NatCen

Social Research that works for society

Using a targeted design and R-indicators to enhance sample quality in a web-CATI longitudinal study

Authors: Curtis Jessop, Dave Hussey, Klaudia Lubian, Martin Wood
Date: June 2019
Prepared for: ESRC

Economic and Social Research Council

Summary

- There is a growing consensus that fieldwork strategies that focus on response rate maximisation may not provide optimal results in terms of survey sample representativeness.
- Responsive or adaptive fieldwork designs offer an alternative approach, using auxiliary data to target fieldwork protocols to sub-groups within a sample with the goal of improving fieldwork outcomes (including sample representativeness).
- However, these designs can be difficult to implement – often relying on efficient, responsive, and flexible fieldwork systems and processes, and the availability of reliable and relevant auxiliary data.
- A ‘targeted design’ is a form of responsive design that targets fieldwork protocols at the start of fieldwork based on pre-existing information available for sample members. As such, it is particularly appropriate for longitudinal studies where substantial information will have been collected about study members at previous.
- This paper documents the application of a targeted design on the NatCen Panel study, using demographic information from the recruitment interview and participation history data to target fieldwork protocols with the aim of improving sample representativeness without affecting overall costs, fieldwork length or response rates.
- To evaluate the impact of the targeted design, we explored and developed alternative measures of sample representativeness including R-indicators and DEFFs
- We conclude that the implementation of a targeted design on a longitudinal study is feasible, even within tight constraints. However, given the relatively engaged nature of longitudinal study members, and that much of any sample bias may have been introduced at the initial recruitment wave, a key challenge is the selection and implementation of protocols strong enough to have a significant effect on sample quality.

1 Introduction

The widespread decline in social survey response rates in the UK and internationally has been well-documented (e.g. de Leeuw et al, 2018), and an increasing proportion of survey budgets are being spent to stem this decline out of concern for the impact it may have on sample representativeness and statistical power. However, response rates are simply a proxy for sample representativeness; focus on them may not improve sample representativeness and may even exacerbate existing bias.

In this context, there is growing consensus that response rates should not be the sole measure of sample quality, and that a 'one-size fits all' approach to fieldwork design may be sub-optimal as 'generic' approaches can introduce bias by being more appropriate for some groups than others.

Metrics such as R-indicators, which provide a summary measure of sample bias, together with responsive/adaptive fieldwork designs offer alternatives to response-rate focused approaches. However, their implementation is not necessarily straightforward: measuring non-response bias and adjusting fieldwork approaches for specific groups is difficult without up-front auxiliary information for issued sample members. For longitudinal studies this is available through data collected at earlier waves. However, the operationalisation of these approaches may still be challenging, requiring relatively complex analysis, systems, and processes, and increasing the difficulty (and costs) of fieldwork

This paper explores the feasibility of implementing a 'targeted design' that is evaluated using R-indicators on the NatCen Panel – a random probability longitudinal panel with a web-CATI fieldwork design. It will document its design and implementation and provide insights for its application in other longitudinal studies.

2 Targeted design

2.1 Responsive and adaptive designs

The terms 'responsive-' and 'adaptive design' can be used interchangeably (Couper & Wagner, 2011). Broadly, responsive designs[1] use auxiliary data to target fieldwork protocols to sub-groups within a sample, with the goal of improving fieldwork outcomes such as reducing costs, improving response rates, or improving sample quality.

These approaches are distinct from 'being responsive' to fieldwork circumstances, where (for example) a lower-than-expected response rate necessitates a change in fieldwork protocols. Rather, responsive designs should be pro-active, with fieldwork protocols pre-determined, and targeted.[2]

2.1.1 Using auxiliary data

Auxiliary data may be information held about cases *ahead of* fieldwork (e.g. survey answers or participation data from previous waves of a longitudinal study or information in or matched onto a sample frame) or information collected *during* fieldwork (e.g. survey responses or fieldwork paradata such as call outcomes).

The role of auxiliary data in a responsive design is to understand the survey sample and monitor fieldwork. They should be used to inform the approach selected, facilitate its implementation, and measure the impact it has on the desired fieldwork outcome. As such, their quality and accessibility dictate what design is possible and are fundamental to its success or otherwise. For example, the impact of a responsive design on sample quality will rely on the auxiliary data accurately identifying cases that are under-represented in the productive sample. The accessibility of those data will determine whether targeted protocols can be implemented at the start of fieldwork, 'live' during fieldwork, or in a subsequent fieldwork phase.

[1] We will use 'responsive design' in this paper.
[2] Appendix A discusses how targeted fieldwork protocols can intersect with a flexible fieldwork approach.

2.1.2 Static & dynamic designs

There are many different forms of responsive design, reflecting the specific circumstances of a study (e.g. auxiliary data available or fieldwork constraints). Couper & Wagner (2011) provide several case studies, and consider these approaches to fit along a continuum. Here, we split them into two categories based on Schouten et al (2013):

- **Static designs** where fieldwork protocols are fixed at the start of fieldwork based on exiting auxiliary data
- **Dynamic designs** where fieldwork protocols can change during fieldwork based on auxiliary data collected

Static designs require enough auxiliary data to be available at the start of fieldwork for effective targeting, and are therefore especially relevant for longitudinal studies, or in studies where the sample frame has extensive, relevant, data.

In contrast, dynamic designs do not rely on existing information but on information collected during fieldwork to target protocols. This can be done at different 'speeds' - for example, fieldwork may be split into two stages, with information collected in stage 1 used to inform treatments in stage 2. Alternatively, protocols could be continuously changing as more data become available.

Depending on their nature, dynamic designs can be especially challenging to implement. They either rely on systems, processes, and resources being in place that can collect and feedback the required data, analyse them, and change fieldwork protocols for a sample unit in a timely fashion, or a fieldwork timetable that allows for this to be done slowly, with a pause between stages.

This paper, focuses on the implementation of a 'targeted design' approach on the NatCen Panel. A targeted design is a form of responsive design akin to a 'static responsive design', using data collected at the recruitment interview and previous fieldwork waves to target fieldwork protocols.

2.1.3 Development of appropriate protocols

The development of appropriate protocols and their effective implementation will also determine the success or otherwise of a responsive design in achieving its goal. For example, to improve the representativeness of a sample, a mechanism must be identified that will improve participation for under-represented groups (when many studies will already be following a 'response maximisation' strategy so marginal effects may be small). Similarly, protocols must be appropriate – a face-to-face follow-up mode may not be appropriate for a study with a limited budget or fieldwork period, and targeted call patterns will not be feasible without systems and processes that can quickly collect and analyse fieldwork paradata and 'push out' new protocols.

3 Using a targeted design on the NatCen Panel

3.1 Introduction to the NatCen Panel

The NatCen Panel is the only probability-based research panel in Great Britain open for use by the social research community. It is designed to collect survey data in a time- and cost- effective manner, maintaining sample quality by using probability-based sampling and covering the offline population.

The Panel currently consists of approximately 8,000 members aged 18+ recruited from the 2015 to 2018 waves of the British Social Attitudes (BSA) survey. All BSA participants are invited to join the Panel, and at a typical panel wave all panel members who have not subsequently left will be invited to take part, maintaining the random probability design. Although designed for cross-sectional studies, the NatCen Panel is fundamentally a longitudinal study, with the same participants returned to over time.

3.2 Goals of implementing a targeted design

The decision to implement a targeted design on the NatCen Panel was taken in mid-2017. Fieldwork monitoring at the time suggested that the overall response rates were gradually declining, most likely caused by attrition from the panel. At the same time, the design effects of the survey

weights (DEFF) were gradually increasing, suggesting this was resulting in a decline in sample quality (Jessop, 2018).

Despite this, the decision was taken *not* to implement a response maximisation strategy such as increasing the incentive level: response rates and sample quality appeared to be declining very gradually, and annual recruitment of new panel members would limit the impact of attrition. Also, a key aim of the NatCen Panel is to collect data at lower cost and in less time than 'traditional' probability-based approaches and it was felt response-maximisation strategies would disproportionately undermine the Panel's ability to deliver on these goals.

A targeted design offered an opportunity to address these issues with response maximisation strategies. By targeting fieldwork protocols based on demographic data from the BSA survey and panel members' participation history it aimed to improve the sample profile while keeping costs, fieldwork length, and response rates neutral. More specifically:

- Minimise the cost/fieldwork length impact by only increasing the effort spent on panel members in under-represented groups
- Minimise the cost impact by not increasing the effort spent on panel members who always participate or who have never participated
- Off-set the cost impact by reducing the effort spent on panel members in over-represented groups
- Minimise any negative impact on response rates by only reducing effort with panel members who either always participate or who have never participated

3.3 Identifying cases for prioritisation & de-prioritisation

3.3.1 Over- and under- represented cases

The first step in implementing the targeted design was identifying cases with characteristics typically over- or under- represented in people participating in surveys administered via the Panel. To do this, unweighted estimates for a range of demographic variables for people participating in panel surveys were compared to weighted estimates for the full BSA sample (representing the target population) to identify variables and categories where bias was most present.

Based on this analysis, a model was developed to measure the extent to which panel members had characteristics which were over- or under-represented in a panel survey[3][4]. This was then translated to a score, with scores of more than 1 indicating the panel member has characteristics typically over-represented, and vice versa. Panellists were then categorised into eight groups from most under- to most over-represented[5].

3.3.2 Accounting for participation history

To improve the efficiency of the design and minimise any negative impacts, panellists' participation history was also accounted for. As a result, the targeted design was limited to people who had been panel members for at least one year so sufficient data was available. This also reflects that the impacts of attrition on sample representativeness are small in the first year.

At each wave, panel members were grouped into those who have taken part in all previous waves they have been invited to, some waves, and no waves. These were then combined with the representativeness groups to create 5 'priority groups' (Table 3:1).

Overall, panel members who were under-represented were given higher priority (and vice-versa) to improve the sample profile. Panel members that participated in no waves or all waves were given lower priority, as reduced effort was less likely to have a negative effect and additional effort was less likely to have a positive effect. Panel members that have participated in 'some' waves were

[3] The descriptive analysis demonstrated that key areas of bias remained stable across waves, so the model was based on the most recent panel survey wave.
[4] A list of the variables used in this model is included in Appendix B
[5] The groups were defined to be balanced, but otherwise separated arbitrarily. Panel members with scores of 0 to 0.33 were put into group 1 (the most under-represented); 0.33 to 0.5 into group 2; 0.5 to 0.66 into 3; 0.66 to 1 into 4; 1 to 1.5 into 5; 1.5 to 2 into 6; 2 to 3 into 7; and 3 or more into 8 (the most over-represented).

given higher priorities as we might expect them to be the most likely to be affected by any targeting[6]. The number of 'cells' allocated to each group was decided based on the estimated marginal impact on response rates and costs, aiming to keep them as close to neutral as possible[7].

Table 3:1 Priority groups definition

	Participated in all waves	Participated in some waves	Participated in no waves
1 (most under-represented)	Medium priority	**Highest priority**	Low priority
2	Medium priority	High priority	Low priority
3	Medium priority	High priority	Low priority
4	Medium priority	High priority	Low priority
5	Low priority	Medium priority	**Lowest priority**
6	Low priority	Medium priority	**Lowest priority**
7	Low priority	Medium priority	**Lowest priority**
8 (most over-represented)	Low priority	Medium priority	**Lowest priority**

3.4 Fieldwork design

3.4.1 'Standard' fieldwork design

The NatCen Panel uses a sequential mixed-mode fieldwork design lasting one month. At the start of fieldwork, all eligible panel members are sent a letter and email inviting them to take part in an online survey and offered a conditional £5 incentive to thank them for their time. If they do not take part, panel members are sent up to one letter, two email, and two text message reminders.

After two weeks, panel members who have not taken part online, and for whom a phone number is available, are issued to NatCen's Telephone Unit to 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 interviewers to make six attempts to contact and interview panel members at a time that suits them.

3.4.2 Targeted design approach

The targeted design follows a similar design but with a different incentive offer, minimum call requirement, and number of reminder letters for each priority group (Table 3:1). These elements were selected because they were expected to impact response rates but would be expensive to implement for the whole sample. Strategies such as additional reminder emails which have relatively small marginal costs would be better as part of a 'response maximisation' design.

Table 3:2 Overview of targeted protocols

	Incentive offer	CATI fieldwork	Communications
Highest priority	£10	Minimum of 8 calls	Two reminder letters
High priority	£5	Minimum of 8 calls	One reminder letter
Medium priority	£5	Minimum of 6 calls	One reminder letter
Low priority	£5	Minimum of 4 calls	No reminder letters
Lowest priority	£5	Not issued to CATI	No reminder letters

[6] As lower priority cases are not sent reminder letters/have reduced CATI fieldwork (Section 3.4.2), those without email addresses or internet access are therefore moved back to 'medium priority' to ensure coverage
[7] These calculations are indicated in Appendix C and Appendix G

4 Measuring representativeness

Response rates have long been used as the primary indicator of survey sample quality. However, they have been shown to have a weak relationship with non-response bias; which may occur if respondents and non-respondents to a survey differ in their characteristics. In fact, there are many examples of increased data collection efforts leading to a higher response rate but also to greater non-response bias. As a result, it is now widely accepted that alternatives to response rates are required for assessing the quality of survey statistics.

4.1 R-indicators

One alternative is the Representativity Indicator (R-indicator), a metric of sample representativeness developed as part of the RISQ (Representative Indicators for Survey Quality) project[8]. The use of R-indicators is not widespread, particularly in the UK, with response rates typically still used as the primary measure of sample quality. One reason for this may be the need for auxiliary information for both respondents and non-respondents; this is a particular problem for cross-sectional surveys of the general population in the UK[9]. In contrast, longitudinal surveys possess a wealth of auxiliary information for those who have responded in the past, making the use of R-indicators far more practical.

4.1.1 Developing R-indicators for the NatCen Panel

The R-indicator we have used for the NatCen Panel is based on the standard deviation of estimated response probabilities. An unweighted logistic regression model of response was used to estimate the response probabilities, with respect to a set of key socio-demographic variables. It is defined as:

$R = 1 - 2\,S(\rho)$

If response probabilities are all equal, then the response data set is “representative” (by definition). In this case the standard deviation of the response probabilities is zero, leading to a value 1 for the R- indicator. If the response probabilities vary then the data set is not fully representative. The degree to which they vary is reflected by their standard error. The maximum value the standard error can assume is 0.5; in this case the value of the R-indicator is equal to 0. This R-indicator therefore varies between 0 and 1 with values closer to 1 indicating a more representative response dataset, other things remaining equal.

The R-indicators were developed using 15 socio-demographic variables collected on the BSA[10]. These include all the variables used in the models to derive the panel non-response weights plus two more, and additional categories[11]. As such, R-indicators can provide a precise and targeted measure of representativeness.

‘Adjusted’ R-indicator

The ideal comparison would be between the panel survey respondents (unweighted) and the population (represented by BSA respondents, weighted by the BSA weight, which is calibrated to GB population estimates). This is not possible under the framework we have used: the model must be weighted or unweighted. To overcome this problem, consideration was given to creating an “adjusted” R-indicator to include the effects of non-response to the BSA, in addition to the two stages following it. A methodology for doing this using a proxy measure (based on a function of the BSA weights) was designed and tested but ultimately rejected after consultation with a key member of the RISQ group.

[8] https://www.cmi.manchester.ac.uk/research/projects/representative-indicators-for-survey-quality/
[9] This is less problematic in countries with population registers that can be used for sampling and include useful information for comparing respondents with non-respondents.
[10] A full list is included in Appendix B
[11] Marital status and long-standing health/disability were also included as available demographic variables that may be associated with topics typically covered by research conducted through the NatCen Panel

4.2 Design Effects (DEFFs)

As an alternative and supplement to R-indicators, the design effects (DEFFs) of the panel survey weights were also used as a measure of representativeness. The weights are used for analysis of a single wave of the NatCen Panel and are created using a two stage process: at the first stage, response to the invitation to join the panel is modelled using a logistic regression model weighted by the BSA weight; at stage two, response to the panel survey is modelled using a logistic regression model weighted by the product of the BSA weight and the non-response weight from the previous stage. The final weight is the product of the BSA weight and the non-response weights from the two stages[12].

4.3 Comparing the two measures

Neither the R-indicators nor the DEFFs are ideal measures of the representativeness of the panel sample. Their relative strengths and weaknesses can be summarised as follows:

The **R-indicators** measure loss of representativeness *following* response to the BSA. Unlike DEFFs, they are unconstrained by prior processes, including all variables and categories deemed useful. In addition, they can be applied to sub-groups, for example to understand how different experimental groups may be affected and can provide variable- and category-level scores to provide deeper understanding of any bias (although this is not used in this paper).

The major weakness of the R-indicators is that they don't include the effect of non-response to the BSA. This is a substantial drawback but is impossible to get around: R-indicators rely on auxiliary information being available for all sample members, whilst the variables used in our model are only available for those who participated in the BSA survey[13]. Apart from this, the R-indicators cover the same sources of potential representativeness-loss as the DEFFs but in more detail and with additional coverage.

The **DEFFs** provide an indication of *overall* representativeness (i.e. they take account of the BSA's weight to the GB population); in this respect they have an advantage over R-indicators by encompassing all stages of non-response.

Unlike R-indicators, however, they were not designed to measure representativeness-loss – they cannot be applied to sub-groups or provide category- or variable- level insights. They are also a blunt instrument in comparison, due to merging of categories (to avoid inflation of standard errors) and trimming of the largest weights (to avoid individuals receiving extreme weights), and are limited to reflecting the variables included in the standard weighting models.

5 Evaluating the targeted design

This section evaluates the targeted design against its objective of improving the sample profile while keeping costs, fieldwork length, and response rates neutral. It will initially look at the survey response rates of the different priority groups to measure the direct impact of the protocols, before looking at the overall impact on the sample profile using the DEFF and R-indicators[14].

The targeted design was not implemented on the NatCen Panel experimentally. As such, this analysis evaluates impact by comparing figures before and after the targeted design was implemented; however, without a counter-factual, any perceived effects (or lack of) may be caused (or muted) by factors external to the targeted design.

5.1 Impact on participation of priority groups

To understand the impact of the targeted design protocols on the different priority groups, we look at the survey response rates[15] for each priority group before and after the targeted design was

[12] The variables included in the regression model for the NatCen Panel weights are included in Appendix B
[13] This drawback would also apply to longitudinal studies, where R-indicators would go back to those participating in Wave 1, but not necessarily the original sample frame.
[14] The impact on overall response rates and costs are discussed in Appendix E and Appendix G respectively
[15] Calculated using a base of all cases issue to that NatCen Panel wave

implemented (split out by the BSA year the panel member was recruited from), and relative to the survey response rates to all issued cases recruited from that BSA year.

The targeted design moves resources to higher priority cases and away from lower priority cases. Were the protocols having the expected impact we would therefore expect the survey response rates to increase for the higher priority, remain constant for medium priority cases and decrease for the lower priority cases (or remain at zero for lowest priority cases), before returning to the previous trend. The patterns are discussed for each priority group below[16], but overall the evidence does not consistently point to the targeted protocols having the anticipated impact on response rates.

Appendix Figure D:1 suggests a decline in survey response rates for panel members in the highest priority groups recruited from BSA 2015 and 2016 was (at least initially) halted by the implementation of the targeted design; but, in contrast, the BSA 2017 panel members show a continuation of a gradual increase in survey response (which may or may not be attributable to the targeted design following implementation).

Figure D:2 suggests the high priority cases follow a similar, but somewhat muted, pattern, with BSA 2015 and 2016 cases showing a more gradually declining survey response rate, which seems to initially stabilise after the targeted design was implemented, but the BSA 2017 cases showing a continuation of a gradual upward trend.

Figure D:3 shows, as expected, no impact on the survey response trend for medium priority cases, and Figure D:4 shows that low priority cases showed an initial drop in survey response rates before returning to their initial levels. Finally, Figure D:5 shows, as expected, that the survey response rates for lowest priority cases mostly remain at zero.

5.2 Impact on sample profile

5.2.1 Design effects of the weights (DEFF)

One way to measure the impact of targeted design on the sample profile is by looking at the DEFF at each wave (Section 4.2). Were the targeted design having the expected impact, we would see an initial decrease in the DEFF following its implementation; before returning to a gradual increase over time, reflecting the gradual decline in overall response rates (Appendix Figure E:1).

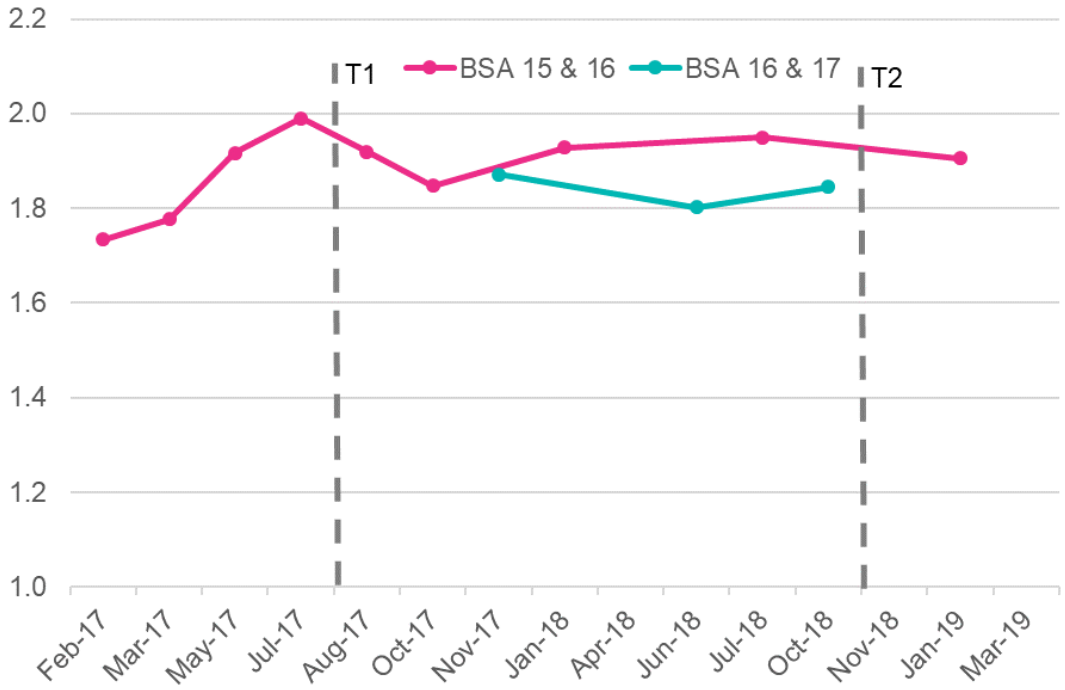
Figure 5:1 shows the DEFF values across panel waves where the BSA 2015 & 2016 panel members were included, before and after the targeted design protocols were implemented for them (T1), and for the BSA 2016 & 2017 cases after the protocols were implemented for the BSA 2016 cases, but before they were implemented for the 2017 cases[17].

This more clearly follows the expected pattern, with the implementation of the targeted design protocols apparently stopping the increase in the DEFF values over time before the use of the targeted design. While we cannot see an ‘impact’ on the 2016/2017 cases as they do not span an implementation, it is notable that they do not show the trend for increasing DEFFs as the 2016/2017 cases prior to T1.

[16] See Appendix D for charts and more detailed discussion

[17] As the weighting models are applied to all survey respondents, it is difficult to reliably split out DEFFs for each BSA sample group. However, for reference, this is done in Appendix F

Figure 5:1 DEFFs over time by sample groups

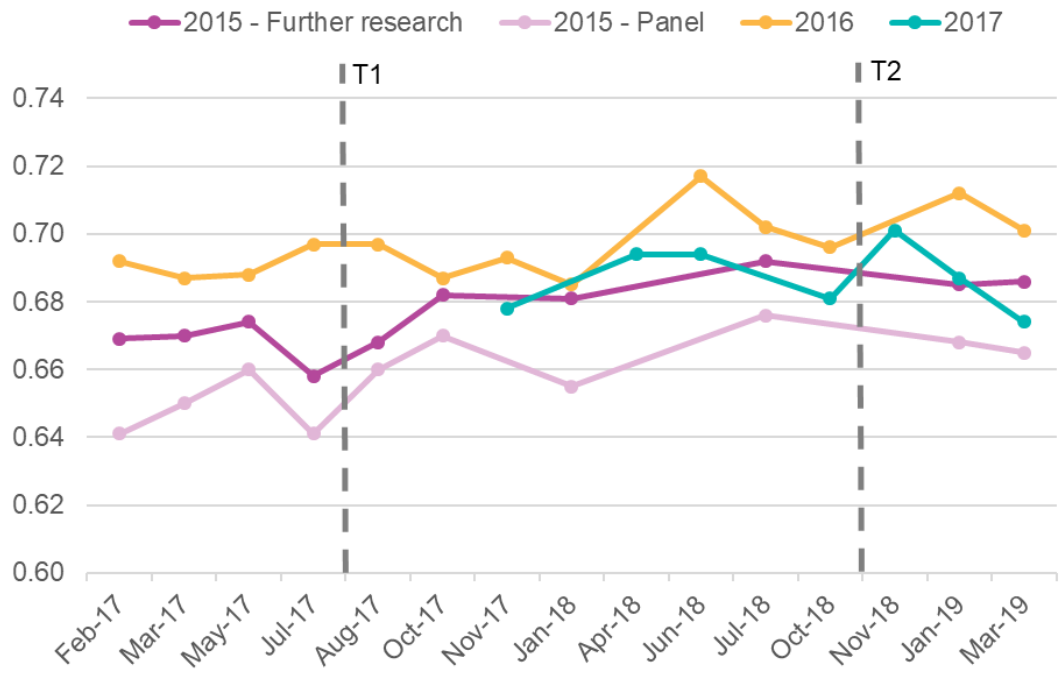


5.2.2 R-indicators

R-indicators are an alternative approach to measuring the sample profile (Section 4.1.1). Figure 5:2 shows the R-indicator values across panel waves by different BSA sample groups, and when the targeted design protocols were implemented for panel members recruited from BSA 2015 & 2016 (T1) and BSA 2017 (T2). Overall, there is little variation: all sample groups have R-indicator scores of between 0.64 and 0.72, with individual sample groups' scores not varying by more 0.035[18].

[18] Typical margins of error around the R-indicators vary between 0.025 (for BSA 16) and 0.035 (for the two BSA 15 groups); a difference of 0.035 is therefore on the borderline of statistical significance.

Figure 5:2 R-indicators over time by sample group[19]



As with the DEFF, were the targeted design having the expected impact, we would expect an overall trend of declining representativeness over time, with the R-indicator scores increasing following the implementation of the targeted design. However, these patterns are not seen in the chart. While the average R-indicator scores following the implementation of the targeted design are higher than prior, there is a pattern of gradual increase over time, rather than an uplift immediately following the implementation. Indeed, it is of note that the decrease in the DEFF and overall response rates seen in the run-up to T1 that prompted the implementation of a targeted design are also not reflected here.

6 Discussion

Implementing a targeted design

This project aimed to explore the feasibility of implementing a targeted design on the NatCen Panel, with the goal of improving the sample profile, while keeping costs, fieldwork length, and response rates neutral. This process has demonstrated that implementing a targeted fieldwork design is possible. However, the analysis in Section 5 does not show a clear or consistent impact on the sample profile:

- There is some indication that the protocols targeted at the higher priority groups halted a decline in survey response rates for the BSA 2015 and 2016 cases, but the same pattern is not seen for the BSA 2017 cases.
- For measures of sample profile, the DEFF trend data also suggests the targeted design may have stopped a decline for the BSA 2015 & 2016 sample but the R-indicator analysis does not indicate the same pattern or any change after the implementation of the targeted design.

This outcome can be explained by a combination of the targeted design not working and us not being able to detect effects:

- The protocols are targeted at the panel fieldwork stage - at this point, the majority of non-response has already occurred via recruitment to the BSA survey, the panel, or subsequent attrition. This may limit the impact the targeted design can have at the overall level. Relatedly, it

[19] Panel members recruited in 2015 are separated into two groups: those who were invited to join the panel, and those agreeing to take part in further research, who showed different response patterns

means the sample is relatively engaged, meaning NatCen Panel members' propensity to participate may not be substantially affected by the 'enhanced' protocols.

- Any effects may also be muted by the proportions receiving the different protocols. Almost half (45%) of cases are issued into the 'medium priority' group, receiving no change in protocol, while for the 29% in the lower priority groups we aim to minimise any impact. We are therefore only trying to affect the response rates of c.25% of cases, with 6% targeted with the highest effort protocols. At this level, any impact on response rates may be muted at the overall level.
- Finally, the lack of an experimental implementation makes it difficult to assert the presence (or absence) of any impact. Survey response rates fluctuate by wave due to survey content, fieldwork timing, and other exogenous factors, even as protocols remain stable. Given effects may be small, this makes it difficult to 'separate the signal from the noise', an issue exacerbated by the conflicting conclusions of different measures of representativeness (discussed below).

Measuring sample representativeness

The analysis in Section 5 and Appendix E appear to show contradictory findings between measures. As the overall response rates decline gradually over time, the R-indicators suggest, if anything, a small overall improvement in sample representativeness while the DEFFs suggest an overall decline, although the variation is marginal in both cases[20].

Appendix Figure H:1 confirms that there is a (weak) positive correlation (coefficient 0.16) between R-indicator scores and DEFFs, i.e., where the R-indicators indicate a lower level of representativeness, the DEFFs indicate a higher level.[21] This pattern may be driven by many differences in the measures; in particular, different variables and categories are included in the R-indicator and weighting models, and they do not cover the same non-response stages. Given the small amount of variation it is also possible we are measuring random variation, and we should not draw firm conclusions from this correlation. Irrespective, the differences exemplify the value of having multiple measures of sample representativeness, as one measure may give a different answer to another, without either one necessarily being 'right'.

Conclusions for longitudinal studies

The implementation of the targeted design on the NatCen Panel was in a very specific context – the relatively engaged sample, tight budget constraints, and restrictions of not impacting costs or fieldwork length limited the extent to which the goal of improving the sample profile could be achieved. These findings may be particularly relevant for emerging discussions of inter-wave designs for the UK's longitudinal panel and cohort studies which may be operating with similar restrictions.

However, these limitations may not be present in other contexts, and the lack of clear positive outcome may not translate to a less-engaged sample, larger overall budget and fieldwork period, or different fieldwork goals. At the same time, the auxiliary data used (demographic data from the 'recruitment wave' and participation history) should be available for all longitudinal studies, as should additional forms depending on the fieldwork approaches used.

Responsive designs can be difficult to implement, increasing complexity and putting additional pressure on fieldwork systems and staff. This study demonstrates the feasibility of implementing a *targeted* form of responsive design in a longitudinal study even under tight constraints. By doing so, those studies can ensure that scarce resources are used in an 'optimal' fashion. The key challenge will be selecting and balancing protocols so they have the desired impact, but this is, of course, a challenge whether a targeted design is used or not.

[20] If we put confidence intervals around the median R-indicator, each of the lines would include the minimum and maximum values. We cannot do the same analysis with the DEFFs.

[21] Appendix H includes additional analysis looking at the correlation between R-indicators and response rates for specific waves

7 References

Couper, M.P. & Wagner, J. (2011). Using Paradata and Responsive Design to Manage Survey Nonresponse.

de Leeuw E., Hox J. & Luiten A. (2018). International Nonresponse Trends across Countries and Years: An analysis of 36 years of Labour Force Survey data. Survey Insights: Methods from the Field.

Groves, RM., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. Public Opinion Quarterly,72(2), 167-189.

Jessop, C. (2018). The NatCen Panel: developing an open probability-based mixed-mode panel in Great Britain' Social Research Practice, 6, 2-14.

Schouten, B., Calinescu, M. & Luiten, A. (2012). Optimizing quality of response through adaptive survey design. Survey Methodology. 39. 29-58.

Appendix A. Flexibility in protocol adherence

Once panel members have been allocated to priority groups, the implementation of the targeted design is the key challenge as the management of multiple fieldwork designs for different groups creates additional logistical burden and increases the likelihood of error[22].

As the NatCen Panel uses a web/CATI fieldwork design, processes are more centralised and errors easier to prevent. However, the delivery of fieldwork in practice does require some flexibility. For example, a fieldwork wave may be under more time pressure than usual, for example due to a need for results to feed into a particular consultation, resulting in fieldwork being compressed to 3 weeks rather than the standard 4. This would mean there would insufficient time to send an additional reminder letter or for the telephone unit to make additional calls to higher priority cases. Alternatively, participation may be lower than expected for a particular wave, for example due to a survey topic being less engaging or more sensitive, and the decision may be taken that it is too risky to reduce effort for some groups, or that additional effort should be placed on more cases to bolster response rates. In either instance, this would be a deviation from, and potentially mute the effects of, the targeted design protocols.

This flexible approach to fieldwork protocols is no different to what would be implemented in with a non-targeted design: while the 'standard' approach may be preferable, it should not be blindly adhered irrespective of actual circumstances. Further, this principle is not incompatible with a targeted design approach. Indeed, the targeted design provides a framework for a more considered approach to flexibility. For example, if it is known ahead of time that an interview may be longer than usual, a £10 incentive might be offered to high priority (as well as highest priority) cases to off-set any negative impacts on response rates, but only for cases that are more vulnerable to non-response and likely to affect the sample profile, minimising the impact on costs. Similarly, if fieldwork is progressing slowly, high priority cases may be sent an additional reminder letter as well as the highest priority cases to bolster response rates in key areas and minimising the impact on fieldwork costs.

[22] Previous research has also identified interviewer non-compliance as a potential problem

Appendix B. BSA variables used in calculating survey weights, R-indicators, and targeted design priority groups

Appendix table B:1 BSA variables & categories used targeted design, survey weights, and R-indicators

BSA Variable	Targeted design	Survey weights	R-indicators
Sex/Age	Male 18-24 Male 25-34 Male 35-54 Male 55-64 Male 65+ Female 18-24 Female 25-34 Female 35-54 Female 55-64 Female 65+	Male 18-24 Male 25-34 Male 35-44 Male 45-54 Male 55-64 Male 65-74 Male 75+ Female 18-24 Female 25-34 Female 35-44 Female 45-54 Female 55-64 Female 65-74 Female 75+	Male 18-24 Male 25-34 Male 35-44 Male 45-54 Male 55-64 Male 65-74 Male 75+ Female 18-24 Female 25-34 Female 35-44 Female 45-54 Female 55-64 Female 65-74 Female 75+
NS-SEC	Managerial and professional occups Intermediate occupations/ Employers in small org; own account workers/ Lower supervisory and technical occupations Semi-routine and routine occupations/ Not classifiable	Managerial and professional occups Intermediate occupations/ Employers in small org; own account workers/ Lower supervisory and technical occupations Semi-routine and routine occupations/ Not classifiable	Managerial & professional occups Intermediate occupations Employers in small org; own account workers Lower supervisory & technical occupations Semi-routine & routine occupations Not classifiable/ Not applicable
Highest qualification	Degree Less than a degree No qualification	Degree Higher educ below degree/ A level or equiv O level or equiv/ CSE or equiv/ Foreign or other No qualification	Degree Higher educ below degree A level or equiv O level or equiv CSE or equiv/ Foreign or other No qualification

Appendix table B:1 BSA variables & categories used targeted design, survey weights, and R-indicators			
Household type	Single-person household 3+ adults, no children Other	Single-person household Adult(s), with children Adult(s), no children	Single-person household 1 adult (with children) 2 adults (no children) 2 adults (with children) 3+ adults (no children) 3+ adults (with children)
Tenure	Owned/being bought/shared owndership Rented (LA+HA) Rented (other) + Other	Owned/being bought/shared owndership Rented (LA+HA) Rented (other) + Other	Owned/being bought/shared owndership Rented (LA) Rented (HA/Trust/New Town) Rented (Other) + Other Rented (Private)
Interest in politics	A great deal Quite a lot Some Not very much None at all	A great deal/Quite a lot Some Not very much/None at all	A great deal Quite a lot Some Not very much None at all
Party identification	Conservative Labour None Other/refused/don't know	Conservative Labour None Other/refused/don't know	Conservative Labour None LibDem UKIP Other/refused/don't know
Household income	£1,200 per month or less £1,201 per month or more Don't know/refused	£1,200 p.m or less £1,201 - 2,200 p.m £2,201 - 3,700 p.m £3,701 or more p.m Don't know/refused	Less than £770 £771 - 1000 £1,001 - 1,300 £1,301 - 1,700 £1,701 - 2,200 £2,201 - 2,700 £2,701 - 3,300 £3,301 - 4,200 £4,201 - 5,600 £5,601 or more Don't know/refused

Appendix table B:1 BSA variables & categories used targeted design, survey weights, and R-indicators			
Region	London Not London	North East North West Yorkshire and The Humber East Midlands West Midlands East of England London South East South West Wales Scotland	North East North West Yorkshire and The Humber East Midlands West Midlands East of England London South East South West Wales Scotland
Internet access	Has internet access No internet access	Has internet access No internet access	Has internet access No internet access
Ethnicity	White Not white	White Not white	White Not white
Economic activity	N/A	In work/waiting to start a job Other	Full time education Paid work Unemployed Retired Other Looking after home
Marital status	N/A	N/A	Married Living with partner Separated/divorced Single/widowed
Whether has a long-standing physical or mental health condition or disability	N/A	N/A	Yes No

Appendix C. Number of cases in different priority groups

Initial allocation

The allocation of resources to the different priority groups aims to be cost- and response rate neutral – any cost or gain in response rates from additional resources for the high priority groups should be balanced by savings or drops in response rates from fewer resources being spent on low priority groups. The net change in costs and response rates are a function of the size of each group, as well as the costs/impact of the protocol changes.

Table C:1 shows the proportions of the issued sample recruited from BSA 2015 and BSA 2016 in each priority groups at the first wave that the targeted design approach was implemented (August 2017). It shows that overall, the highest proportion of cases were 'medium priority', while similar proportions of cases were high/highest priority or low/lowest priority. This reflects the relatively cautious approach, with few cases receiving extreme changes as the scale of impact was unknown and these protocols were being implemented on 'live' studies.

Appendix table C:1 Proportion of issued cases in each priority group (August 2017 wave)

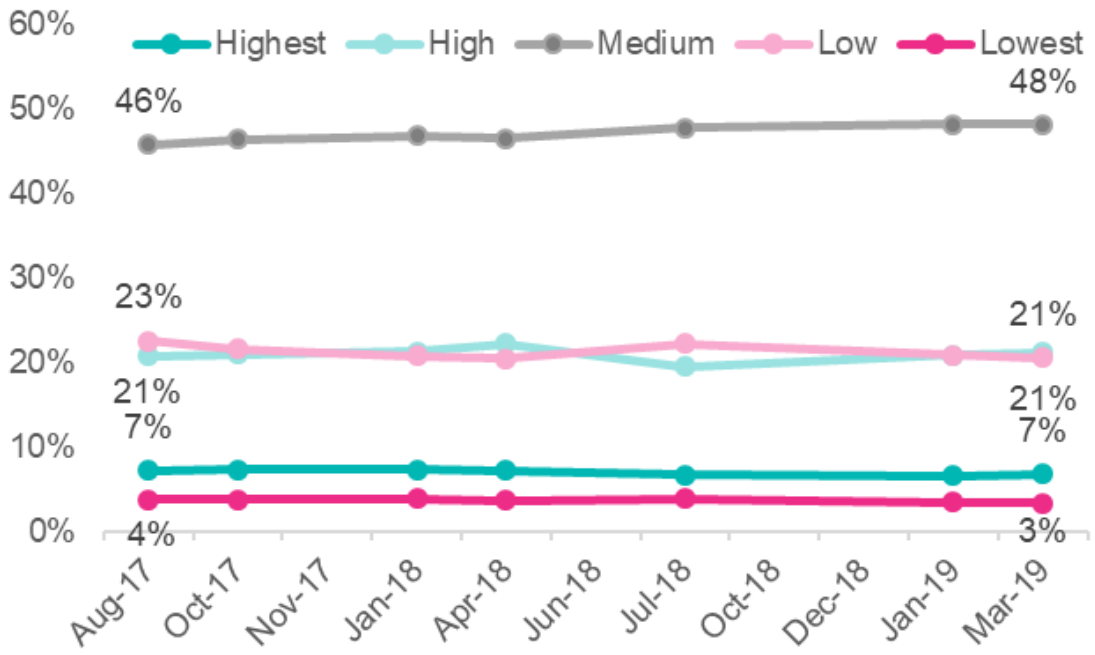
Priority group	BSA 15	BSA 16	Total
Highest	7%	4%	6%
High	21%	16%	19%
Medium	46%	44%	45%
Low	23%	30%	25%
Lowest	4%	6%	4%

Changes over time

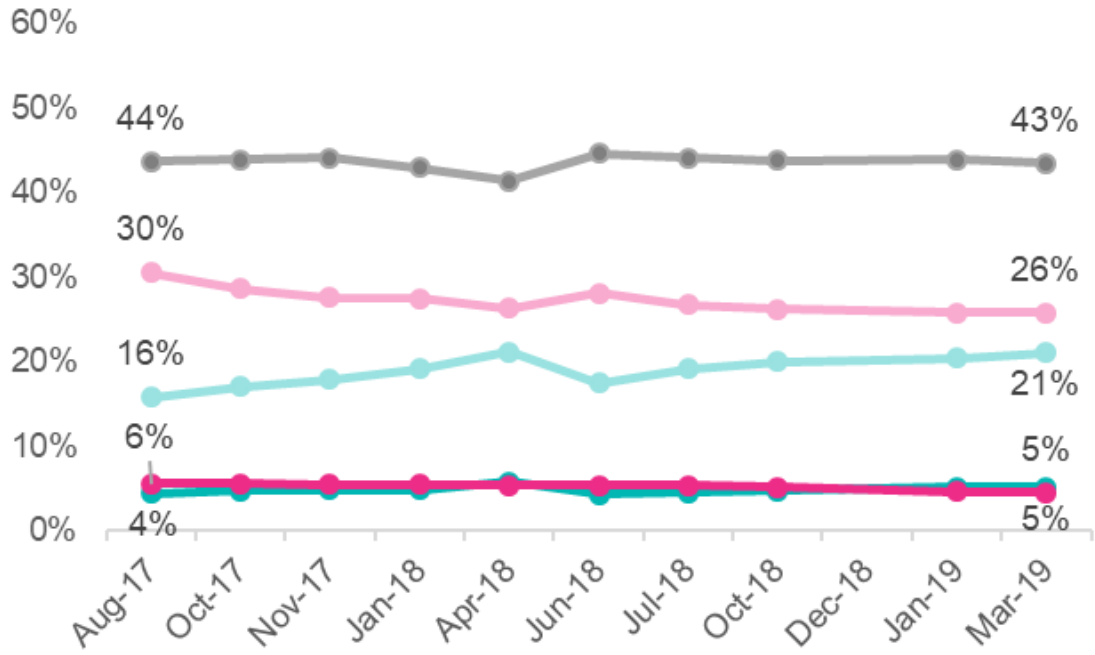
Table C:1 also shows that a higher proportion of those in recruited from BSA 2015 were in high priority groups than those from BSA 2016. The priority group that a panel member is allocated to can change over time. Overall, the model, indicating whether a panel member has characteristics that are over- or under- represented, is run annually (as each new BSA cohort has sufficient data to be targeted). At a finer level, panel members' participation history changes on a wave-by-wave basis, with more panellists move into the 'participated in some waves' group, as they take part or miss a wave for the first time. As well as this, all groups decrease in size as panel members request to leave the panel.

Figures C:1 and C:2 show how the proportions of the issued sample recruited from BSA 2015 and BSA 2016 in each priority group changed over time. While the BSA 2015 sample has remained fairly stable (reflecting that panel members had had more time to leave the panel/not participate/participate for the first time), for the sample recruited from BSA 2016, the proportion in the 'high' and 'low' priority groups increase and decrease respectively, making the targeted design approach increasingly expensive over time.

Appendix figure C:1 Proportion of BSA 2015 issued cases in each priority group across waves



Appendix figure C:2 Proportion of BSA 2016 issued cases in each priority group across waves



Appendix D. Impact of protocols on survey response rates for different priority groups

This section discusses in more detail the impact of the targeted design protocols on the survey response rates of each priority group, as summarised in Section 5.1. To do so, we look at the survey response rates for each priority group at each wave, before and after the targeted design was implemented (T1 for BSA 2015/2016 cases; T2 for BSA 2017 cases) and split out by the BSA year the panel member was recruited from. In addition, we plot the survey response rates for *all* issued cases to try to account for some fluctuation due to a wave having especially high or low response rates (although they are broadly stable).

Since panel members can move between different priority groups between waves (or leave the panel), the panel members in each priority group will change from wave to wave. There are therefore multiple survey response rates associated with a priority group at each wave prior to the implementation of the targeted design. For simplicity, we include the average of these in each chart[23].

Highest priority cases

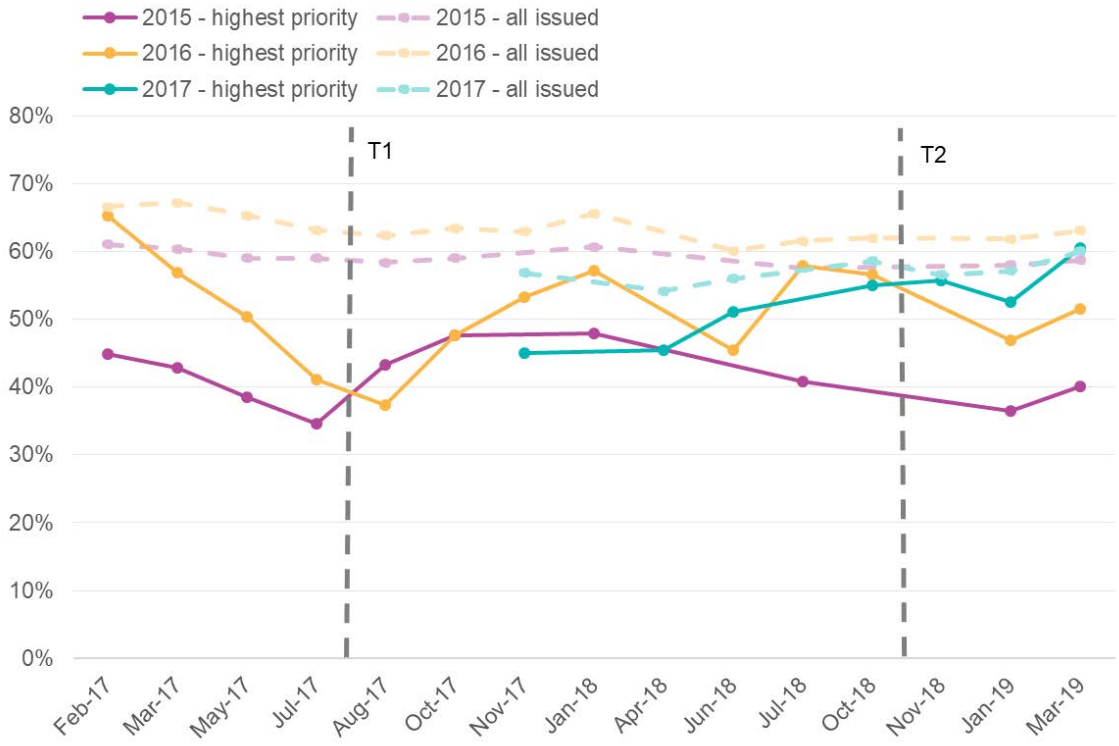
The highest priority cases are those which were identified as most under-represented and that have participated in some, but not all, preceding waves. The targeted design moves additional resources to them to improve their response rates. We would therefore expect them to have lower survey response rates than the issued sample overall prior to the implementation of the targeted design, and for these survey response rates to be higher after its implementation.

Figure D:1 paints a mixed picture. Panel members in the highest priority groups recruited from BSA 2015 and 2016 show declining survey response rates prior to the targeted design being implemented at T1; in contrast those recruited from BSA 2017 show, if anything, a gradual increase in survey response rates prior to the implementation of the targeted design.

Following the implementation of the targeted design, BSA 2015 and 2016 cases no longer show a trend of declining response rates, instead fluctuating from wave-to-wave, while BSA 2017 cases appear to continue on an upward trend.

[23] For example, in Figure D:1, the response rate for the BSA 2017 highest priority cases at the October 2018 wave is the average of the response rates of BSA 2017 highest priority cases issued in November 2018, the BSA 2017 highest priority cases issued in January 2019, and the BSA 2017 highest priority cases issued in March 2019

Appendix figure D:1 Survey response rates over time by BSA recruitment year for highest priority cases and all issued cases



This suggests no clear impact of the targeted design on survey response rates of the highest priority cases. Certainly, the response rates for the highest priority cases from BSA 2015 & 2016 are no higher following the implementation of the targeted design than they were before, but it may have stopped an apparent decline. Conversely, the BSA 2017 cases do have higher response rates with the targeted design approach, but this looks to be a continuation of an existing upwards trend.

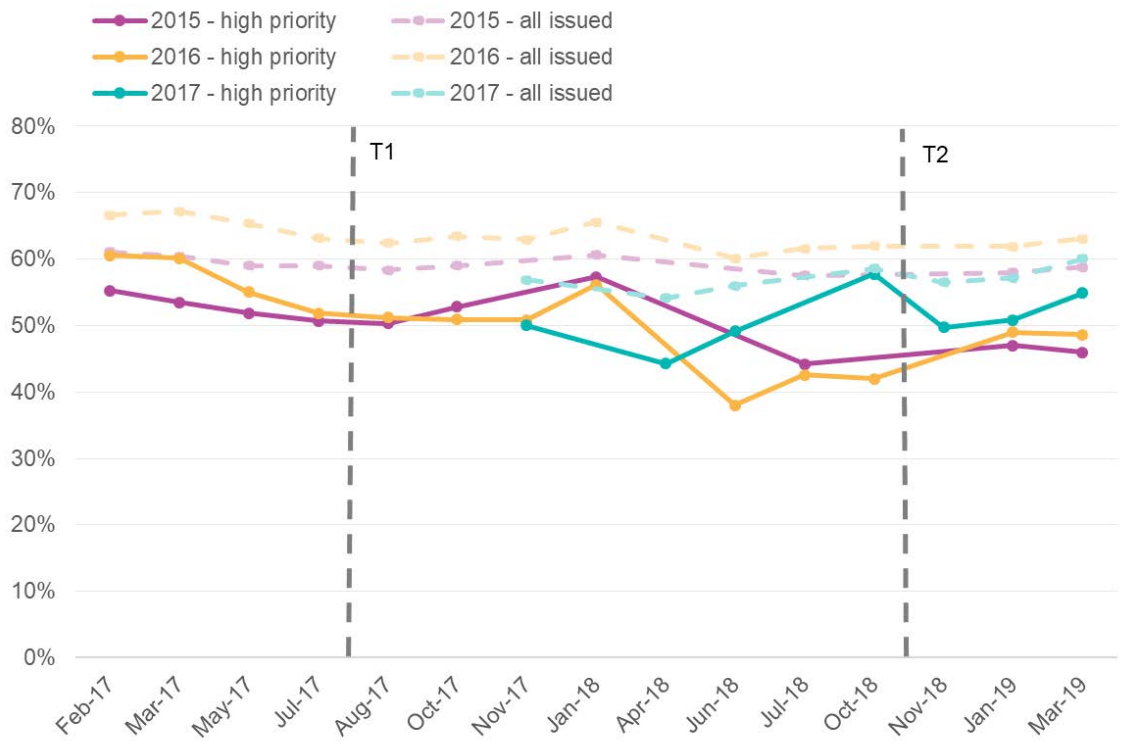
High priority cases

The high priority cases are those which were identified as under-represented (but not as under-represented as the highest priority cases) and having participated in some, but not all, preceding waves. The targeted design moves additional resources to them to improve their response rates (but not as many resources as for the highest priority cases). On balance, we would therefore expect them to have lower survey response rates than the issued sample overall prior to the implementation of the targeted design, and for these survey response rates to be higher after its implementation.

Figure D:2 shows the high priority cases following a similar, but muted, pattern as seen with the highest priority cases. BSA 2015 and 2016 cases show a trend of gradually declining survey response rates prior to the implementation of the targeted design, after which they fluctuate, but are no higher than previously. In contrast, BSA 2017 cases show slightly higher survey response rates following the implementation of the targeted design, but this may have been on an upward trend anyway.

Again, this shows no clear or consistent indication of an impact from the targeted design.

Appendix figure D:1 Survey response rates over time by BSA recruitment year for high priority cases and all issued cases

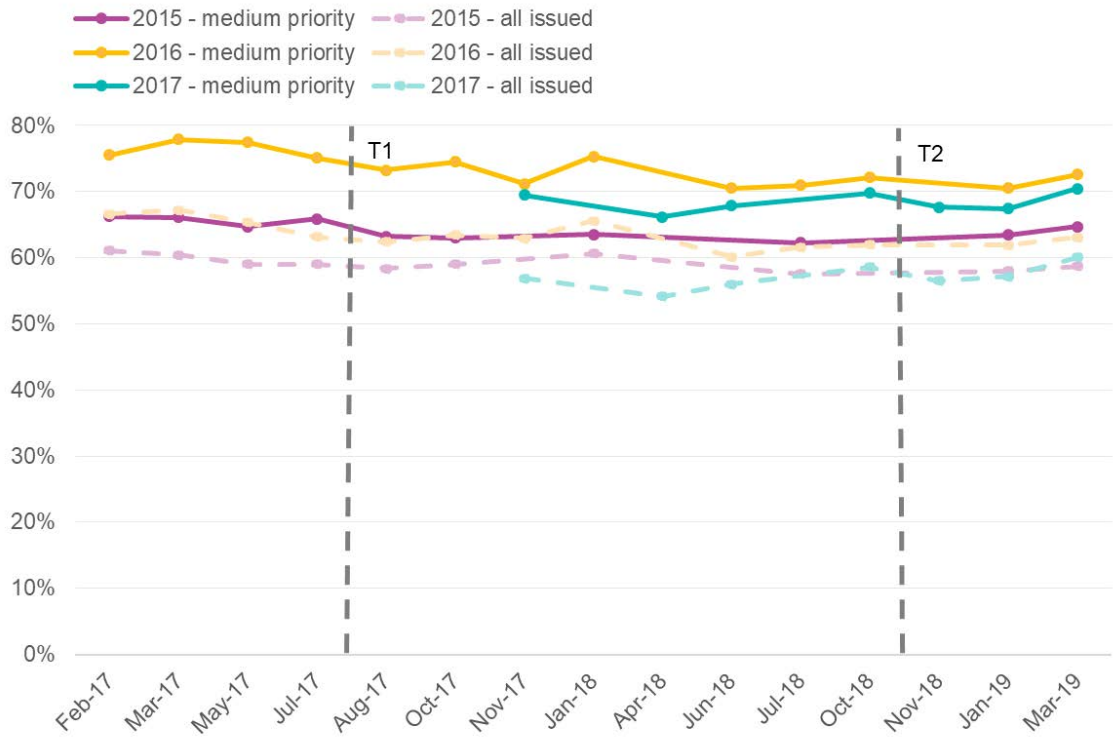


Medium priority cases

The medium priority cases are those which were identified as under-represented but having participated in all preceding waves, or identified as over-represented, but having only participated in some preceding waves. These cases experienced no change in fieldwork protocol. We would therefore expect them to have higher response rates than the issued sample overall prior to the implementation of the targeted design, and for these response rates to remain higher after its implementation.

Figure D:3 shows that for sample from all BSA waves, the medium priority cases had higher survey response rates than the issued sample as a whole both before and after the implementation of the targeted design, and (as would be expected) suggests no impact from the implementation of the targeted design.

Appendix figure D:1 Survey response rates over time by BSA recruitment year for medium priority cases and all issued cases

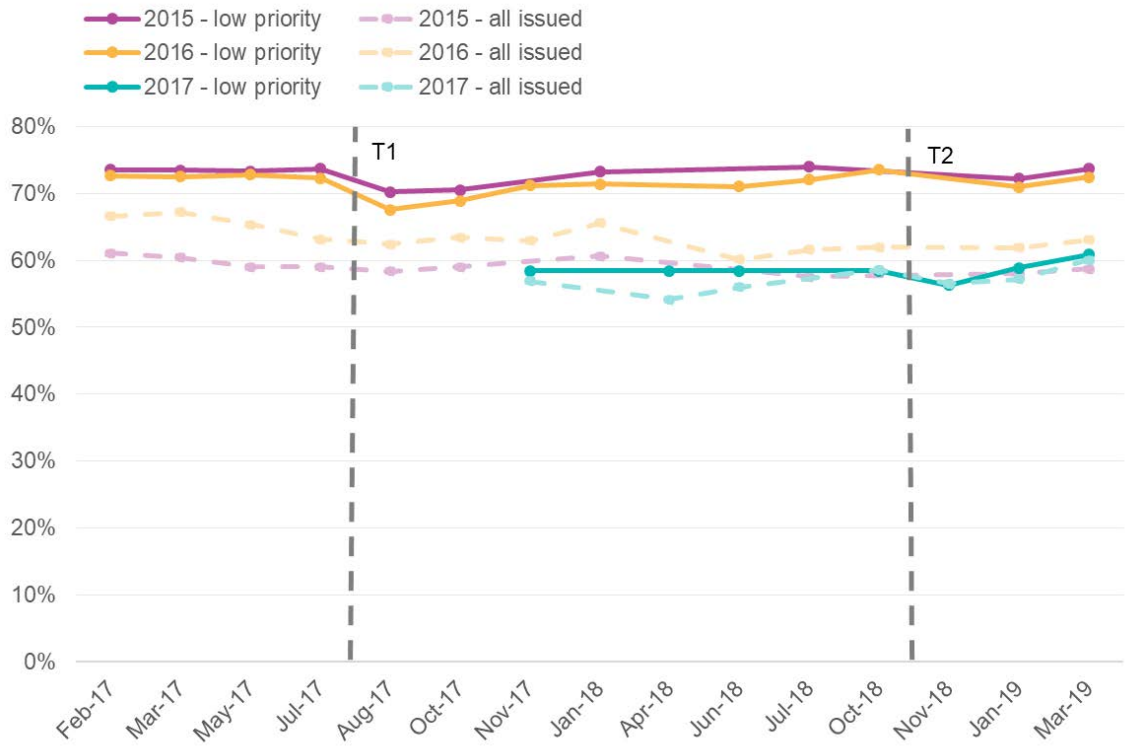


Low priority cases

The low priority cases are those which were identified as over-represented and having participated in all preceding waves, or identified as under-represented, but having not participated in any preceding waves. The targeted design reduces the level of effort and resources spent put into getting these cases to participate. On balance, we would therefore expect them to have higher response rates than the issued sample overall prior to the implementation of the targeted design, and for these response rates to drop slightly after its implementation.

Figure D:4 shows that for sample from BSA 2015 and 2016 waves, the low priority cases had higher survey response rates than the issued sample as a whole, both after the implementation of the targeted design and before. However, cases recruited from BSA 2017 show survey response rates similar to the issued sample as a whole. None of the sample groups' response rates appear to be affected by the implementation of the targeted design, though it should be noted that the design aims to have minimal (negative) impact on this group.

Appendix figure D:1 Survey response rates over time by BSA recruitment year for low priority cases and all issued cases

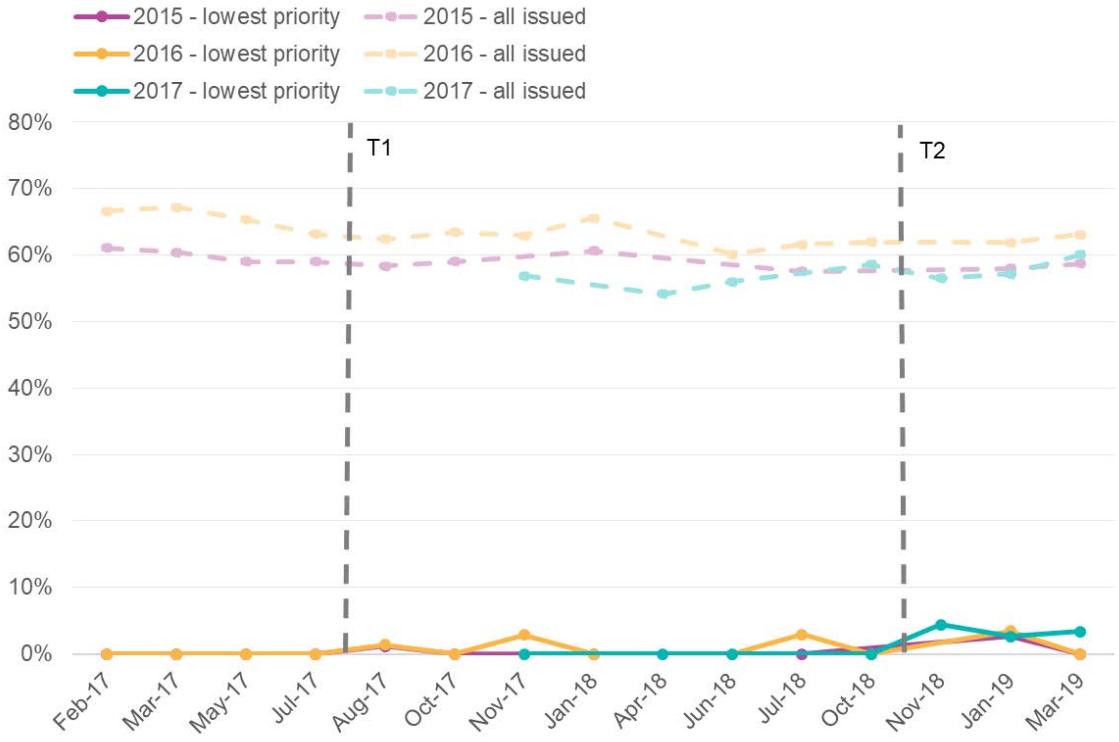


Lowest priority cases

The lowest priority cases are those which were identified as over-represented and having not participated in any preceding waves. The targeted design reduces the level of effort and resources spent put into getting these cases to participate. By definition, these cases should have a 0% survey response rate prior to the implementation of the targeted design, and we might expect it to subsequently increase only on occasion.

Figure D:5 shows the survey response rates across panel waves by different BSA sample groups for the 'Lowest priority' groups. As expected, it shows that for sample from all BSA waves, the lowest priority cases had survey response rates close to zero both before and after the implementation of the targeted design.

Appendix figure D:1 Survey response rates over time by BSA recruitment year for lowest priority cases and all issued cases



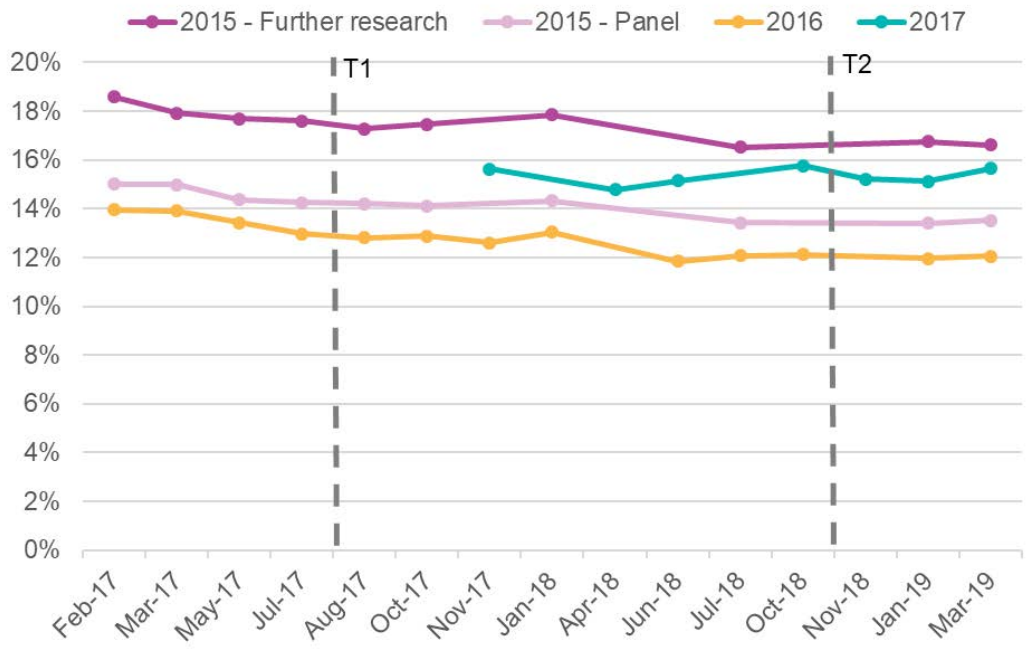
Appendix E. Impact on overall response rates

Figure E:1 shows the overall response rates[24] across panel waves by different BSA sample groups and the point from which targeted design protocols were implemented for panel members recruited from BSA 2015 & 2016 (T1), and BSA 2017 (T2).

Overall, we see very little variation in the response rates, with a long-term gradual decline in the proportion of participants recruited from BSA 2015 and 2016 participating, but little change in the proportion of those recruited from BSA 2017

Based on this, there is no indication that the implementation of the targeted design approach had an impact on the overall response rates for panel surveys, with the gentle decline seen for BSA 2015 and BSA 2016 samples seen prior to T1, and the stability seen for BSA 2017 prior to T2, continuing.

Appendix figure E:1 Overall response rates over time by sample group



[24] Calculated using a base of all eligible cases issued to the original BSA interview and, therefore, accounting for initial (non)participation in the recruitment survey, recruitment rates, attrition and panel survey response at a particular wave. However, all variation between waves will be due to attrition and panel survey response rates.

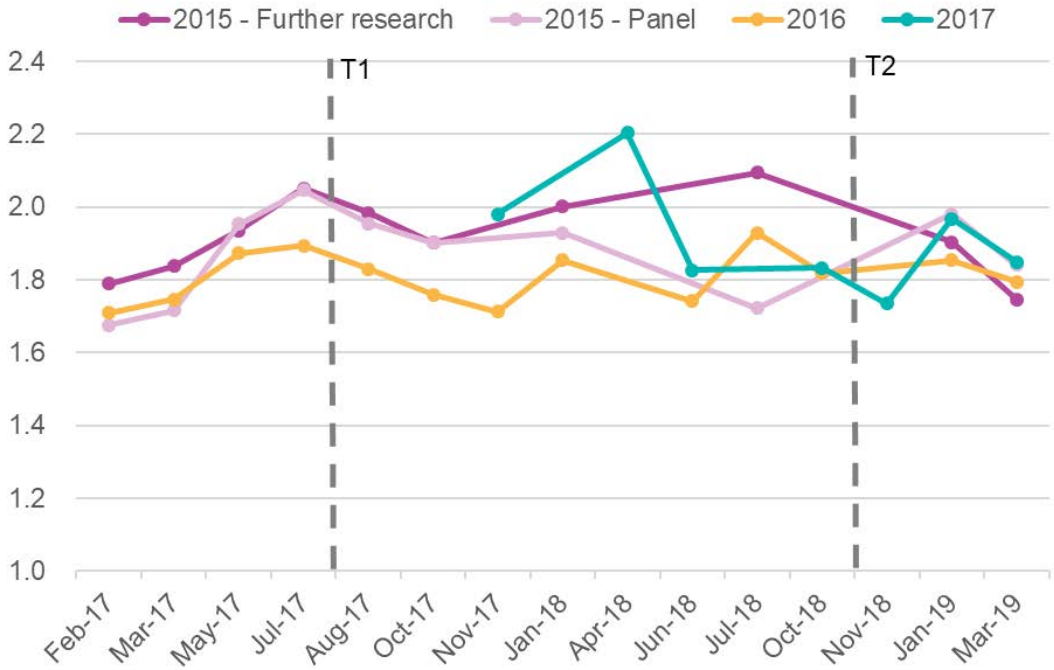
Appendix F. DEFFs over time by sample group

As outlined in Section 5.2.1, as the weights for the NatCen Panel are computed for all panel members, rather than within BSA year, it is problematic to split out the DEFFs for each group – the model may inadvertently down-weight cases from one group to balance the up-weighting of cases from another. While we would the effect of this to be minimal as non-response patterns appear to be similar across BSA years, it is possible and may confound results.

However, not constraining the analysis to waves where the entire issued sample is the same does allow us to look at additional data points, and in finer detail. Figure F:1 shows the DEFFs of the different sample groups over time. The pattern for cases recruited from BSA 2015 & 2016 appears much the same as indicated in Section 5.2.1, with an apparent increase in the DEFFs stopping after the introduction of targeted design at T1, followed by fluctuation.

However, the BSA 2017 sample shows a different pattern. Again reflecting the pattern for the BSA 16 & 16 DEFFs outlined in Section 5.2.1, there is no apparent increase in the DEFFs prior to the implementation of targeted design for this group at T2, and there does not seem to be any impact resulting from that implementation.

Appendix figure F:1 DEFFs over time by sample group



Appendix G. Costs

As well as aiming to increase the sample representativeness while maintaining response rates, the targeted design aimed to be cost-neutral. This section approximates the impact of the different elements of the targeted design on costs per wave, assuming an issued sample of 1,000.

Incentives

One of the more costly changes to the protocols (per participant) is the increase in incentives from £5 to £10 on completion. As such it was only applied to the highest priority group, and relatively few cases were included in this group.

We can estimate the impact on costs by multiplying the number of highest priority cases that 'normally' participate by 5 and adding the number of additional interviews expected multiplied by 10.

As outlined in Appendix C, c.6% of all issued cases were identified as 'Highest priority'. The proportion of those issued completing is more difficult to estimate, varying between c.35% and 65% depending on the wave/sample source, as outlined in Appendix Figure D:1, as is the impact of the increased incentive, as discussed in Section 5.1. Table G:1 therefore summarises the impact on costs of increasing the incentive to £10 with varying response rates and changes in response rates for the highest priority groups, showing that, for every 1,000 issued cases, the costs would increase by £360 to £660.

Appendix table G:1 Cost impact/1000 issued cases of increased incentives for highest priority group for different 'normal' response rate and response rate changes

'Normal' response rate	**Increase in response rate**		
	0pp	**10pp**	**20pp**
40%	£360	£420	£480
50%	£450	£510	£570
60%	£540	£600	£660

Mailings

Although less expensive at an individual level, the changes in mailing affect a larger proportion of the issued sample. As well as increasing costs by sending an additional letter to the highest priority cases, it aims to save costs by not sending reminder letters to low and lowest priority cases.

Typically, c.55% of highest and low priority cases issued, and 100% of lowest priority cases will be sent a reminder letter. This accounts for those that have already taken part before the reminder letter is scheduled, which will be none of the lowest priority group as they 'never' take part.

Table G:2 estimates the cost impact per 1,000 issued cases based on the proportions of the sample in each group, and assuming a marginal cost of 69p per letter.

Appendix table G:1 Cost impact/1000 issued cases of changing letter reminder protocols				
	% of issued sample	**% receiving reminder letter**	**Marginal cost per letter**	**Total change in costs**
Highest priority cases	6%	55%	£0.69	+£23
Low priority cases	25%	55%	£0.69	-£95
Lowest priority cases	4%	100%	£0.69	-£28
Total				*-£100*

Number of calls

Estimating the cost impact of the number of calls is most challenging. For this we have assumed that equal proportions of each priority group are issued to CATI and have non-contact outcomes (c.25%). In reality, this may vary – for example lowest priority cases may be more likely to be issued to CATI as they will not complete online, but this may be balanced by them being less likely to have provided a phone number.

Based on this, and proportions of the sample in each group, Table G:3 estimates the cost impact per 1,000 issued cases, assuming a marginal cost of c.33p per call.

Appendix table G:1 Cost impact/1000 issued cases of changing telephone contact protocols					
	% of issued sample	**% issued to telephone fieldwork and non-contact**	**Change in no. of calls per non-contact case**	**Total change in no. of calls**	**Change in cost**
Highest priority cases	6%	25%	2	+30	+£10
Highest priority cases	20%	25%	2	+100	+£33
Low priority cases	25%	25%	-2	-125	-£42
Lowest priority cases	4%	25%	-6	-60	-£20
Total				*-55*	*-£18*

Based on these figures, although the changes in letter and telephone contact protocols do save some money, there is therefore a net increase in costs of c.£400 per 1,000 issued cases, driven by the increased incentive costs. Although this is not entirely 'cost-neutral' as intended, it is a relatively small proportion of the total costs.

Appendix H. Correlation analysis

Figure H:1 shows the correlation between the 'overall' R-indicator score for a given panel wave and the 'overall' DEFF for a given panel wave (i.e. not separated out by BSA year).

As both are measures of sample representativeness, despite there being differences between the precise nature of what they measure, we would expect them to be negatively correlated, but we instead see a slight positive correlation – though given the small variation in each, this may be due to random fluctuation.

Appendix figure H:1 R-indicators against DEFF

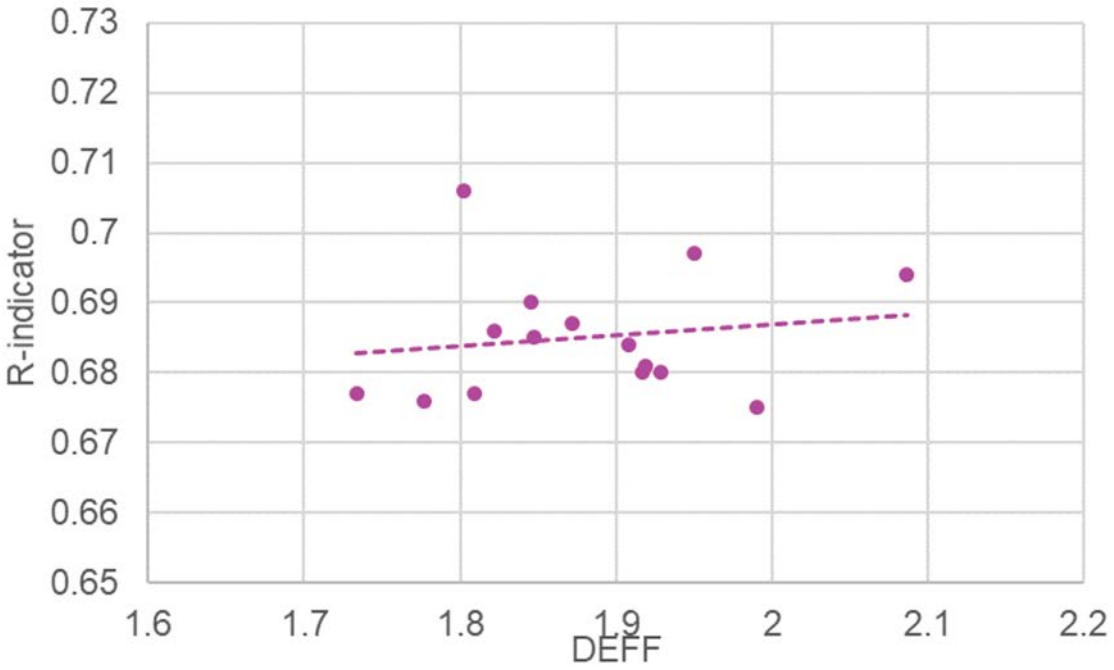


Figure H:2 shows the correlation between R-indicators and the 'recruited' response rates (both using a base of participants in BSA). Contrary to what may be expected, these show a negative correlation for participants recruited from BSA 2015, 2016, and 2017, with coefficients of -0.80, -0.76, and -0.79 respectively. This supports the assertion in Section 4 that response rates may be poor indicators of sample representativeness

Appendix figure H:2 ‘Recruited’ response rate against R-indicators by BSA wave