportively.

As a result we:

- Secured considerable media coverage, with reports on BBC News and in The Guardian (which also interviewed John Hills and took an article from us for its Public Manager column), Daily Telegraph, Daily Mail, The Independent and Evening Standard
- Blogged for Democratic Audit, the LSE British Policy and Politics (again), The Conversation, the UCL Constitution Unit, Lib Dem Voice and Whitehall Watch
- Met with the chair of the Public Accounts Committee and with National Audit Office officials to seek to engage their interest in the poor practices in the appointment and management of tsars that our code sought to address
- Wrote seeking a meeting with the newly appointed director general for Civil Service Reform in the Cabinet Office but received no reply

Despite these efforts to make our research useable as well as useful, it stimulated awareness but no action.

Impact

We estimate our efforts at publicising and promoting the research took 50% of the time we spent on undertaking the research. That is far more than most researchers do. We worked hard at relating our research results to the practical concerns of different audiences. We honed our skills at writing journalistically, blogging and giving interviews. We learned to respond to media wishes to have anecdotes about individual tsars without compromising our proper commitment to objectivity and confidentiality. Tsar appointments have continued since we completed our data collection in mid-2012 in the same way that we revealed, criticised and sought to reform. Could we have done more, or differently, in our search for impact?

Research impact itself has been extensively researched over the last two decades. How did our practice match up to what is recommended? Take as a benchmark the ESRC's current advice on 'how to maximise impact' (ESRC, n.d.). It offers a number of 'key factors' which it relates to:

- Process: 'impact works best if you can tap into pre-existing networks and relationships with research users.'
- Context: 'the environment in which you communicate your messages has a bearing on any potential impact... an awareness of policy and practice debates and initiatives will help you to time your work most effectively to achieve the best end results.'
- Content: 'the extent to which the content of your research fits with the context in which it is disseminated will have a bearing on its capacity to generate impact.'

The publicity and promotion we undertook for our research met these tests. We contextualised our work in the apparent commitment to 'open policy making' and a greater contribution from 'external experts'. We focused our efforts quite directly on Whitehall and Westminster and the media (print, broadcast and social media) which relate to them – rather than on academic outlets and audiences[4]. And we sought, with the drafting of the code of practice, to make the results of our research not just useful but useable.

There is a view that changes in practice – like most innovations – occur over longer timescales than is generally assumed. Certainly our research has, for the first time, revealed the role of policy tsars as part of the advisory processes which can shape public policy. Drawing on our work, recent commentary on processes and structures in Whitehall departments by some journalists and academics has acknowledged their existence. It may be that, in time, demands will arise for reform. This could well be if and when the appointment or conduct or results of a tsar's work become contentious. Over the period we researched in detail, there had been a few causes célèbres: Emma Harrison's resignation as Families Champion in 2012 after her organisation A4e became subject to fraud investigations; the embarrassment when it was revealed that the entrepreneur James Caan, appointed in 2013 by Nick Clegg to promote equal opportunities in the workplace, had given his daughter a job in one of his companies; Mary Portas being given a tough time at a select committee about the relationship of her tsar appointment to her TV and consultancy work. Perhaps it needs more of such cases to demonstrate more forcibly that the informality and intimacy of tsar appointments lack appropriate degrees of transparency and accountability.

We also draw another, related, conclusion. In a 1995 paper, the late great Carol Weiss argued that the forces shaping public policy could be characterised memorably as the four Is – ideology, interests, information and institutions (Weiss, 1995). Research evidence is one kind of information, and it finds itself in competition with other kinds of information and with political ideologies, interests (essentially self-interests) and the culture of institutions. The practices of tsars as external advisers are modified by the competing forces of ideologies, interests and institutions. Above all, the power of interests in shaping changes in practice is not just to be thought of in corporate terms – for example, as lobbyists. There can also be interested individuals. What we needed for our research to achieve greater impact was such an individual; someone who had the position and motivation to take our evidence and analysis about tsars and get something done about it – in the face of the self-interested resistance of ministers and the complicity of officials in that.

[4] In recent months, we have submitted evidence to a review of the work of the commissioner for public appointments being undertaken by a tsar: Sir Gerry Grimstone who has had two such earlier appointments.

We conclude from this case that, as researchers, we can act as very effective irritants and drivers, putting sound evidence and argument into the public domain, targeting it at key actors, and thereby informing the understanding of and debates about public issues, policies and practices. Yet this can prove insufficient to achieve desired change. Here the implementation of the changes, which we believe our research justifies, still depends on harnessing the commitment and power – in short, clout – of a political actor with something personal to gain.

References

ESRC (n.d.). How to maximise impact. [online] Available at: **www.esrc.ac.uk/research/evaluation-and-impact/how-to-maximise-impact/** [Last accessed 23/10/15].

HM Government (2012, June) The Civil Service Reform Plan. London: HM Government. [pdf] Available at: **www.gov.uk/government/uploads/system/uploads/attachment_data/file/305148/Civil-Service-Reform-Plan-final.pdf** [Last accessed 23/10/15].

Levitt, R., Martin, S., Nutley, S. and Solesbury, W. (2010) Evidence for accountability: Using evidence in the audit, inspection and scrutiny of UK government. London: Nuffield Foundation. [pdf] Available at: **www.nuffieldfoundation.org/sites/default/files/files/Evidence_for_accountability_web_PDF.pdf** [Last accessed 23/10/15].

Levitt, R., and Solesbury, W. (2006) 'Outsiders in Whitehall.' Public Money and Management 26(1): 10–12.

SRA (2015) What is high quality social research? [pdf] Available at: **http://the-sra.org.uk/wp-content/uploads/what-is-high-quality-social-research.pdf** [Last accessed 23/10/15].

Weiss, C. (1995) 'The Four 'I's of School Reform: how interests, ideology, information and institutions affect teachers and principals.' Harvard Educational Review 65(4).

The Social Research Association (SRA)
24-32 Stephenson Way
London NW1 2HX

0207 998 0304
admin@the-sra.org.uk
www.the-sra.org.uk

@TheSRAOrg

the-sra.org.uk/journal-social-research-practice

Copy-edited by Shirley Henderson www.shirleyhenderson.co.uk Design by www.graphics.coop