Social Research Association

Issue 6
Summer 2018

Social Research Practice

The SRA journal for methods in applied social research

Social Research Practice

Issue 6 Summer 2018

The Social Research Association journal for methods in applied social research

Contents

Editorial

Richard Bartholomew

Editor

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

Web-based panels have become the predominant method for conducting surveys in market research as well as some areas of social research, but their dominance poses major challenges for maintaining the quality 'gold standard' of social surveys: random probability sampling and, primarily, face-to-face interviewing. The speed and low cost of web-based panel surveys can be highly seductive for those who commission research and want data quickly but who understand less about how sampling has a crucial effect on the reliability of that data. The challenge for social research is to find ways of reconciling the speed and lower cost offered by web-based panels with the quality provided by random probability sampling. Curtis Jessop describes how NatCen has developed one way of doing this by establishing a research panel recruited via the British Social Attitudes Survey. He explores the problems of attrition and maintaining response in repeat panels, and assesses the effect this has on overall representativeness.

It is vital that the public, and especially employers, parents and young people, have confidence in qualifications and the system for awarding and quality assuring those qualifications. But what are the key factors which drive that confidence and how do they inter-relate? This is the challenge faced by Qualifications Wales, the independent regulator of qualifications in Wales. To explore ways of understanding the relative importance of the different drivers, Vasile Rotaru and his colleagues have investigated the potential of the Comparative Judgement method first developed by psychologists in the 1920s. Vasile describes how this approach has been trialled as a way of assessing the relative influence of different statements and concepts on people's confidence in qualifications and in the qualification system. Those who have ever tried to influence people's views will not be surprised by the conclusion that it is rather easier to judge the effect of negative statements than of positive ones.

In this issue we publish the first of our new series of Research Notes. These provide a format for social researchers to discuss aspects of work in progress or lessons learnt. Sally Gowland of BBC Media Action describes the challenges of conducting qualitative research inside war-torn Syria in order to understand the lives and perspectives of audiences for a new radio drama. This must be one of the most extreme and testing environments in which social research has been conducted: unstable and often dangerous situations with potentially extreme risks faced by both researchers and participants. Sally's note provides some useful pointers on how to conduct research in even the most difficult situations of conflict and uncertainty.

Many researchers who have contemplated conducting studies involving NHS patients, staff or organisations, even if this is not central to the purpose of their research, will have found the processes for gaining ethical and governance approvals rather daunting. In our second Research Note, Hannah Hartley and Emma Bolton provide practical guidance on how to navigate the requirements of the new combined NHS and Health Research Authority (HRA) processes, based on their own recent experience. Their note provides an invaluable source of reference on what you need to do.

I hope you enjoy reading this issue. We welcome proposals for new articles or shorter **Research Notes**. Our next issue will be published in December 2018. If you are interested in offering a Research Note or a full article please visit the dedicated page of the SRA website **http://the-sra.org.uk/journal-social-research-practice/**

If you have an idea for an article but are not sure if will be suitable, just drop me a line: **rabartholomew@btinternet.com**

The NatCen Panel: developing an open probability-based mixed-mode panel in Great Britain

Curtis Jessop, Research Director, Survey Research Centre, NatCen Social Research

Abstract

The NatCen Panel is the only probability-based research panel in Great Britain open for the social research community to use. It currently consists of approximately 6,000 panellists aged 18+ recruited from the British Social Attitudes survey. Panellists are invited to take part in surveys online or on the phone approximately once every one to two months.

This paper describes the development of the NatCen Panel, outlining the rationale for setting it up, its standard methodology, and information on response rates and sample representativeness, and how these have changed as the panel matures.

Research context

In Great Britain, the 'gold standard' for social research surveys has been to collect data face-to-face, with participants selected using random sampling methods. However, while this approach produces relatively high response rates, it is also comparatively slow and expensive. As a result, this methodology is not always appropriate for a given project – budgets may not stretch far enough, or data may be required more quickly in order to respond to unfolding events.

If a face-to-face probability design is not appropriate, telephone random-digit dialling (RDD) or non-probability web panels can provide alternatives. However, there are question marks over the quality of these approaches (see Yeager et al, 2011). Already typically lower than face-to-face surveys, declining response rates for telephone surveys (Nelson, 2012) risk reducing the quality of their samples, while the increasing need to combine mobile with landline samples has increased costs. The use of quotas – as opposed to random samples with call backs – also increases the risk of bias. With web panels, the use of self-selecting samples in addition to quotas and short fieldwork periods can also lead to biases, while web-only fieldwork excludes a sizeable minority of the population whose experience of society is very different to the rest of the population.

The development of an open, probability-based, mixed-mode panel aims to address the limitations of face-to-face probability-based fieldwork, without compromising quality. Web-first fieldwork employed in a standardised manner, provides efficiencies that enable data to be delivered quicker and at lower cost than face-to-face designs. The random probability design, coupled with the use of telephone fieldwork to include the offline population, reduces the risk of bias and allows the panel to maintain quality. This is an approach that has been successfully implemented internationally[1] and explored as a possibility in the UK[2], but never put into practice.

[1] For example, the GESIS panel in Germany (**www.gesis.org/en/gesis-panel/gesis-panel-home/**) or the AmeriSpeak panel in the USA (**https://amerispeak.norc.org/**).

[2] **http://gtr.ukri.org/projects?ref=ES%2FM010031%2F1**

Standard methodology

This section outlines the 'standard' methodology which has been used by the NatCen Panel since May 2016. In practice, there have been some deviations from this – for example due to the requirements of a particular study, or as part of an experiment designed to test improvements to efficiency/quality of the design – but, for the most part, the method has remained consistent over time.

'Piggy-back' recruitment

The approach to recruitment for the panel needed to meet the following criteria:

- A random-probability design to avoid biases of self-selection and 'convenience' samples and allow legitimate statistical inferences to be made to the wider population
- High recruitment rates to reduce the risk of non-response bias and provide a larger sample size for analysis of sub-groups
- Low costs to keep the panel accessible to a wider range of researchers

Fresh recruitment was considered but a face-to-face approach was judged to be too costly and it was anticipated that web, paper or telephone recruitment would result in low recruitment rates (and poor protocol compliance in self-completion modes), undermining the sample quality.

A 'piggy-back' approach (that is, recruiting participants from an existing study) was, therefore, decided upon: panellists were recruited from participants in the British Social Attitudes (BSA) survey.[3] The BSA is a high-quality, random-probability face-to-face survey of people aged 18+ in Great Britain.[4] By recruiting from the BSA survey, we are able to:

- Maintain a probability-based recruitment design
- Achieve high recruitment rates
- Keep recruitment costs low as the only costs were the marginal costs of asking additional questions
- Obtain a large quantity of background information about the panellists
- Ensure that recruitment protocols (for example, the selection of a random adult in the household) were followed

People interviewed for the BSA survey were asked to join the panel at the end of their BSA interview, and those who agreed were asked to confirm their contact details. Once recruited, all panellists were sent a letter confirming that they had joined the panel, with an information leaflet providing more detailed information about what taking part would involve.

Sampling

BSA participants who agreed to join the panel, and have not subsequently dropped out (either by asking to leave or becoming ineligible) are considered 'available panellists'. Under the standard design, all available panellists are invited to take part in each new wave of the survey, sustaining the principle that the population has a known and non-zero chance of being selected, and thus the random probability design.[5]

[3] This approach had recently been used by the Pew Research Centre, recruiting their American Trends Panel from the Political Polarization and Typology RDD survey (Pew Research Centre, 2015).

[4] More about the BSA survey at: **www.bsa.natcen.ac.uk/**

[5] Unless we are trying to boost the sample size, or conduct longitudinal analysis, we typically invite only panellists recruited from the most recent two waves of BSA. We may invite a random or targeted sub-sample, for example to reduce the sample size or reach a particular demographic group which we can identify in advance. In these situations, the random probability design is still maintained for the target population.

Fieldwork

The NatCen Panel employs a 'sequential mixed-mode' fieldwork design, lasting slightly over four weeks. To keep things simple to administer, and avoid confusing panellists, only one wave of fieldwork is run at a time. Eight waves are scheduled approximately 1.5 weeks apart throughout the year, though there is some flexibility in precisely when they take place.

At the start of fieldwork, all available panel members are sent a letter and email inviting them to take part in the research online. These include a link to the web survey and a unique log-in code to access the questionnaire. Panel members are offered a conditional £5 incentive to thank them for their time. If they do not take part after receiving the invitation, panel members are sent a reminder letter, two reminder emails, and two reminder text messages to encourage them to take part.

After two weeks, all available panel members who have not taken part in the survey online, and for whom we have a phone number, are issued to the NatCen Telephone Unit to follow-up by phone and either support them to take part online or complete an interview over the phone. Telephone fieldwork lasts for a little over two weeks to allow telephone interviewers to make a minimum of six attempts to contact and interview all panel members at a time that suits them.

By employing this sequential mixed-mode design with a four-week fieldwork period, this approach aims to balance quality and efficiency:

- By issuing all cases to web first, and employing multiple reminders in multiple modes, we maximise the number of cases completing online, minimising interviewer costs
- A four-week fieldwork period allows people time to take part (minimising bias towards the 'readily available'), while providing data in a timely fashion
- Telephone fieldwork boosts response rates, and allows those who are not comfortable with, or do not have access to, the internet to take part
- Standardised systems mean approximately eight weeks from agreeing a set questions to delivering a clean weighted dataset

Weighting

As with all surveys, surveys conducted using the NatCen Panel are subject to a degree of non-response bias, due to the fact that those who choose to take part tend to differ from those who do not. Non-response can occur at three stages in the process: 1) refusal to take part in the recruitment survey, 2) refusal to join the panel, and 3) refusal to take part in a survey issued to panel members (including through attrition).

To account for this, weights are computed to adjust the sample. The final weight is the product of three separate weights: the BSA weight and two further non-response weights, each of which is designed to adjust for non-response at the three stages described above. The standard BSA weight is itself a composite weight: a weight to account for unequal selection probabilities; a non-response weight to adjust for household level non-response; and calibration to populations estimates of sex, age and region. The stage two and three weights employ the wealth of data available from the BSA survey, and have been standardised since February 2017.[6] These data allow non-response at these stages to be modelled effectively, thereby accounting for much of the bias that may otherwise remain unadjusted for.

[6] The standard models include: age grouped within sex; government office region; household type; household income quartiles; highest educational qualification obtained; political party identification; internet access; respondents' economic activity; ethnicity; housing tenure; NC-SEC analytic class; interest in politics; BSA year.

Response rates and sample quality

Recruitment

The existing NatCen Panel has been recruited from three waves of the BSA survey – 2015, 2016 and 2017. Overall, 60% of people who took part in these BSA surveys agreed to join the panel. However, slightly different wording has been used for the invitation question at each wave and this appears to have affected recruitment rates. In 2015, a random half of BSA participants were asked if they would be willing to be contacted for 'follow-up studies' and the other half were asked for consent for us to contact them 'as part of a research panel'; the latter approach was repeated in 2016. In 2017, the wording was simplified, and changed to be more engaging.

BSA participants were more likely to agree to be contacted 'for follow-up studies' than as 'part of a research panel', while the wording changes to the recruitment question in 2017 appeared to improve recruitment rates (Table 1). The recruitment question wording also appears to have affected levels of attrition and non-response to particular panel surveys, thereby reducing (but not removing) the overall difference between the invite approaches in terms of response (Table 2).

Table 1: recruitment rates to the NatCen Panel by sample group

	BSA 2015			
	Follow-up studies	Join the panel	BSA 2016	BSA 2017
Completed BSA	2188	2140	2942	3988
Recruited to the panel	1734	1049	1422	2580
Recruitment rate	79%	49%	48%	65%

The reason for continuing with an approach which elicited lower overall response rates was that the lower recruitment rates resulted in a more efficient panel: having to send fewer invite/reminder letters and issue fewer cases to telephone interviewers reduced its costs while having only minimal impact on the sample profile. Indeed, Table 8 suggests that, despite the lower response rates, asking people to join the panel may **improve** the sample profile (that is make it look more like the population). We also judged it to be more thorough in ensuring informed consent, and we therefore maintained this approach, while also adapting to the wording to try to increase recruitment rates in 2017.

Response rates

Whilst the response rate to any individual survey wave may be useful for monitoring fieldwork, an 'overall response rate' using a base of all eligible cases issued to the original BSA interview and, therefore, accounting for initial (non)participation in the recruitment survey,[7] recruitment rates, attrition and response at a particular wave is a more accurate indicator of the true level of non-response.[8]

The overall response rates for surveys conducted using the panel vary from wave to wave as the questionnaires themselves change (for example in length or topic), and as we experiment with different elements of the fieldwork design (for example the number or timing of reminder communications). Broadly, however, when using the standard design, the overall response rate has been around 15%; a little higher than this for those agreeing to be contacted for follow-up studies than as part of a research panel (Table 2).

[7] The response rates to the BSA 2015, 2016 and 2017 surveys were 51%, 46% and 45% respectively.

[8] We would normally also adjust for cases that become ineligible (for example through death or leaving the country), but, for simplicity, as these numbers are very small, we have not done this here.

Table 2: overall response rates across waves by sample group

Wave	BSA 2015		BSA 2016	BSA 2017	TOTAL
	Follow-up studies	Join the panel			
May 16	22%	17%	–	–	19%
Sep 16	18%	15%	–	–	16%
Nov 16	19%	15%	14%	–	16%
Feb 17	19%	15%	14%	–	16%
Mar 17	18%	15%	14%	–	15%
May 17	18%	14%	13%	–	15%
Jul 17	18%	14%	13%	–	15%
Aug 17	17%	14%	13%	–	14%
Oct 17	17%	14%	13%	–	15%
Nov 17	–	–	13%	16%	14%
Jan 18	18%	14%	13%	–	15%
Base	*4,292*	*4,197*	*6,408*	*8,781*	*Varies*[9]

Attrition

Although the overall response rates are broadly stable, Table 2 shows a gradual downward trend for each sample group over time. As the proportion of available panellists taking part does not show a similar pattern, we can infer that this drop in the overall response rates is due to attrition – people asking to leave the panel, or leaving as they become ineligible[10] – and a resulting decline in the number of people being invited to take part.

Table 3 tracks the levels of attrition over time for the different sample groups. As outlined above, it shows that it is higher for those agreeing to be contacted for follow-up studies than as part of a research panel. The rates are also higher than the gradual decline in overall response rates might suggest. This is because the majority (currently 58% of those recruited from BSA '15 or '16) of those who have dropped out did not take part in any panel waves, so by leaving the panel they did not affect the overall response rate.

[9] The total response rate for a particular wave of panel fieldwork is calculated on a base of all eligible cases issued to the BSA wave(s) from which sample issued at that wave was recruited. For example, in July 17, the base will have been 14,897 (4,292 + 4,197 + 6,408).

[10] For example, due to death or leaving the country.

Table 3: attrition rates across waves by sample group

Wave	BSA 2015		BSA 2016
	Follow-up studies	Join the panel	
May 16	6%	3%	–
Sep 16	8%	4%	–
Nov 16	11%	6%	0%
Feb 17	13%	8%	2%
Mar 17	15%	9%	3%
May 17	17%	10%	5%
Jul 17	19%	12%	7%
Aug 17	20%	13%	7%
Oct 17	21%	14%	8%
Nov 17	–	–	10%
Jan 18	22%	15%	10%
Base	*1,734*	*1,049*	*1,422*

Table 3 also shows that attrition rates have continued to increase, even for sample recruited over two years ago, although the rate of increase has slowed. This is associated with a gradual decline in sample representativeness (Table 8), which, along with the increased potential for panel conditioning, and the fact that the sample is ageing, has led us to decide that our standard design will typically only use sample from the latest two waves of the BSA survey (that is currently those recruited from BSA 2016 and BSA 2017). However, the BSA 2015 sample has not been 'dropped' altogether; despite ongoing attrition, its overall response rates and DEFF are similar to the more recent sample groups (Table 8), so it is still used as a 'boost' sample when larger sample sizes are required.

Longitudinal response rates

As well as retaining the BSA 2015 sample as a boost sample where required, it has been maintained for longitudinal research. Although the NatCen Panel was developed for cross-sectional studies, as it tracks individuals over time its design is akin to a longitudinal study, and longitudinal analysis is possible. However, to do so, we need sufficient numbers of people taking part in all waves of interest, to give statistical power for analysis and ensure estimates are representative of the population.

As with cross-sectional response rates, re-interview rates between panel waves vary from survey to survey. Considering the seven waves conducted between November 2016 and October 2017 (waves using sample recruited from BSA 2015 and 2016), between 82% and 92% of panellists taking part in any one wave took part in any one other (Table 4), whilst 66% of panellists interviewed in November 2016 took part in all seven waves, suggesting longitudinal analysis is feasible with the NatCen Panel design.

Table 4: proportion of panellists interviewed at a wave who were also interviewed at another wave

	Nov 16	Feb 17	Mar 17	May 17	Jul 17	Aug 17	Oct 17
	Col %	Col %	Col %	Col %	Col %	Col %	Col %
Nov 16	100%	90%	90%	89%	90%	90%	89%
Feb 17	88%	100%	91%	91%	91%	90%	90%
Mar 17	87%	90%	100%	92%	92%	92%	91%
May 17	84%	87%	89%	100%	91%	92%	91%
Jul 17	82%	86%	88%	90%	100%	92%	91%
Aug 17	82%	84%	86%	89%	91%	100%	92%
Oct 17	82%	84%	86%	89%	90%	92%	100%
Base	*2,375*	*2,322*	*2,290*	*2,223*	*2,184*	*2,159*	*2,168*

Impact of telephone fieldwork

The telephone fieldwork stage of the NatCen Panel standard design aims to improve sample quality by encouraging participants who are less engaged to take part, and by enabling those who are not comfortable with, or who do not have access to, the internet to take part. The proportion of interviews conducted on the phone varies between around 13% and 22% from wave to wave, contributing approximately three percentage points to the overall response rate.

There is some indication that the contribution of telephone fieldwork to the overall response rate is declining as the sample matures (Table 5). This was initially thought to be due to panellists moving online as they provided contact details (and were, therefore, more likely to receive reminders) and became more used to web fieldwork processes. However, we do not see a corresponding increase in the contribution of web fieldwork to the overall response rate. An alternative explanation is that this is due to ongoing attrition, with those 'less engaged' panellists who were being picked up by telephone fieldwork also being the ones to leave the panel.

Table 5: overall response rates by mode of interview across waves by sample type

Wave	2015				2016		2017		TOTAL	
	Follow-up studies		Join the panel							
	Phone	Web	Phone	Web	Phone	Web	Phone	Web	Phone	Web
May 16	7%	14%	4%	13%					6%	14%
Sep 16	4%	14%	3%	12%					3%	13%
Nov 16	4%	15%	3%	12%	3%	11%			3%	12%
Feb 17	4%	15%	3%	12%	3%	11%			3%	12%
Mar 17	3%	15%	2%	13%	3%	11%			3%	13%
May 17	3%	15%	2%	12%	2%	11%			2%	13%
Jul 17	3%	14%	2%	12%	2%	11%			3%	12%
Aug 17	2%	15%	2%	12%	2%	11%			2%	13%
Oct 17	2%	15%	2%	12%	2%	11%			2%	13%
Nov 17					2%	11%	3%	12%	3%	12%
Jan 18	3%	15%	2%	12%	2%	11%			2%	13%
Base	*4,292*		*4,197*		*6,408*		*8,781*		*Varies*	

Overall, 10% of those recruited to the NatCen Panel reported in their BSA interview that they did not personally have access to the internet. This group is different from the rest of the population in a range of demographic, behavioural and attitudinal measures, so giving them the opportunity to take part is important to minimise bias in the survey sample. Table 6 shows that despite the use of telephone fieldwork, the overall response rate for this group is lower than for those with personal access to the internet. However, it also shows that telephone fieldwork disproportionately increases the response rates for this group. Around 60% of interviews conducted with people who reported not having personal access to the internet are completed over the phone, compared with around 15% of the rest of the population.

Table 6: overall response rates by mode of interview across waves by whether the participant reported personally having access to the internet

	Web access			No web access		
	Total	Phone	Web	Total	Phone	Web
May 16	20%	5%	15%	13%	9%	4%
Sep 16	17%	3%	15%	11%	7%	4%
Nov 16	17%	3%	14%	11%	8%	4%
Feb 17	16%	3%	13%	10%	7%	4%
Mar 17	16%	2%	14%	10%	6%	4%
May 17	16%	2%	14%	10%	6%	4%
Jul 17	15%	2%	13%	10%	6%	4%
Aug 17	15%	2%	14%	9%	5%	4%
Oct 17	15%	1%	14%	10%	6%	4%
Nov 17	15%	2%	12%	12%	7%	4%
Jan 18	15%	2%	14%	10%	6%	4%

Sample profile

Although response rates are often used as a proxy for sample quality (assuming that as response rates decrease, non-response bias increases), they do not indicate the degree or nature of bias in the underlying sample. Table 7 shows how the unweighted sample from the November 2016 wave of the panel compares to the unweighted BSA sample and the weighted BSA population estimates.

There are a number of places where the unweighted panel survey sample differs from population estimates. People who took part in the November 2016 panel survey were more likely to be women, older, in managerial and professional occupations, have a degree, live in a single-person household, and own their own home. For the most part, these reflect the biases often seen in survey samples, and also reflect the bias found in the original BSA survey (achieved) sample. In some instances, the bias has been reduced (for example household type). However, there are also a few instances where the additional non-response adds to the bias seen in the demographic profile – the biases in social grade, levels of education, and tenure seen in the panel survey sample are not seen to the same degree in the BSA survey sample.

Table 7: sociodemographic profile of November 2016 panel sample compared to unweighted BSA sample and populations estimates

	BSA 15/16 population estimate (weighted)	BSA 15/16 sample (unweighted)	November 2016 panel survey sample (unweighted)
Sex			
Male	49%	44%	44%
Female	51%	56%	56%
Age			
18-24	12%	6%	5%
25-34	17%	15%	13%
35-44	16%	17%	18%
45-54	18%	18%	21%
55-64	14%	17%	19%
65+	22%	28%	24%
Region			
North East	4%	5%	4%
North West	11%	13%	11%
Yorkshire and The Humber	9%	8%	8%
East Midlands	7%	9%	9%
West Midlands	9%	9%	8%
East of England	10%	9%	10%
London	13%	10%	9%
South East	14%	15%	16%
South West	9%	9%	10%
Wales	5%	5%	5%
Scotland	9%	8%	8%

Social classification			
Managerial and professional occupations	40%	39%	50%
Intermediate occupations	13%	13%	14%
Employers in small org; own account workers	9%	9%	9%
Lower supervisory and technical occupations	9%	9%	8%
Semi-routine and routine occupations	29%	29%	20%
Not categorised	1%	0%	0%
Highest level of education			
Degree or above	24%	23%	32%
Higher education below degree	11%	11%	14%
A level or equivalent	19%	17%	19%
O level/CSE or equivalent	26%	26%	24%
Foreign or other	3%	2%	2%
No qualification	17%	21%	9%
Household type			
Single person household	17%	29%	25%
Lone parent	4%	6%	7%
2 adults (no children)	37%	34%	36%
2 adults (with children)	21%	18%	20%
3+ adults (no children)	15%	8%	8%
3+ adults (with children)	7%	4%	4%
Other	0%	0%	0%
Economic activity			
In full-time education/training	5%	3%	2%
In work, waiting to take up work	56%	52%	56%
Unemployed	5%	4%	4%
Retired	23%	29%	26%
Other	11%	12%	11%

Tenure			
Owned/being bought	64%	64%	71%
Rented (LA)	10%	11%	7%
Rented (HA/Trust/New Town)	7%	8%	7%
Rented (other)	18%	16%	15%
Other	1%	1%	1%

As the weighting approach used for the NatCen Panel includes background BSA variables, the differences between the unweighted panel survey sample and the weighted BSA population can be effectively summarised by its design effect (DEFF), with a larger DEFF indicating that the weights have had to do more 'work' to make the unweighted sample look like the population. The single figure DEFF calculations presented in Table 8 below are based on the final weight for each wave, summarising all the variables used in the modelling. This allows us to look at the changes in the sample profile as a whole without looking at many individual variables (although it should be noted that this isn't their purpose).

Table 8 summarises the DEFF for each of the sample groups across waves of panel fieldwork. In some respects, these reflect the patterns shown in Table 2: each of the sample groups show a gradual increase in DEFF across waves, suggesting a slight decline in sample representativeness.[11] However, it differs in other respects: the overall response rates for those asked for permission to be contacted for follow-up studies were consistently higher than those asked for consent to be contacted as part of a research panel, suggesting a **more** representative sample. In contrast, their DEFFs tend to be slightly higher, suggesting a **less** representative sample. Similarly, at the January 2018 wave, the BSA 2017 sample had both a higher overall response rate and DEFF than the BSA 2016 sample.

Table 8: DEFF due to weighting across waves by sample group

	BSA 2015				
	Follow-up studies	Join the panel	BSA 2016	BSA 2017	TOTAL
Feb 17	1.8	1.7	1.7	–	1.7
Mar 17	1.8	1.7	1.7	–	1.8
May 17	1.9	2.0	1.9	–	1.9
Jul 17	2.1	2.0	1.9	–	2.0
Aug 17	2.0	2.0	1.8	-	1.9
Oct 17	1.9	1.9	1.8	–	1.8
Nov 17	–	–	1.7	2.0	1.9
Jan 18	2.0	1.9	1.9	–	1.9

[11] This seems to stop increasing, and slightly decrease after the July 2017 wave. From August 2017 onwards we implemented a 'targeted design' approach aimed at improving the sample profile which may explain this, though more analysis is needed.

Conclusions

As the fieldwork costs associated with face-to-face (or RDD) probability samples have increased, and questions have been raised over the quality of non-probability alternatives, probability-based research panels have become increasingly popular internationally, with several examples in the USA and Europe. However, the high set-up costs associated with a design that would elicit a high quality sample has prevented one being set up in the UK. By employing a 'piggy-back' recruitment approach, the NatCen Panel is able to greatly reduce these costs.

This paper outlines the methodology used by the NatCen Panel, and provides some simple metrics to evaluate its quality and how it has changed over time. Our analysis suggests that the approach is effective at achieving its goal of collecting survey data in a time- and cost-efficient manner, while maintaining quality at an appropriate level. While response rates are lower than face-to-face surveys, they are comparable to alternative probability-based approaches, with the added benefit of being able to use background data to model non-response effectively. In terms of cost and speed, a 15-minute survey of 2,000 people would currently cost approximately £60,000, with less than two months between questionnaire content being agreed and the production of a clean, weighted dataset – considerably quicker and cheaper than existing probability-based approaches.

Next steps

We will continue to monitor and evaluate the quality of the panel sample. For example, we are currently undergoing a programme of work to compare estimates produced by the panel to external benchmarks. By developing an adjusted version of R-indicators[12] which will use background BSA data to measure the similarity between the survey sample and respondents, we will also be better able to summarise the sample profile and evaluate experiments intended to further improve the fieldwork design and sample quality, such as different incentive or communications strategies.

A key area of further work will be to look at the measurement effects associated with the design. For example, using a mixed-mode design is effective at producing a more representative sample as it encourages the less engaged to take part and includes those uncomfortable with using the internet, as well as those unable to. However, it also means that the design is susceptible to mode effects as participants may answer questions differently online and over the phone. Experimentation is required to understand the impact these mode effects may have, and how they contribute to total survey error relative to sampling effects.

Finally, the re-interview rates outlined above indicate that the panel may be a valuable resource for projects looking at short-term longitudinal change. A small number of projects have taken advantage of this feature, but we are keen to explore this use further. As well as traditional cross-sectional and longitudinal surveys, the panel has been used for question testing, experimental designs, and recruitment to qualitative studies, and we intend to continue to explore new ways to make the most of this new infrastructure.

References

Nelson, E. (2012). 'Great Britain'. In Häder, S., Häder, M. and Kühne, M. (eds.) Telephone surveys in Europe: research and practice. (p47-58). Heidelberg: Springer.

Pew Research Center. (2015). Building Pew Research Center's American Trends Panel. [online] Available at: **http://assets.pewresearch.org/wp-content/uploads/sites/12/2015/04/2015-04-08_building-the-ATP_FINAL.pdf**. [Accessed 5/7/2018].

Schouten, B., Morren, M., Bethlehem, J., Shlomo, N. and Skinner, C. (2009). How to use R-indicators? [online] Available at: **http://hummedia.manchester.ac.uk/institutes/cmist/risq/RISQ-Deliverable-3.pdf** [Accessed 5/7/2018].

Yeager, D., Krosnick, J., Chang, L., Javitz, H., Levendusky, M., Simpser, A. and Wang, R. (2001). 'Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples'. Public Opinion Quarterly 75(4): 709-747.

[12] Developed by Schouten et al (2009).

Using comparative judgement to explore drivers behind confidence in qualifications and the qualification system

Vasile Rotaru, Senior Research Officer, Qualifications Wales

Abstract

The article describes a pilot study conducted to test the usability of the comparative judgement (CJ) method to explore public confidence in qualifications. Participants were presented with pairs of statements, and were asked to choose from each pair, the statement that had the higher impact on their confidence. An examination of the participants' responses indicates that, aside from measuring the (comparative) value of objects, CJ can also be used to explore or compare concepts. The article signposts some of the methodological aspects that need to be considered when employing CJ in this way. Substantive findings are used to clarify the potential of CJ in this regard.

Introduction

The research reported in this paper took place at Qualifications Wales (QW), the independent regulator of qualifications in Wales. One of the two principal aims of QW is promoting public confidence in qualifications and in the qualification system. A key methodological issue in fulfilling this aim is that public confidence is complex and hard to pinpoint.

First, confidence could include multiple aspects such as trust, satisfaction and fairness. Therefore, relying either on a global measure or only on certain aspects of confidence can be risky in relation to validity (see Cowell et al, 2012). Second, perceptions of the education system could be influenced by the public's satisfaction with other government institutions, as people may amalgamate their judgements into one overall perception of how society is functioning (Loveless, 1995:16). Finally, different groups could emphasise different, and often conflicting, factors when thinking about confidence in the education system (Wayson et al, 1988:13). Thus, viewing public confidence as a homogeneous concept might be counterproductive. On the other hand, the public are likely to share basic values about the outcomes of schooling and, therefore, might hold a common, aggregate sentiment about the value of qualifications (Achilles et al 1989, Wayson et al, 1988:13) which also needs to be considered.

This study aimed to test whether the comparative judgement (CJ) method could be used to explore and better understand public confidence. The CJ method was chosen given its increasing use in educational contexts, which is, in itself, of interest to QW, but also because, if proven efficient, it promises a quick and reliable way of identifying the drivers of public confidence in qualifications and how these drivers interrelate. This, in turn, is important for designing an efficient strategy for monitoring and fostering public confidence.

Comparative judgement method

Comparative judgement, adaptive comparative judgement and paired comparison are some of the terms most often used to describe the method used in this study. We use the term comparative judgement in this paper.

The CJ method was developed by the psychologist Louis Thurstone who formulated the law of comparative judgement (Thurstone, 1927). He assumed that each time a person compares a pair of stimuli, each stimulus gets placed on a scale or continuum by a 'discriminal process'. This is a 'kind of process in us by which we react

differently' to stimuli and categorise them according to degrees of possessing a certain attribute. By placing the stimulus on the continuum, we assign it an instant value. This is not an exact value, but by repeatedly comparing a stimulus, we get a set of values which form a normal distribution. The mean value of the distribution is the approximate value of the stimulus (Thurstone, 1927). This value can be used to rank order the stimulus in relation to other stimuli.

In a CJ exercise, a group of judges is asked to compare pairs of objects drawn from the same set in terms of an attribute. This could be any 'quantitative or qualitative attribute about which we can think 'more' or 'less' for each specimen' (Thurstone, 1927), which loosely could also mean 'better', 'greater', 'more appealing' and so on. In an educational context, for example teachers, instead of assigning a score to each essay, would be asked to compare and decide in subsequent rounds which of two students' essays is better.

After each comparison, the results are fitted into a statistical model and several standardised estimates for each object are produced (Jones and Iglis, 2015). This study used the No More Marking (NMM) online platform[1] which employs the Bradley and Terry model (Bradley and Terry, 1952) to analyse the preference response data.[2] One of the estimates produced by the model is the expected score for each object that predicts the outcomes of the comparisons that this object is involved in. This score is calculated iteratively and is based on the actual number of times the object won or lost (Whitehouse and Pollitt, 2012). The Bradley-Terry model calculates a strength parameter and, using its value, determines the probability that the object will be preferred when compared to another object (Whelan, 2017). Using these estimates, each object is assigned a rank order from the one that has the least of the attribute of interest to the one that has the most of it (Jones and Alcock, 2014).

CJ uses an established psychological principle that 'the relativity of judgement (...) is a feature of absolute judgements' (Cohen, 1937:100) since '[a]ll judgements are comparisons of one thing with another' (Laming, 2003:9). When teachers, for example, grade essays they would still compare students' performance to a certain assessment criterion or reference.

The relativity of judgement has been confirmed in many studies ranging from evaluating physical stimuli to making decisions (see Laming, 2003). The advantage of CJ is that it makes comparison easier and more reliable by requiring that the same feature is compared in two objects, instead of asking people to measure the feature in isolation (Jones and Alcock, 2014), that is, to compare it to a reference criterion. For instance, it is easier to determine which of the two weights is heavier than to state their approximate weight (Jones and Iglis, 2015). The advantages of this approach become even more evident when no (apparent) objective measure exists, such as judging colour intensity.

This leads to another benefit of CJ: removing the effect of subjective measurement standards (Pollitt and Crisp, 2004). For example, if a person is rating the intensity of two colours as 7 and 8 while another as 5 and 6, when comparing these colours there is a good chance that both would say that the second colour is more intense.

Finally, due to statistical modelling of the results, the differences between the expected and observed outcomes of each comparative judgement could be used as a quality-control tool of the judging process. Big deviations from expected outputs (either a judge makes an unexpected decision, or an object receives inconsistent values) – called misfit – could be evaluated and dealt with accordingly (see Pollitt, 2012; Whitehouse and Pollitt, 2012).

[1] 'No More Marking' is an online platform that offers a free tool to implement CJ. The tool covers the whole process: from uploading the scripts to be judged and a facility to invite judges to inbuilt software that analyses the results. See more at **https://www.nomoremarking.com/**

[2] The model is a special case of the Rash model for dichotomous data (Agresti, 1992). It postulates that the probability of preference for a certain alternative is proportional to some underlying unobservable utility or quality. Using the Bradley and Terry model one can obtain the estimates of these qualities.

A full CJ implementation requires that judges compare every possible pair (Thurstone, 1927). As this can be time consuming even with a relatively low number of objects,[3] in practice the Adaptive Comparative Judgement (ACJ), which provides similar results[4] at a fraction of the number of all possible pairs, is used. This approach avoids comparing objects that are further apart (that is an object which almost certainly has a greater value than the other) as this does not add useful information for the ranking purposes. As most of the time we do not have any information about an object's value, the pairs are initially picked at random. Once the judging process provides some information about the value of the objects, an algorithm guides the selection of pairs (for example pair only the objects that won or lost in the preceding round). With each round of this adaptive selection, better object separation occurs to the extent that reliable results can be achieved with a relatively reduced number of comparisons (see Pollitt, 2012).

NMM suggests that one needs roughly ten judgements per object to achieve a reliability of 0.8 (that is for 20 objects one needs 200 judgements).

The study

In the study, our 'objects' were short statements (15 to 50 words each) containing either a real or a hypothetical fact or opinion about qualifications, but also about the state of the education, as confidence in qualifications depends on public opinions about the larger education context. The statements were mainly based on mass-media headings or QW press releases. Each statement, depending on the expected impact on public confidence, was categorised as positive, negative or neutral. The positive and negative statements reflected areas that previous research[5] has indicated to be important for driving public confidence in education or qualifications: educational policy, attainment, school environment (facilities, resources, safety), quality of education and people's experience of interacting with the system. Neutral statements contained scenarios that, depending on the view taken, could be categorised as either negative or positive, or had no apparent impact on the qualifications.

Table 1: examples of statements used in the study

Positive	Negative	Neutral
More than £4 million is to be invested to establish a new national network of excellence for teaching science and technology to raise standards in Welsh schools.	The school in your community is forced to cut back on trips and clubs and to axe some GCSE and A-level subjects, due to funding.	GCSEs, AS and A levels in Wales, England and Northern Ireland are changing. In future, there will be differences in most school subjects taught in these three countries.
Parents at your local school feel that their children are safe in their learning environment.	The first results for the new Wales-only maths GCSEs have been revealed – and most students got a D or below.	Head teachers and examination officers across Wales are being invited to attend a one-day event in Cardiff looking at the current thinking around assessment.
The number of appeals about the GCSE exam results that led to a change in the marks fell by 40%.	Most adults in Wales are not confident in the quality of GCSE qualifications, according to the results of a study published by Qualifications Wales, the independent regulator of non-degree qualifications.	Qualifications Wales, the independent regulator of non-degree qualifications, is in the process of reforming several GCSEs, A and AS levels for teaching in the nation's schools and colleges.

[3] The number of pairs being n*(n-1)/2, where n is the number of objects.

[4] But see below the discussion on ACJ inflating reliability coefficients.

[5] The domains were selected based on findings from the *Confidence in qualifications and the qualification system in Wales*, Qualifications Wales report, March 2017 and *Views on education in Wales: re-contact survey*, Welsh Government report, November, 2016.

There were 33 positive, 37 negative and 21 neutral statements. To compare the impact of negative and positive news on public confidence, 27 statements in the negative and positive groups were exact opposites.

The QW staff (70 people) were randomly assigned to judge either statements from positive and neutral groups or statements from the negative and neutral groups. Each judge had to answer the question: 'What has a greater positive (or, for the second group, negative) impact on your confidence in qualifications in Wales?' To compare the similarity of the confidence in qualifications and confidence in the qualification system, judges from each group also completed a separate CJ task evaluating the impact of statements from the same set but on the qualification system.

Each judge had to compare 15 pairs for each task. Eventually we had 25 judges comparing positive and neutral statements, and 33 judges comparing negative and neutral statements. This resulted in a lower number of comparisons than the recommended ten. Nonetheless, the achieved reliability (defined as a measure of how consistent judges were) was relatively high (ranging from 0.62 to 0.83) although in three of the four tasks, the reliability coefficient was still lower than the usual >0.8 values achieved in other CJ exercises (for an overview see Bramley, 2015). Judgements took place over a period of two weeks. Although judges could have accomplished their tasks in separate rounds, most participants opted to complete both tasks in one session.

Findings

This study was different from the usual use of CJ in that it explored a concept, rather than simply comparing and ranking a set of objects. The purpose of this exercise was not to generate substantive findings but to trial the method before using it more widely. This section describes some methodological aspects in using CJ in these types of inquiries and uses the substantive findings only to clarify the potential of CJ.

Statements

Ensuring that the objects/features to be compared are interpreted in broadly similar ways is not something one would usually need to worry about in a typical CJ exercise. But, unlike the typical CJ in which objects contain all the information necessary to make a comparison, in this study many statements were merely representations of what people could be exposed to in real life. In addition, many of the statements compared in this study were not real. This was partly due to our aim of testing whether it was possible to draw any conclusions about the impact of statements which were identical in content but opposite in meaning (for which the opposite statements needed to be created) and partly because we could not identify news stories that would cover all areas of interest.

Consequently, we needed to ensure that statements were interpreted broadly along the same lines and that participants would have as few 'it depends' type reactions as possible when comparing the impact of a statement. We checked the initial set of statements with some QW staff for wording, relevance and plausibility. We then recruited 12 members of the public to explore their understanding of the updated statements, using an approach similar to cognitive testing of surveys. We also collected feedback from participants after CJ was completed. Judging and testing was always based on voluntary informed consent.

Throughout this process, it became clear that providing sufficient detail was important to allow judges to discriminate between statements. People asked us to name the surveys mentioned and who ran them, to indicate the baseline when an increase or decrease in a figure was reported, to name universities mentioned in statements, and so on.

Having relatively short statements was particularly relevant for our exercise as we assumed that in real life people would most of the time be exposed to short headlines rather than more detailed accounts. The challenge, when adding plenty of detail, was to ensure that this assumption was not breached, and that the statement's length remained reasonable, as extended statements could require more time for judgement and increase attrition. Also, as more detail is added, it becomes harder to identify the driver behind people's decisions. One way to attempt to solve this problem (untested in this research) is to vary the same statement by the type of detail and to investigate how this affects the results.

Another issue that became apparent was that some statements, especially neutral ones, were perceived by the judges to have no apparent relevance to the question asked.[6] Such statements seemed to have negatively affected CJ because judges became confused about the purpose of the exercise. In retrospect, informing judges that, to better explore the concept, a broad range of statements was included, might have been useful. In addition, some of the created statements seemed too implausible, which made judges question the accuracy of the exercise. In the guidelines, we warned judges about some statements not being real and asked them, when encountering such statements, to assume that they were true. However, the feedback collected post CJ indicated that judges still felt uneasy when judging those statements. So, if possible, we recommend avoiding using such statements and, if you have to do this, to exclude those that seem too unrealistic.

Preparing judges

Studies using CJ usually recruit suitable judges, and familiarise them with the method. Sometimes judges even make the judgements at the same time and location. We did not organise a familiarisation meeting as we were testing the method as if it were to be used by judges across Wales, and such meetings would be expensive in time and cost. Instead, we opted for creating and testing detailed guidelines to help each judge complete the task remotely and independently.

We sent the initial draft of the guidelines to several colleagues, and invited them to access the online platform and complete a CJ task. We modified the guidelines to reflect their feedback, and we also collected feedback from several judges once the real exercise was completed. Several aspects emerged as being important to include in the guidelines.

Judges need to receive quite detailed instructions about how to complete the CJ exercise as the online platforms, even if quite intuitive, could be confusing. In a real study, this could result in participants dropping out. We ended up copying a screenshot of the relevant part of the web page and giving step-by-step instructions while referring to the screenshot. The particularities of the platform used should also be considered. In our case, for example, the platform shows the average judging time. Some judges felt that it was linked somehow to the need to perform the judgements within certain time limits, so we included advice for judges to disregard the time. The platform also offered the opportunity to stop the exercise at any time and resume it later. Informing participants about this feature is useful as it allows more flexibility and could increase the participation rate.

Judges in the mock CJ mentioned that they sometimes struggled to discriminate between two statements. Given that in ACJ the more judgements which are made the more difficult it could become to compare subsequent pairs, we included guidance on what needed to be done when judges were unsure which statement to select. We suggested that they use whatever rationale they might find useful to differentiate the two statements and, if still unable to compare, choose at random (for example toss a coin). To our knowledge, during the actual CJ no one resorted to randomly selecting their answer.

There are no rules for how many judges need to be involved. However, each judge should have a reasonable workload. It is also important to consider the nature of the task. Comparing students' essays might need fewer judges as the standards of what is good are highly likely to be shared amongst teachers. However, the type of exercise undertaken in the current study, when no clear 'quality criteria' exist, might require more judges to ensure reliability. We are not aware of any research into this and, until shown otherwise, we think that having a larger number (that is exceeding the number of judgements suggested for achieving good reliability) is better, although, beyond a certain number, the marginal impact of every additional judge is probably limited.

Limitations

There are at least several limitations that need to be considered when using CJ in similar types of studies.

CJ provides a snapshot of the statements' impact at the time of the exercise. As such, CJ has a limited use for exploring how lasting the impact of a statement is or how the confidence is changing if two or more factors occur one after the other. Along the same lines, in this study, we assumed that people in real life will

[6] For example, percentage of teenagers being happy with their lives.

predominantly get information about education from news headlines. But even if this assumption were true, people would still not receive and absorb the information in ways similar to those in this exercise. In contrast to other uses of CJ in which the context has less importance or no bearing at all on judgements, it is highly likely that the context has a role in affecting public confidence and, in some cases, has an even bigger impact on people's opinions than the information itself. However, the statements in this exercise were mostly provided without context, not to mention that many of them were altered to make the wording simpler and clearer (for example rewording long sentences). This should be considered when interpreting the results of such a study.

Another aspect to consider when using CJ to explore a concept is the issue of construct validity. We strived to ensure that each statement was relevant to the concept of interest and, more importantly, that all relevant aspects were covered. However, as the concept is quite complex, the risk of missing important factors is real.

Although external validity was not important for the current pilot, it is clearly an important aspect when such a study is undertaken with the population of interest. As already mentioned, using a larger than usual pool of judges, perhaps using quotas to ensure a representative sample along with using infit values (see next section) as a diagnostic tool, could be a way to deal with this issue. However, this needs further research.

Finally, as indicated earlier, ACJ could overestimate the reliability coefficients. As Bramley (2007) explains, this happens because the objects' values are not known beforehand. Thus, all parameters are being estimated concurrently as the data from judgements are fed into the model. This produces 'spurious separation' of the objects. Overestimation gets worse the sooner the objects are paired algorithmically (instead of randomly),[7] to the point that the reliability coefficient becomes 'at best misleading and at worst worthless' (Bramley, 2015). The NMM platform uses a Progressive ACJ, in which adaptive selection increases gradually. This is thought to reduce the problem of inflated reliability coefficients.[8] To improve the reliability estimates, one can guide the initial selection of pairs by providing, if possible, information about the (estimated) value of the objects or rely on additional measures of reliability such as correlating the scores estimated by two groups of judges comparing the same pairs (Bramley, 2015).

Conclusions and potential for CJ use

The first published use of CJ in education was in 1993 (Pollitt, 2012). Since then, the method has been gaining ground in assessment (see Jones and Alcock, 2014) and awarding processes (see Bramley, 2007). CJ has also been used in other fields to rank different objects or statements (for example weighting the seriousness of perceived health problems (McKenna et al, 1981)) and suggested as an alternative to Likert scale survey questions (see Bockenholt, 2004).

Aside from measuring the (comparative) value of objects, we propose that CJ can also be used to explore or compare concepts by employing the infit statistics which in Rash-type models is derived from residuals. The infit indicates if a response is at odds with what the model would expect. The conventional rule is to consider an infit estimate acceptable when it lies within two standard deviations apart from the mean (Pollitt 2012).

An object's infit value[9] can be used to judge its relationship with the concept: a high infit deviance can be an indication that its belonging to the concept is ambiguous. For example, in the current study, most of the statements with high infit deviance in the positive group were statements about decision-makers' intentions. This may indicate that, on its own, such information might not increase people's confidence, and its value, instead, depends on other variables, such as the context or a judge's characteristics. High infit values can have other plausible explanations, such as confusing wording of the statement, and this should be considered as well.

[7] As an oversimplified example, let's imagine that we want to compare the numbers from 1 to 4 in terms of which is larger. By random selection, in the first round, we are to compare 3 with 4 and 1 with 2. In the next round, we select the pairs algorithmically, based on the information from previous judgements. Let's say that the rule is to compare winners with winners and losers with losers (i.e. 4 with 2 and 1 with 3). In all four comparisons we get reliable results, but the achieved rank order – 1 3 2 4 – is not the one that we would expect.

[8] **https://blog.nomoremarking.com/progressive-adaptive-comparative-judgement-dd4bb2523ffe**, accessed 19/12/17.

[9] The expectation of good infit value is 1 within a range from 0 to infinity. Anything less or more than 1 departs from this ideal. For example, the NMM platform suggests, as a rule of thumb, that an infit is fine if it is less than 1.2 and an infit value of 1.3 is considered significantly inconsistent and is to be separately investigated.

Examining how objects are ranked and scored offers an additional opportunity to gain further insight about the concept of interest. The ranking itself is quite important, as we can hypothesise that statements at the lower end are less related to the concept. But we can also examine other aspects. For example, when inspecting the true scores, we found that many statements formed clearly delimited clusters. Examining those clusters for a common theme can be an avenue for adding to the understanding of the concept.

Varying the statements across different variables can also offer interesting insights. In our study, we altered the statements according to the information source, the certainty of impact (now versus in the future or unclear), and the scale of impact. For example, we explored whether the same opinions expressed either by students or by parents would have a similar impact. The existence of stark ranking differences could be an indication that respondents trust one source of information more than the other. Another option (not used) would have been to explore the impact of the same message if it were published through different mass-media channels.

Our study also provided information that had not been anticipated. For example, participants were more consistent when judging the impact of negative statements (higher reliability and the same two statements having the first two ranks in both tasks), compared to the judging of positive statements which did not yield such consistent results. This could indicate that people are better at assessing the impact of negative factors than at determining the effect of a positive development. In terms of building confidence, this could mean that managing negative risks is paramount.

References

Achilles, C. M., Lintz M. N. and Wayson W. W. (1989). 'Observations on building public confidence in education'. Educational Evaluation and Policy Analysis 11(3): 275-284.

Agresti, A. (1992). 'Analysis of ordinal paired comparison data'. Journal of the Royal Statistical Society. Series C (Applied Statistics) 41(2).

Bockenholt, U. (2004). 'Comparative judgments as an alternative to ratings: identifying the scale origin'. Psychological Methods 9(4): 453-465.

Bradley R. A. and Terry M. E. (1952). 'Rank analysis of incomplete block designs I: the method of paired comparisons'. Biometrika 39: 324-45.

Bramley, T. (2015). 'Investigating the reliability of Adaptive Comparative Judgment'. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. [online] Available at: **http://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/** [Accessed 5/7/2018].

Bramley, T. (2007). 'Paired comparison Methods', in Techniques for monitoring the comparability of examination standards, edited by: Paul Newton, Jo-Anne Baird, Harvey Goldstein, Helen Patrick and Peter Tymms. [online] Available at: **https://www.gov.uk/government/publications/techniques-for-monitoring-the-comparability-of-examination-standards** [Accessed 5/7/2018].

Cohen, N. E. (1937). 'The relativity of absolute judgments'. The American Journal of Psychology 49(1): 93-100.

Cowell, R., Downe, J., Martin, S. and Chen, A. (2012). 'Public confidence and public services: it matters what you measure'. Policy & Politics 40: 123-143.

Jones, I. and Alcock, L. (2014). 'Peer assessment without assessment criteria'. Studies in Higher Education 39 (10): 1774-1787.

Jones, I. and Iglis, M. (2015). 'The problem of assessing problem solving: can comparative judgement help?'. Educational Studies in Mathematics 89(3): 337-355.

Laming, D.R.J. (2003). 'Human judgment: the eye of the beholder'. Cengage Learning EMEA. 2006 edition.

Loveless, T. (1995). 'The structure of public confidence in education'. John F. Kennedy School of Government, Harvard University Faculty Research Working Paper Series.

McKenna S.P., Hunt S.M. and McEwen J. (1981). 'Weighting the seriousness of perceived health problems using Thurstone's method of paired comparisons'. International Journal of Epidemiology 10(1): 93-97.

Pollitt, A. (2012). 'Comparative judgement for assessment'. International Journal of Technology and Design Education 22:157-170.

Pollitt, A. and Crisp, V. (2004). 'Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions?'. Paper presented at the British Educational Research Association (BERA) annual conference. UMIST, Manchester: UK.

Thurstone, L.L. (1927). 'A Law of comparative judgment'. Psychology Review 34: 273-286 [online] Available at: **https://brocku.ca/MeadProject/Thurstone/Thurstone_1927f.html** [Accessed 13/12/2017].

Wayson, W.W., Achilles, C., Pinnell, C.S., Lintz, M.N., Carol, L.N. and Cunningham, L. (1988). Handbook for developing public confidence in schools. Phi Delta Kappa Educational Foundation. Bloomington: Indiana.

Whelan, J. (2017). 'Prior distributions for the Bradley-Terry model of paired comparisons'. arXiv:1712.05311.

Whitehouse, C. and Pollitt, A (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment. Manchester: AQA Centre for Education Research and Policy.

RESEARCH NOTE

The challenges of conducting research inside Syria

Sally Gowland, BBC Media Action

Abstract

Since 2011, Syrians have been living through a war and the challenges of prolonged conflict. BBC Media Action's radio drama, Hay el Matar (Airport District), aimed to bridge divides and help people deal with the pressures of prolonged conflict. Before creating any programme, BBC Media Action conducts formative research with audiences to understand how they live, what they believe and the issues that affect them. However, operationalising research inside Syria in order to understand the Syrian audience brought many challenges and research problems. This research note discusses some of these challenges, such as access to respondents; researcher and respondent safety; training; ethics and dealing with trauma; and how these were dealt with, in order to offer learning and advice to others planning to conduct research in conflict situations.

Acknowledgements

This research was funded by the European Commission's 'Instrument Contributing to Stability and Peace'. BBC Media Action would like to thank all of the research participants for the generous donation of their time. Special thanks too to the NGO staff who helped to conduct the research and the many BBC Media Action staff members who contributed.

Background

Since 2011, Syrians have been living through a war and the challenges of prolonged conflict. Those who have remained in the country have lost family members and their homes; lack access to education and health facilities; and are living with the daily threat of violence. In such a context, ethnic and religious divides have intensified; armed groups have been increasingly recruiting young people in Syria and neighbouring countries; and the political situation is constantly changing. BBC Media Action's radio drama, Hay el Matar (Airport District), aimed to bridge divides and help people deal with the pressures of prolonged conflict. Funded by the European Commission, the thrice-weekly drama aired on BBC Arabic radio and online in 2016 and 2017. During this period, some parts of Syria were controlled by the regime; some by opposition forces; and some by the so-called Islamic State (IS).

BBC Media Action's research approach

Prior to creating any programme, BBC Media Action conducts formative research with audiences to understand their lives – how they live, what they believe and the issues that affect them. Following this, we conduct research to monitor and evaluate the impact of a programme on the intended audience to ensure it is achieving its outcomes.

In the countries in which BBC Media Action operates, the local BBC Media Action research team usually conducts this research. However, BBC Media Action was unable to operate inside Syria and, instead, worked with its local office in Beirut, Lebanon with researchers also from Lebanon. This presented several challenges: how to design and plan qualitative research which would help inform the development of the drama, and evaluate its impact with its key target audience – people living inside Syria.

The research aimed to conduct a series of focus groups with people from a diverse range of backgrounds inside Syria, to inform the shape of the drama. We then planned to ask them to listen to the broadcast of the radio drama for several months, and conduct follow-up focus groups to understand how people had engaged with the programme and what, if any, influence it had had on their lives. However, with the research team based in Lebanon and unable to conduct the fieldwork itself, they had to plan how best this could be done.

Operationalising the research

Firstly, the BBC Media Action research team had to decide if it could – and should – try to conduct primary research with people living inside Syria over the course of the drama's broadcast. The team spoke to other NGOs, researchers and donors to understand how they were reaching people inside Syria and conducting research.

The team found that many aid agencies were not risking sending their own staff into a war-torn field environment to conduct research, and instead, were using a local contractor. Though the team found a limited number of (eager) research contractors that could operate inside Syria and receive money from abroad, many of them did not have the geographic reach or the staff necessary to carry out rigorous research. Bypassing research contractors and hiring individual freelance researchers were not possible because this would have meant asking people to gather information in a hostile environment – without being able to offer them any security.

As a result, the team chose to work with a local trusted NGO (Dawlaty) operating in opposition-held areas which could conduct the research safely. It had a network of trusted practitioners working on the ground. This approach meant that large parts of Syria (regime-held and IS-held areas) would be excluded from the research. However, the benefits of conducting research in these types of areas had to be weighed against the risks of recruiting researchers in areas where armed groups might summarily execute someone for being an 'informant' or in areas in which researchers would need to inform the regime of the work and would then be monitored consistently. This could have posed a risk to researchers, affected the integrity of the research, and influenced the respondents' answers, skewing results.

Training

Even when working with a trusted NGO in opposition-held areas, there were security and safety implications. The BBC Media Action team in Lebanon was not able to travel to Syria to train the NGO. It was not possible for the NGO staff to come to Beirut for face-to-face training in how to conduct research as it was dangerous and costly for Syrians to travel across borders or governorates.

Therefore, the team in Beirut conducted Skype training sessions so that the NGO staff could be trained in qualitative techniques. This was challenging as the poor internet connection meant that training was often interrupted. Further, the training had to be comprehensive as some of the NGO staff had no previous experience in conducting research. Researchers trained the NGO staff in data protection, confidentiality, and safety procedures; and in how to talk to respondents and conduct focus group discussions effectively. NGO staff were also coached in how to safely share data without giving away identifying information, as well as to recognise when it was too dangerous to continue.

Ethics

In implementing the research, the team had to think about the safety of the respondents and the NGO staff. For example, it was vital to assess whether a participant telling a researcher about their daily life might seriously compromise either of them.

Another consideration was the risk of bringing together the same groups of people for focus groups, repeatedly, over a prolonged period. The context of conflict was continually changing with control shifting between the different factions, and people moving home. There was also concern that convening groups might arouse suspicion amongst local authorities or increase the chance that representatives from local authorities would infiltrate the groups.

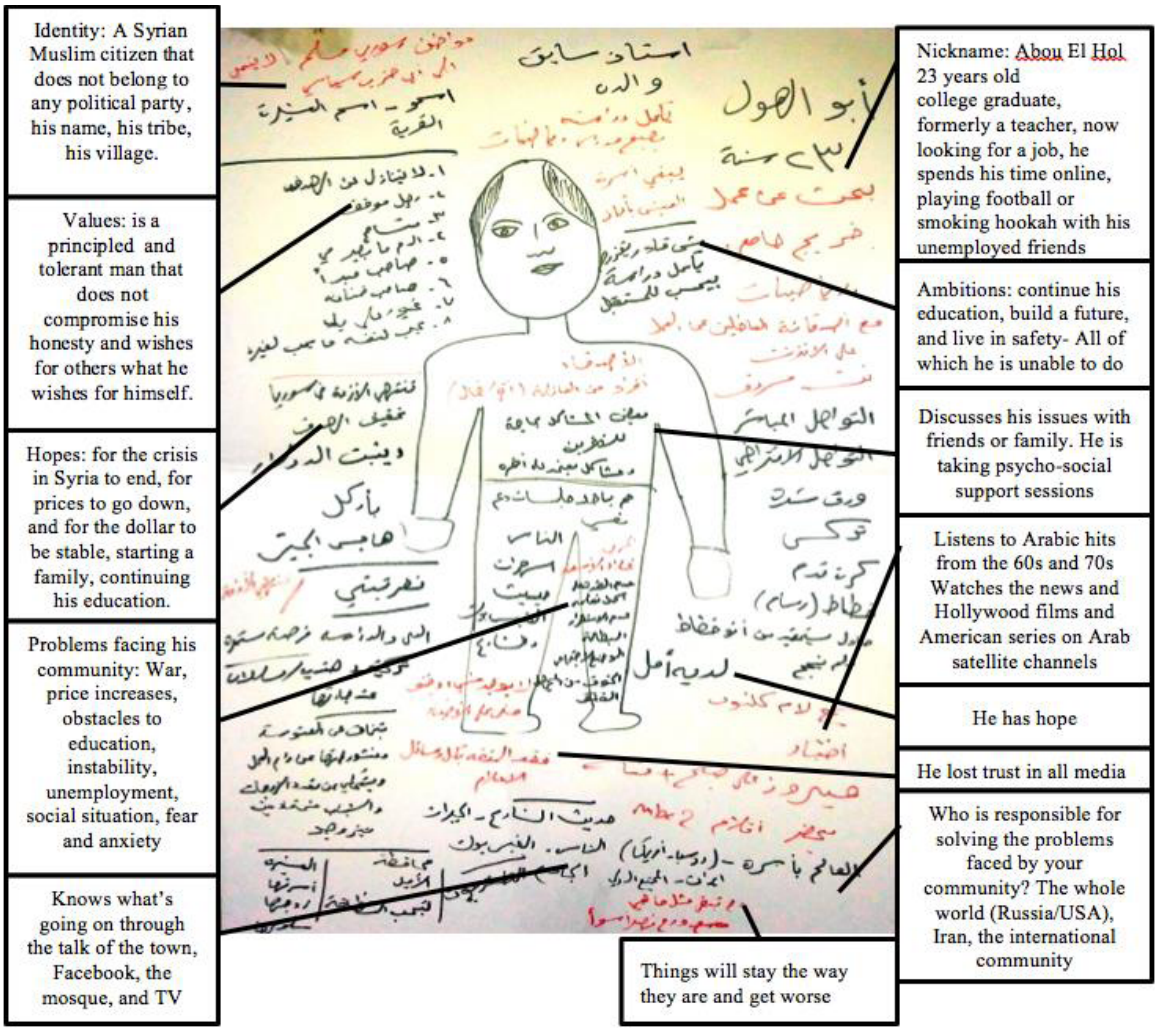
The original design aimed to revisit the same focus groups over the course of the entire broadcast period (nearly 12 months). However, for the reasons outlined above, Dawlaty advised that this would be risky, and instead, suggested that the research was split into three different cycles with different participants in each cycle. Each group would attend two meetings. The first would be a briefing session at the start of the listening period, in which people would be briefed on participation, fill in a short survey, and discuss their media use, information needs, the concerns they face, and the communities they live in. Three months later, after listening to the drama, people would be invited to another meeting with the NGO to evaluate the content and give feedback. The next cycle would repeat the same exercise but with a different set of people from different areas. In total, nine different groups were conducted – three in each cycle – with an equal number of male and female participants and involving a variety of age-groups.

Dealing with trauma

In the first meeting with each focus group, respondents were asked about their daily lives. This was through a participatory exercise: participants drew a picture of a typical person in their community and labelled it with details about this person's life – specifically, the challenges they face, their values and ambitions, and how they engage with the media. This exercise encouraged discussion about issues and concerns of participants and the people around them; built trust; and helped participants feel comfortable talking openly with one another and the NGO staff. Given that these participants had experienced significant trauma, this exercise allowed them to express some of their concerns and issues without disclosing highly personal experiences. In the second meeting, three months later, NGO staff asked people what they thought of the drama's characters and storylines and whether this had prompted them to think differently or discuss the issues raised.

The content of the focus groups was potentially harrowing and upsetting for NGO staff so there was a debrief system. After each group, the researchers in Lebanon checked in with the NGO staff and these sessions helped the team deal with some of the stresses.

At data analysis, it was sometimes difficult for researchers involved in the fieldwork to objectively assess the data. Frequent calls with London-based researchers helped them look at the data from different angles and to see beyond personal stories.



Above: results from one of the participatory exercises with a focus group of young men in Daraa governorate, south-western Syria

Conclusions

Conducting research in conflict environments is challenging. There are many barriers to designing and implementing research that can be carried out ethically and practically. Some ways which we have found to be effective have been:

- Understand the local context with local researchers. Working with local staff in the region who understood the context, dynamics and complexities of the conflict was essential for creating a context-specific research design that would work from a logistical and ethical perspective
- Work with trusted, local partners. This was important in order to reach respondents and ensure that data would be as high quality as possible because respondents were engaging with an organisation and people with whom they had an existing relationship
- Prepare for change: active conflict means that people move around and territories change hands. Prepare for drop-outs by over sampling but also for a final dataset which may not be as comprehensive as you anticipated
- Ensure there is support for researchers. Talking to people who are experiencing conflict can be traumatic, and it can be difficult for researchers hearing their stories. Make sure that debrief sessions are timed appropriately and that other support services are available if researchers need this

More at: **http://www.bbc.co.uk/mediaaction/publications-and-resources/research/briefings/middle-east-and-north-africa/syria/syrian-drama**

RESEARCH NOTE

Navigating the NHS and HRA ethics and governance process: a worked example

Hannah Hartley and Emma V Bolton, University of Leeds

Abstract

Researchers who plan to carry out research in the NHS or with NHS patients or members of staff are likely to need NHS and Health Research Authority (HRA) ethical and governance approvals. The recently combined NHS and HRA process is unfamiliar to many researchers. Whilst there is extensive information and advice on the HRA websites, this paper aims to provide an insight into the experiences of researchers who have recently gained the necessary approvals using the NHS and HRA ethics and governance process, and to offer direction and guidance on navigating the process.

Funding acknowledgement

The University of Leeds

Getting started: identifying the appropriate pathway

There are two types of approval that may be needed to undertake research in the NHS: ethical approval and Health Research Authority (HRA) approval. You may need to apply for both of these, or just one.

Ethical approvals are essentially focused on the participants, and whether the research is ethical, that is, that no harm will come to participants as a result of taking part in the research. Ethical approvals are assessed by research ethics committees (RECs).

HRA approval is needed for research projects involving NHS organisations for which the NHS has a duty of care for patients or staff who are recruited into the study as participants. It is orientated around governance and legal compliance. For example:

- If you plan to recruit participants who are NHS patients or vulnerable adults (for example, pregnant women or users of social care services), then you will most likely decide to apply for NHS and Health Research Authority (HRA) ethics and governance approval
- If you plan to recruit participants who are NHS staff members (for example, midwives), you may not need NHS ethical approval. Instead, you might decide to apply for local ethics approval, for example, your university REC, but you would also need to apply for HRA governance

Documentation

Before embarking on the ethical approval process, you should have a detailed research protocol agreed by you and your research team. You should know who your target sample is; have clear procedures for participant recruitment, data collection and data management; be confident that the research is feasible; and have carefully considered the ethical implications of the research.

The Integrated Research Application System (IRAS) is the place to start. The IRAS website creates various different forms, based on the initial 'project filter questions' – answer these carefully to generate the right form(s), and visit the 'help' section if confused. The new system means that you create only one IRAS

application that will be assessed by a REC and the HRA. You need to complete an IRAS form, as well as any other forms generated by the project filter questions, for whichever REC/HRA pathway you have identified. The IRAS form is relatively straightforward to complete, providing you have a comprehensive understanding of the research protocol. It is worth investing the time to get your protocol right, as it is usually the document which is requested by individuals you come into contact with during and after the ethics process. It provides an overview of your research, and will make filling in the IRAS form as painless as possible. The checklist function on the online IRAS platform prompts for the necessary documents to be attached to the application, such as the protocol and participant information documents, though you may not need to attach all the documents listed in the checklist.

Key people

It is important to identify the key people who will review your study and to seek their guidance throughout the process. This may include staff at your organisation and in the NHS Trust (for example, your sponsor's representative, or staff in the research and innovation/development department). If you are doing research in primary care, you should make contact with the local research and development (R&D) team, which will be based at a clinical commissioning group (CCG). It is good to be in touch early and to make sure you know everyone and how they can help you. These people will be able to help you complete forms, collate the relevant documentation and answer your questions about the process. Seeking information from peers who have successfully navigated the process is likely to be informative. However, bear in mind that all projects are individual, and there are likely to be differences somewhere along the line.

Submission to NHS REC

Once the IRAS application is complete, files have been attached and you are ready to submit, the application must be signed off by the appropriate people in your research team and organisation (for example, chief investigator, supervisors and sponsor). You must verify your application using the 'verification tool' before you can electronically submit. Make sure you check your form carefully and that everything is filled in. If the verification tool identifies any missing information, your electronic authorisations will become invalid and you will need to get the form signed by everyone involved again. The 'check your form' function can identify missing information, and can be used before the verification step.

You are then ready to book the REC and submit your application. These steps must be done on the same day. Firstly, you need to book the REC; check the locations and dates of the REC meetings on the HRA website; and note the meetings you would prefer to attend. Consider the location of the REC, and the appropriateness of the REC for the study. For example, there are qualitative-specific RECs and social care RECs which cover social care research. The RECs usually take place at least two weeks after you submit your application. The RECs have limited appointments each day and can, therefore, become fully booked, so it is better to have two or three meetings in mind for the next step. You must then ring up the central booking service (CBS) to book a REC meeting when you are ready to, but before you submit your application.

When you ring the CBS, you will be asked to answer questions in a yes/no format about your research. It is useful to have your IRAS form available, so you can provide consistent answers. The answers you provide will determine whether your study will go to full or proportionate review, and the type of REC it will go to. The CBS will confirm your REC details, which need to go on your IRAS form before you submit it. You can then submit your IRAS form using the e-submission tab.

Full and proportionate NHS REC review

If your study is sent for full review, you will be required to state your preferred RECs. You will be invited to attend the REC and answer any questions the committee may have about the research – this is not mandatory, but it offers the opportunity for you to answer questions in person and resolve matters raised by the committee. You may also attend by a tele-conference. If you want to attend by tele-conference, check with the REC and give two phone numbers, in case there is a problem with one, and make sure you are available to sit by the phone for a few hours. Once you have attended the REC and answered any questions, the REC will write to

you with an 'opinion' and/or any suggestions or amendments to the application or the documents. REC opinions are one of the following:

- Favourable opinion
- Favourable opinion with additional condition
- Provisional opinion
- Unfavourable opinion

You will also be contacted by the HRA, which may require additional changes to the document before it approves the study. Once the changes have been made and sent to the REC and HRA with a covering letter, you will receive ethical and HRA approval separately. You should wait for approval from both the REC and HRA before you start recruitment.

If your study poses minimal ethical risk and burden for participants, your study may be sent for proportionate review. This is a quicker process than a full REC, but is still rigorous. If your study is sent for proportionate review, all correspondence will usually be through email or telephone; you most likely will not attend a REC. Your study is first checked for validation, which means its appropriateness for proportionate review is checked. You should receive an email within ten working days stating the outcome of this. If it is not valid for proportionate review, then the study will be reviewed through the full review process. If it is reviewed through proportionate review, you will be contacted by email with an opinion and/or any amendments or suggestions to the application or the documents, and the HRA, parallel to those described in the full review process above.

What next? After REC and HRA approval

Once you have received ethics and governance approvals, you may need to seek local approvals, for example your identified NHS site(s) may need to confirm they have 'capacity and capability' to undertake the study. The HRA letter provides information on any further approvals you need before starting your research, such as pre-engagement, and capacity and capability checks. You may be required to send a 'local information pack' consisting of key study documents to the study site, for them to assess their 'capacity and capability'. In primary care, the local R&D team will facilitate this stage. If you are not employed by the NHS organisation, you may need to obtain a 'research passport' or 'letter of access' in order to carry out research activities onsite. You will need the support of the sponsor and the NHS Trust/local R&D team to obtain this. It may be useful to begin this alongside your ethics application, particularly if you are working to a limited timescale.

Once you have all the necessary approvals in place you may begin recruitment. Congratulations! Ensure you are confident in, and mindful of, what you can and cannot do under the remit of your approvals throughout the research project. If you decide to make any changes to the procedures identified in your application, you may need to apply for an amendment before you can do so.

The Social Research Association (SRA)
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