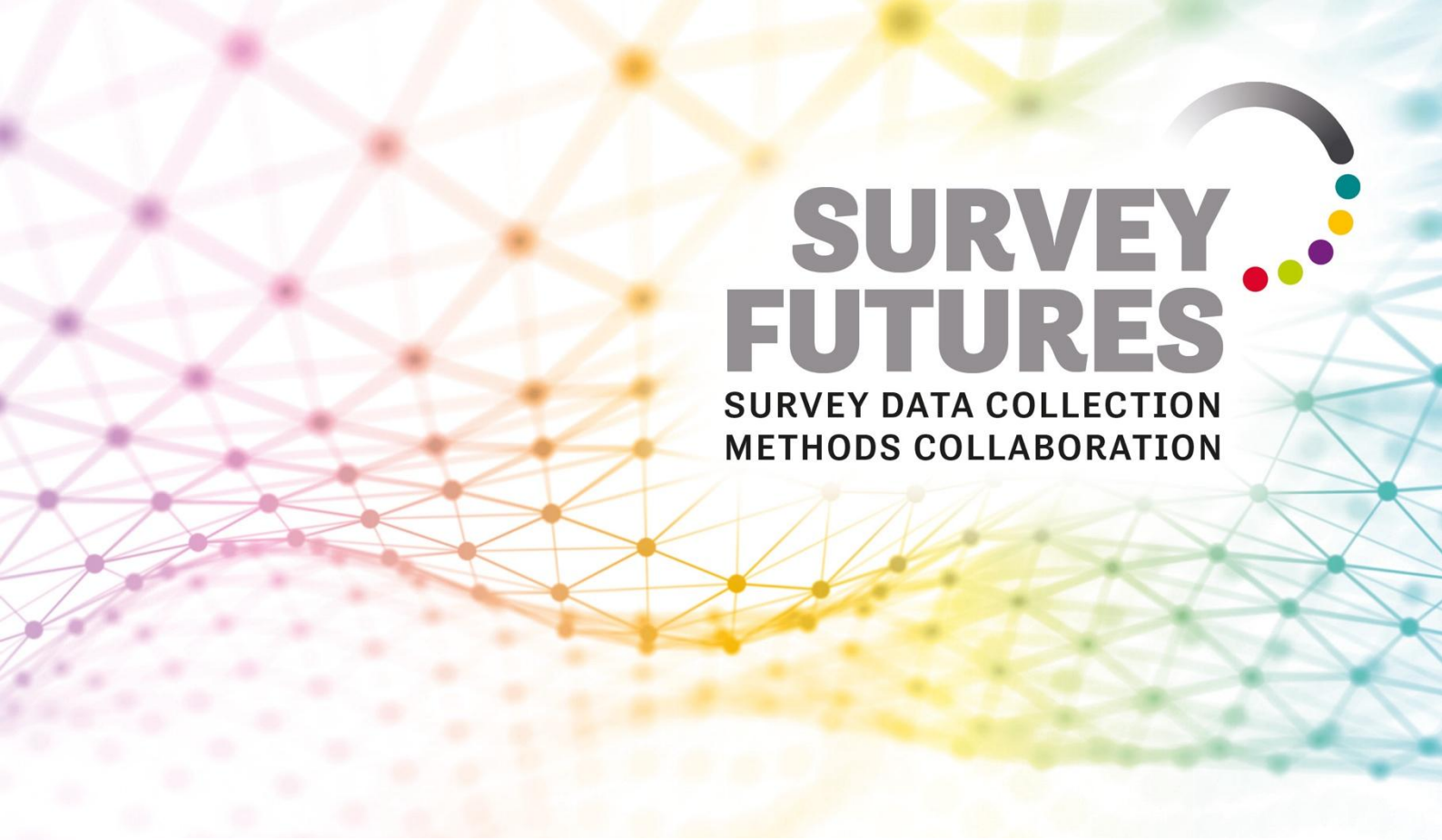



# Survey Practice Guide 2: How to mitigate against measurement effects when surveys move online

*A Measurement Effect Risk Framework (MERF) and web questionnaire guidance*

Jo d'Ardenne, Richard Bull, Aditi Das, Zac Perera & Olivia Sexton (National Centre for Social Research)

June 2025

www.surveyfutures.net

*Survey Futures* is an Economic and Social Research Council (ESRC)-funded initiative (grant ES/X014150/1) aimed at bringing about a step change in survey research to ensure that high quality social survey research can continue in the UK. The initiative brings together social survey researchers, methodologists, commissioners and other stakeholders from across academia, government, private and not-for-profit sectors. Activities include an extensive programme of research, a training and capacity-building (TCB) stream, and dissemination and promotion of good practice. The research programme aims to assess the quality implications of the most important design choices relevant to future UK surveys, with a focus on inclusivity and representativeness, while the TCB stream aims to provide understanding of capacity and skills needs in the survey sector (both interviewers and research professionals), to identify promising ways to improve both, and to take steps towards making those improvements. Survey Futures is directed by Professor Peter Lynn, University of Essex, and is a collaboration of twelve organisations, benefitting from additional support from the Office for National Statistics and the ESRC National Centre for Research Methods. Further information can be found at www.surveyfutures.net.

# Contents

# Glossary

**Accessible design:** Accessible design, in the context of this document, refers to questionnaires that are designed to be usable by disabled people who have additional needs arising from, for example, visual, motor or cognitive impairments. Accessible design aims to remove barriers that may prevent these individuals from being able to participate in web surveys.

**Acquiescence bias:** The tendency for survey respondents to be more likely to answer 'Yes' or 'Agree' to questions compared to 'No' or 'Disagree' regardless of the content of the question.

**Automated prompts:** Computerised messages that are displayed to interviewers or respondents. These are triggered by a user action e.g. messages about inconsistent or out of range answers.

**CAPI: Computer-Assisted Personal Interviewing**. A face-to-face interviewer administered survey, where interviewers enter respondents' answers into a questionnaire run on the interviewer's laptop or tablet.

**CASI: Computer-Assisted Self Interviewing.** Where respondents complete a self-completion questionnaire as part of CAPI interview. This is usually done on the interviewers' device (laptop or tablet) but without the interviewer being able to see the respondents' answers.

**CATI: Computer-Assisted Telephone.** Where interviewers administer a questionnaire over the telephone. Data is entered into the interviewer's PC, laptop or tablet.

**CAWI: Computer-Assisted Web Interviewing.** A self-completion mode where respondents are provided with a link to an online questionnaire that they can complete on their own web-enabled device (mobile, tablet, laptop or PC).

**Device effect:** Where data are influenced by what type of device the survey is completed on (e.g. PC, laptop, tablet or mobile). Differences in screen-size, visual display and input methods (e.g. touch screen versus keyboard) can contribute to device effects.

**Interviewer effect:** Where data collected are influenced by whether or not an interviewer is present.

**Measurement effect:** See 'Mode effect.'

**Mode:** The way in which survey data are collected. Modes will either be interviewer administered (i.e. where interviewers ask the questions and conduct data entry) or self-administered (i.e. where respondents read the questions and conduct data entry). Example survey modes include CAPI, CATI, CAWI and so on.

**Mode effect:** Where data are influenced by the mode of data collection. Mode effects are attributable to both 'selection effects' and 'measurement effects.' Selection effects occur when the mode impacts *who* participates in a survey, for example by impacting on response rates or the final sample composition achieved. Measurement effects occur when the mode impacts *how* participants answer questions, for example via presentation effects or interviewer effects.

**Mode transition:** Where survey designers redesign specific elements of a survey to accommodate a change in mode.

**Non-differentiation:** A form of satisficing behaviour where respondents give the same, or very similar answers, to multiple questions.

**Positivity bias:** The tendency for respondents to give positive answers to attitudinal questions (indicating they are happy or satisfied) compared to negative answers.

**Presentation effect:** When the visual design of a question impacts the data collected. Examples of presentation effects include 'primacy effects' -the tendency for response options displayed at the top of a list to be selected more often than those at the bottom- and 'straight-lining' -the tendency to pick the same answer for every question if responses are displayed in a matrix or grid format.

**Primacy effect**: See 'Presentation effect.'

**Satisficing:** Where survey respondents provide 'lower effort' answers rather than fully considered responses. This phenomenon occurs when respondents aim to complete the survey with minimal effort due to lack of interest, lack of motivation or perceived time constraints. Examples of satisficing behaviours include not reading response options in full, non-differentiation, skipping questions, giving approximations and so on.

**Selection effects:** See 'Mode effect.'

**Social desirability bias:** The tendency for respondents to give answers that conform to social norms, or to present themselves in a positive light. This leads to over-reporting of some behaviours and attitudes, and under-reporting of others

**Straight-lining:** See 'Presentation effect.'

# 1. Background and context

Many probability-based surveys, which have historically only been conducted face-to-face, are now transitioning to alternative modes. Online modes are increasingly being used as a cost-efficient mechanism for data collection, either as a standalone mode or in combination with another modes (e.g. web followed by a CAPI follow-up, or web-CATI combinations). A key issue for survey practitioners is how to effectively introduce online modes of data collection to existing surveys, without introducing avoidable measurement effects.

## 1.1 What is a mode effect and what is a measurement effect?

Mode effects are differences in data that are attributable to the mode of administration (be it CAPI, CATI, web or postal), rather than a real differences in whatever construct the questions are attempting to measure

There are two components of mode effects: selection effects and measurement effects (e.g. Schouten *et al* 2023). Selection effects occur when the mode of data collection impacts *who* participates in a survey. It is known that mode choice impacts sample frame availability, response rates and the final sample composition achieved (e.g. Dillman 2017). In contrast, measurement effects occur when the mode of data collection impacts *how* participants answer the survey questions. The presence (or absence) of interviewers, how the questions are presented and how 'user-friendly' the instruments are can impact the data collected, in addition to any selection effects.

All survey data may be impacted by both selection and measurements errors to some extent. However, introducing a secondary mode, or changing mode of data collection part way through a survey time-series, can change the nature of the *errors*, thereby introducing mode *effects* that act as a confounding factor during data analyses. It is important for survey practitioners to consider the risk of mode effects when transitioning a survey to a new mode, and to take mitigating steps.

## 1.2 Purpose of this document

The primary purpose of this document is to provide survey practitioners with a 'Measurement Effect Risk Framework' (MERF) that can be used to help transition interviewer administered questionnaires to online modes. A secondary purpose of the MERF is to help practitioners identify risks that could occur when combining online modes with other modes of administration (e.g. web-CAPI and web-CATI combinations). Using the framework, practitioners can identify questions that may be at a higher risk of measurement effects occurring and take steps to mitigate against some of these risks as far as practicable. This document also contains practical guidance for the development and testing of web questionnaires.

Please note this document does not attempt to address how to reduce selection effects when introducing an online mode, nor how to test for mode effects during analysis.

## 1.3 How the Measurement Effect Risk Framework was developed

The National Centre for Social Research (NatCen) first developed a measurement risk framework for the UK Household Longitudinal Study (Understanding Society) to support

the transition of the survey to a mixed mode design (d'Ardenne *et al*. 2017). This original framework was based on a review of sources of measurement error presented by Campanelli *et al* in 2011 and the Questionnaire Appraisal System –QAS (Willis and Lessler, 1999). The original framework has been used to review a variety of surveys (including the English Longitudinal Study of Ageing, the English Housing Survey and the Skills and Employment survey) as one method of assessing their suitability for mode conversion.

As part of the Survey Futures programme, this framework has been refined and updated. Changes made have been made as a result of:

1. A literature review on measurement effects and mitigations
2. Collecting feedback from survey practitioners who had used the original framework (to gain insight into how it could be improved)

This document presents the revised framework.

## 1.4 How to use the Measurement Effect Risk Framework

The Measurement Effect Risk Framework (MERF) is described in section two of this document. It lists fourteen features that make a survey more at risk of measurement effects occurring, if it is transitioned to an online mode or asked in a mixed-mode survey. Risks are classified under three headings: Interviewer effects, respondent satisficing and presentation effects.

The MERF includes:

1. A description of each type of risk, including examples of literature where this risk is referred to.
2. How to establish if a risk factor is present or not; and
3. Options for mitigations.

When using the MERF the objective is to systematically review every question in the survey. If risks are flagged practitioners can take steps to introduce mitigations where possible. Where resources allow, we recommend double coding (where two reviewers independently assess each question). Strict application and dual coding will maximise the likelihood that all potential issues are detected. A simplified MERF checklist is provided in Appendix A, this can be used alongside the full MERF instructions to document findings.

## 1.5 Caveats to the Measurement Effect Risk Framework

In this section we present caveats to the use of the Measurement Effect Risk Framework.

### Overlapping risk factors

The framework groups risk factors under three headings; interviewer effects, respondent satisficing and presentation effects. In practice, where the literature identified examples of measurement effects, it is not always clear what mechanism is causing these. For example, our literature review found multiple authors describing how attitudinal scales generate different responses between modes i.e. with more positive

responses in interviewer administered modes, compared to more negative responses in self-completion modes (e.g. Dillman 2017, Schork *et al* 2021, Stefkovics, 2022, Fergusson *et al* 2022). However, different authors attribute this finding to different causes. Some describe how respondents want to appear more agreeable to interviewers (e.g. the effect is caused by positivity bias and/or social desirability bias). In contrast other authors attribute this to presentation effects (e.g. primacy effects). In practice it could be either or a combination of these factors driving measurement differences. Our aim in developing the framework was to map all possible risks, and to provide as many mitigations as practicable. The allocation of risks to headings remains subjective and we are not attempting to definitively attribute causal mechanisms.

## Selection effects

In many cases, where mode effects were documented in the literature review, it was not possible to tell whether differences in data were caused by measurement effects, by selection effects (i.e. differences in the final sample composition achieved that could not be corrected for during analyses) or a combination of both. Questionnaire design cannot mitigate against selection effects. Therefore, following the mitigations listed in the framework will not guarantee that all mode effects can be avoided.

# 2. Measurement Effect Risk Framework (MERF)

## 2.1 Risk of interviewer effects

Interviewer effects are when the presence or absence of an interviewer creates differences in how people respond. In some cases a transition to a self-completion mode (e.g. from CAPI to web) could improve data quality by reducing interviewer effects, such as socially desirable reporting and positivity bias.

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **A1: Socially desirable responses** | Embarrassing, illicit or illegal behaviours are more likely to be reported in self-completion modes compared to interviewer administered modes (Tourangeau *et al* 2000).<br><br>Socially desirable responses are more likely in interviewer administered modes compared to self-completion modes e.g. not smoking, practicing food hygiene, healthy eating, sexual behaviour and contraceptive use (Kim and Cooper 2021, Fergusson 2022, Adali et al 2022).<br><br>Reports may also vary between modes if there are different levels of risk of responses being seen or overheard by bystanders (e.g. Adalı *et al* 2022). | Could participants edit their answers to 'look good' in front of an interviewer? This applies to both behavioural and attitudinal topics.<br><br>Could there be negative consequences for the participant if the information given was overheard? | A move from a CAPI mode to a self-completion mode may improve data quality (i.e. by reducing socially desirable reporting) but could compromise survey time series data. If transitioning to an online survey, the change of mode should be flagged when conducting time-series analysis.<br><br>If mixing modes i.e. combining web and CAPI, consider asking all the questions susceptible to this risk in self-completion mode (e.g. in CAPI these items should be included in CASI) in order to prevent between-mode measurement effects.<br><br>It is considered good practice to encourage honesty and confidentially when introducing a survey to attempt to address socially desirable reporting. However, we cannot fully mitigate against interviewer effects via including such introductions. |

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **A2: Sensitive question** | Some questions may be considered 'personal' 'intrusive' or 'taboo' even if they do not include socially desirable responses (Tourangeau, 2000). We refer to these questions as being sensitive. Many socio-demographic questions are considered sensitive and there are higher levels of refusal for socio-demographic questions in interviewer modes compared to self-completion modes (e.g. Soszynski, 2023a). Higher levels of refusal are particularly notable for income questions (e.g. Valet et al 2019 found 19% refusal in CAPI compared to 10% refusal in self-completion). | Could the question be considered personal or intrusive?<br><br>Does the question attempt to measure DOB, age, gender, ethnicity, health, disability income or any other socio-demographic that could be considered sensitive? | We cannot fully mitigate against this interviewer effect. If transitioning to an online survey, the change of mode should be flagged when conducting time-series analysis.<br><br>If mixing modes i.e. combining web and CAPI, consider asking all the flagged questions using a self-completion mode (e.g. for CAPI surveys these items should be included in CASI). This is to prevent within-wave measurement effects. Interviewer training regarding how to handle refusals should also considered in cases where a self-completion module is not viable.<br><br>Differences in socio-demographics profiles between modes should always be included in reports where a survey has changed mode or if a new mode has been introduced. Note that selection effects will also drive differences in socio-demographic profiles between modes. |
| **A3: Rating scale** | Rating scales appear sensitive to mode effects. Participants are more likely to 'agree' to agree/disagree scales in interviewer administered modes (e.g. Fergusson et al 2022). Reported levels of satisfaction are lower in self-completion modes compared to both CAPI and CATI interviewer administered modes (e.g. Schork et al 2021, Sozynski 2023, Agraib 2023).<br><br>Differences in rating scale data between modes have been noted even in cases where questions do not appear to have a socially desirable response (Dillman, 2017). | Is the question asking the participant to give an attitude using a scale?<br><br>Include of agree/disagree scales and satisfaction scales (excellent-poor or 0-10 rating scales etc.) | We cannot fully mitigate against this interviewer effect. If transitioning to an online survey, the change of mode should be flagged when conducting time-series analysis.<br><br>If mixing modes i.e. combining web and CAPI, consider asking all the flagged questions using a self-completion mode (e.g. for CAPI surveys these items should be included in CASI) in order to prevent measurement effects. |

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **A4: Knowledge and skills tests** | Measurement effects have been noted for both questions that attempt to measure objective knowledge of a factual topic (e.g. Agraib et al 2023) and measures of cognitive functioning (e.g. Al Baghal, 2019).<br><br>Higher scores in knowledge and cognitive functioning tests have been noted in CAWI compared to interviewer administered modes. It is unclear whether this is being driven via selection effects, respondents 'cheating' in self-completion tests (e.g. looking up answers/conferring) or whether interviewer presence somehow impedes test performance. | Is the question aiming to assess objective levels of respondent knowledge on a factual topic? Is it a test of cognitive functioning (a memory test, etc.)? | We cannot fully mitigate against this interviewer effect. If transitioning to an online survey the change of mode should be flagged when conducting time-series analysis.<br><br>There is no evidence that this measurement effect can be mitigated against via changing question wording. If mixing modes i.e. combining web and CAPI, consider asking all the flagged questions using a self-completion mode (e.g. for CAPI surveys these items should be included in CASI). |

## 2.2 Risk of satisficing

'Satisficing' refers to respondents adopting non-optimal approaches to questionnaire completion (Krosnick, 1991). Examples of satisficing include, not reading questions in full before deciding on a response, skipping questions, non-differentiation and so on. It is generally assumed that satisficing is greater if questionnaires are perceived as burdensome. Interviewers can partially mitigate against burden as they can read out the required text, help with definitions and conduct all data entry. Perceptions of cumulative burden can trigger break-off (abandoning the questionnaire).

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **B1: Question length and complexity** | In CAPI/ CATI, interviewers are trained to read out the entire question. In self-completion modes participants may not read the entire question, especially if it is long-winded. Respondents are more likely to break-off at longer questions (Peytchev, 2009).<br><br>Aside from length, use of jargonistic or technical language can also increase burden. Questions that use 'plain language' have been shown to decrease item non-response, and increase levels of differentiation i.e. respondents give more varied responses (Bauer et al 2023)<br><br>Some CAPI/ CATI scripts include optional interviewer read outs such as clarifications and definitions. When transitioning questionnaires to self-completion modes it is preferable to avoid optional information, but rather to design questions to avoid reliance on extra help. | Is the question more than one sentence in length?<br><br>Does it include multiple inclusion and exclusion criteria?<br><br>Does it include jargonistic or technical language?<br><br>Does the question contain any optional interviewer help-text? | The aim is to cut all superfluous text from a question stem whilst still retaining the same meaning. Ideally the question should be under 250 characters.<br><br>Simplify language as far as possible.<br><br>If a question includes multiple inclusion/exclusion criteria, consider breaking the question down into a series of single clause questions.<br><br>Help screens can be included as alternatives to optional interviewer read-outs but it should not be assumed that participants will consistently read these. The default position should be to design questions that do not rely on help screens where possible (Wilson & Dickenson, 2021). If used, help screens should use a heading or hyperlink that explicitly says what extra information they provide e.g. 'How to answer if you have more than one home' rather than 'Help' or 'More information.' |
| **B2: Looped questions** | Looped questions (where people are asked the same series of questions multiple times) are considered burdensome by some respondents. There is evidence that break-offs on web surveys are more likely to occur with a looped as opposed to a non-looped question, with the effect being more marked for smartphone users (Emery, 2023) | Is it a looped series of questions, with the same series asked multiple times? Examples include series of questions about each household member, a series of questions about different products or services etc | Reduce the number of looped questions as far as possible to minimise risk of break-off. If loops are needed, reduce length of each loop, and consider introducing a maximum number of loops per person. Consider asking each loop in different questionnaire sections to reduce perceptions of monotony. |

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **B3. Calculations** | Some questions involve participants having to recall different events and add them up (e.g. number of days affected by health problems in last year, number of alcoholic beverages drunk per week). Some questions involve transformations (e.g. asking participants to provide income in a weekly format when they receive it monthly or vice versa). Interviewers can play a role in encouraging participants to give an answer or to give more accurate responses.<br><br>In self-completion modes respondents are more likely to adopt 'short-cutting' strategies to avoid calculations. They are more likely to give 'don't know' responses (e.g. Lipps & Monsch 2022) or rounded answers (Schnell et al 2022). Schnell et al. (2022) found that automated prompts can be effective in online surveys to reduce rounding of answers (e.g. checks asking if this is an exact or rounded number, in cases where response ends in a 0). | Is a numerical response required. either a number or a response from a list of numeric bands?<br><br>Does the question require mental calculations e.g. adding, subtracting, division or averaging? | Consider whether there is any way that the calculation element can be removed or reduced. For example, redesign questions where participants can select their own reference period (per week/ per month/ per year) rather than having a fixed reference period.<br><br>Consider the granularity of information requested e.g. could open numeric questions be converted to banded numeric questions? Could highly granular numeric bands be collapsed into less granular bands?<br><br>Consider the use of follow-up questions in case of 'don't know' responses (e.g. if people decline to provide an exact number, ask them to provide an estimate in a less granular banded format)<br><br>Consider the use of prompts in CAWI to reduce rounding (see section 3.5 for more guidance on prompts). |
| **B4. Open questions** | Less information is given in open questions in self-completion modes compared to interviewer administered modes. In web surveys, break-offs are more common at open questions compared to other question formats, due to the perceived burden of answering such questions (Peytchev, 2009).<br><br>Respondents on smartphones tend to give shorter answers than respondents on larger devices (e.g. Wenz 2024) and limiting the need for typing is recommended (e.g. Antoun et al 2019). If open questions are retained, longer data entry fields (i.e. bigger boxes) can increase the number of characters entered into the response field (Toepoel, 2016).<br><br>One reason for using open questions is to allow for in-office coding against more complex code frames (e.g. | Is a completely open textual answer required? Exclude short textual answers (e.g. name or address fields).<br><br>Does the question collect data that is later coded in office? | Keep the number of questions that require open textual responses to a minimum for web questionnaires to reduce burden.<br><br>Use larger data entry fields (text boxes) to encourage longer responses.<br><br>Consider the use of alternative formats to open questions i.e. pre-coded lists.<br><br>Consider drop-downs and automatic lookups for more complex code frames. Alternative formats for complex code frames will require development resources and user-testing. Always offer an open text box alternative (e.g. Other specify) for people who fail to find codes they are looking for using alternative formats. |

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| | coding of occupation, industry or medical conditions). Short responses in this case can prevent data being successfully coded.<br><br>There is some evidence to suggest that alternative formats to open text entry (drop-downs and auto lookup functions) can increase the volume of codable data achieved online (e.g. Couper & Zhang 2016). However, there is also evidence that these formats can introduce break-offs if they are difficult for respondents to use (Couper & Zhang 2016**).** | | Consider the use of LLMs to probe responses and/or to pre-code open responses that can then be used in subsequent question filters. |

## 2.3 Risk of presentation effects

Some measurement effects are associated with how questions are presented in different modes. Questions may have differences in visual presentation between modes (i.e. CAPI, web and paper) and within modes (e.g. web surveys completed on different device type). Some modes (i.e. CATI) have no mechanism for visual presentation at all. In the following section we discuss presentation effects and mitigations against these.

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **C1: Long lists** | Mode is linked to the risk of order effects occurring. If participants are provided with a visual list of response options, they are more likely to pick responses near the top of the list (a primacy effect), whereas if they hear a response list they are more likely to pick a response from later in the list (a recency effect). These order effects are attributed to respondents not always reading lists in full in visual modes (e.g. Hohne & Lenzner 2018) or more cognitive burden in auditory modes (e.g. Schouten 2022).<br><br>Web questions do not consistently appear more prone to primacy effects compared to other visual self-completion modes (e.g. Clement 2023). Device type (mobile versus large screen) does not appear to consistently exacerbate primacy effects (e.g. Clement 2020) | Are five or more answer options offered to participants?<br><br>Are you transitioning from a non-visual mode (CATI or CAPI without a showcard card) to a visual mode (CAWI)?<br><br>Are you combining non-visual and visual modes? (e.g. web and CATI) | For rating scales: consider randomising scale direction between participants (i.e. some get negative responses first and others get positive responses presented first). This is to ameliorate order effects on aggregate. This can be done for all electronic modes (e.g. CAWI and CATI).<br><br>For categorical lists: consider reducing the number of response options as far as possible to reduce primacy effects. Consider randomising non-ordered or hierarchical lists.<br><br>If combining CAPI and CAWI ensure that response options in both modes are always presented visually (e.g. always using showcards or show-screens in CAPI) |

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **C2: Check all that apply (CATA)** | Check all that apply (CATA) questions yield different response distributions between CAWI and CATI modes (e.g. Dillman, 2017 & Fergusson 2022). More responses could be selected in a CATI 'Yes or No' format compared to a visual CATA format.<br><br>Differential reporting between CATA lists and 'Yes or No' formats are due to respondents being less likely to fully read long lists of CATA responses. In 'Yes or No' formats participants are forced to consider all response options on a list. | Are you transitioning from a non-visual mode (CATI or CAPI without a showcard card) to a visual mode (CAWI)?<br><br>Are you combining non-visual and visual modes? (e.g. web and CATI)<br><br>If so, is the question a check all that apply (CATA) question? | Keep CATA lists as short as possible to encourage full reading in CAWI/other visual modes.<br><br>Consider randomising order of CATA options – this is to ameliorate primacy effects on aggregate level rather than at the respondent level.<br><br>If combining CAWI and CAPI ensure that response options in both modes are presented visually to ensure consistent presentation (e.g. using showcards or show-screens in CAPI rather than a Yes or No format)<br><br>If combining CAWI and CATI note that measurement effects could occur. Consider using Yes or No formats in both modes. However, note the trade-off here is increased administration times in CAWI, which have associated risk of increasing break-off and increasing item non-response (Peytchev, 2009, Lipps & Monsch 2022). The decision as to whether to use Yes or No in CAWI should be made with consideration of what impact this could have on overall administration time based on the number of CATA items in the questionnaire as a whole. |
| **C3: Hidden codes** | Some CAPI/CATI questionnaires include interviewer observations or hidden codes (i.e. response options that are not overtly offered to respondents but can be selected by interviewers). When transitioning questionnaires to online modes, decisions need to be made regarding whether these hidden codes are adjusted for self-completion modes (i.e. so they can be shown to respondents) or whether they are dropped entirely. | Does the item include an interviewer observation that is not read out to respondents?<br><br>Are there any spontaneous or 'hidden' answer codes which are not shown to participants? | It is recommended that if a survey transitions to an online mode all hidden interviewer codes and checks are reviewed to see whether they should be dropped or adapted to be 'respondent facing.' In some cases this may involve developing new respondent facing code frames. Questions this applies should be flagged when conducting time-series analysis.<br><br>The treatment of hidden 'Don't know' or 'Refusal' codes is discussed separately in section 3.4. |

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **C4. Interviewer checks** | Some CAPI/CATI questionnaires include interviewer check messages that are not shown to respondents. These are to prevent out of range responses or 'unlikely' combinations of responses.<br><br>During a mode transition decisions need to be made regarding whether these checks are retained in CAWI. High volumes of check messages could detract from respondent experience and trigger break-off, especially in cases where it is not clear how check messages should be resolved. The visual design and positioning of dynamic check messages is important regarding whether these are effective (e.g. Kunz and Fuchs, 2019).<br><br>Checks that may require respondents to change answers in earlier questionnaire sections may be particularly problematic (e.g. participants might be expected to change an answer on the current web page but not to navigate back to an earlier point in the questionnaire to correct an inconsistency). | Are there any interviewer checks for this question? | All interviewer check messages should be reviewed to ascertain whether they are appropriate for respondents answering in CAWI.<br><br>Check messages should be used sparingly as overuse could be a source of survey termination.<br><br>Check messages should give clear instructions on resolution appropriate for lay users. They need to be displayed in the same area of the screen as the answers which require amending.<br><br>Check messages that ask respondents to navigate back to earlier sections should be avoided.<br><br>Check messages should undergo user-testing to ensure that the format works on different devices and that users are able to resolve the messages if they occur.<br><br>Further guidance on checks is provided section 3.5 |

| Type of risk | Description | Is risk factor present? | Implications, mitigations and contingencies |
|---|---|---|---|
| **C5: Visual aids** | Some questions include visual aids to explain concepts (for example diagrams or pictures).<br><br>Some questions have response options that include a visual component. These include visual analogue scales (e.g. a slider where people indicate where they fall along a line or bar), smiley face rating scales or ranking tasks (i.e. where people are asked to drag and drop responses into a new preferred order) | Does the question rely on a visual aid?<br><br>Do response options rely on a visual component – e.g. a slider or getting people to drag and drop responses to form a ranked list?<br><br>Is there anything else about the question that makes it unusual in terms of visual design? | Questions with visual components should be subject to user-testing to ensure that the format works on different devices and that users are able to understand/ interact with the visual feature as intended.<br><br>Questions that rely on visual prompts should be avoided in mixed mode surveys that involve CATI. Consider switching to a question format that is appropriate for all modes of administration; For visual analogues scales and/or smiley face scales consider 0-10 scales as an alternative. For ranking tasks consider a battery of discrete choice questions (would you prefer X or Y?). Use the same question format for all modes. |
| **C6. Batteries of questions/ Grids** | Many questionnaires involve batteries of questions that repeatedly use the same scale. Non-differentiation (i.e. where respondents give the same answer response to every question in a battery) is impacted by visual design in CAWI. For example, Verbree *et al* (2020) describes how non-differentiation is higher in desktops compared to smartphones and tablets. This is likely to be due to differences in display e.g. large screens use grid-based formats that are prone to straight-lining whereas smaller screens use alternative to grid formats that are less prone to straightlining. | Is your question part of a battery of questions using the same scale (i.e. that could be displayed as a grid in CAWI?) | If batteries of questions are included it is important to make formats consistent across devices as far as possible. This means avoiding grid-based formats, including on larger screen devices. Alternative displays to grids are discussed in more detail in Chapter 3. |

# 3. Guidance on the development of web questionnaires

As a result of our literature review, we identified a number of recommendations that apply to all questions that are developed for an online mode. In this final chapter we summarise these additional considerations. This chapter will discuss:

- The importance of adopting **mobile-first design** principles.
- **Grid formats and their alternatives** on different device types.
- **Accessibility considerations**, and their role in making web questionnaires inclusive and compliant with accessibility standards.
- Options for the presentation of **'Don't Know'** and '**Refusal**' and the relative strengths and weaknesses of different approaches.
- The role of **automated prompts.**
- The use of **Computer-Assisted Self-Interviewing (CASI**) in web-CAPI surveys.
- The importance of **pre-testing** web questions.

By highlighting these topics, we aim to provide practitioners with guidance that will enhance the quality of their web surveys from both a user-perspective and a data quality perspective.

## 3.1 Mobile-First Design

A repeated theme in the literature reviewed for this study was the importance of adopting mobile-first design principles when designing web questionnaires. Researchers must assume (for general population surveys) that a significant proportion of respondents will attempt to complete the survey via a smartphone. Approximately one in five adults in the UK only ever go online using a smartphone (Ofcom, 2024). 'Smartphone only' individuals are more likely to come from semi-skilled or unskilled manual occupations and/ or unemployed. Therefore, failing to allow for smartphone completion could bias any data collected by excluding key groups of interest.

We would argue that it is crucial to ensure that smartphone completion is not only permitted but optimised. Emery *et al*. (2023) highlight the need for significant improvements in the smartphone experience for online surveys to reduce drop-off rates. This is a very timely and important aspect of current survey design, as the prevalence of smartphone use continues to grow, and ensuring a seamless mobile experience is essential for maintaining respondent engagement and data quality in the future.

Dillman et al. (2017) recommend addressing visual display issues by designing web questionnaires to be compatible with all screen sizes. They suggest removing graphics and logos that might take up a disproportionate amount of screen space on mobile devices and detract from the respondent's focus on the questions. Certain layouts, that may not render well on smaller screens should also be avoided (e.g. such as grids, horizontal scales). Additionally, limiting the number of questions per screen can prevent

respondents from feeling overwhelmed and to reduce scrolling, which can be particularly challenging on smartphones. These measures collectively help create a streamlined survey experience.

## 3.2 Grid formats and their alternatives

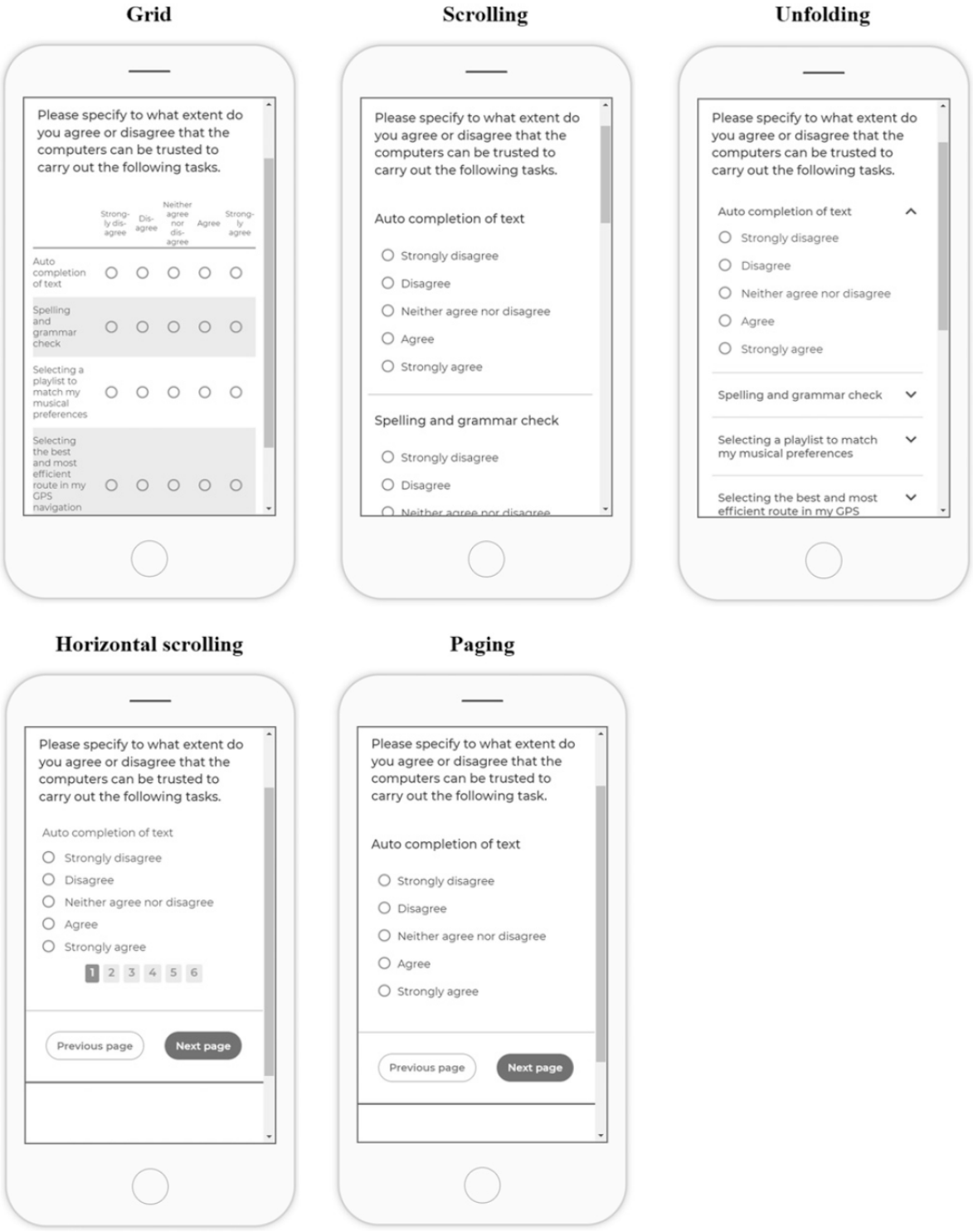
Traditional grids (i.e. table or matrix-based grids) should be avoided in CAWI surveys because they cannot be consistently well rendered on smartphones. Survey designers should prioritise layouts that render well on both larger screens and mobile devices. Using the same design for both large and small screens can prevent device effects, meaning mobile designs should be applied to large screens as well as well as smartphones (e.g. Antoun, 2019).

Vehovar *et al*. (2023) have run experiments looking at the relative merits of grid formats, and their alternatives, in online surveys. They compared the following options:

- **Traditional grids**
- **Scrolling alternatives**, where multiple questions are displayed on a single page, in a vertical list
- **Unfolding alternatives** (often referred to as accordion grids), where multiple questions are displayed on a single page, but each new question only unfolds to become visible after the previous item has been answered
- **Horizontal scrolling alternatives** (often referred to as carousel grids) where multiple questions are viewable on a single page, but each question will appear only if a respondent answers a question, or swipes left or right
- **Paging**, where one question is shown per page.

Figure 1 overleaf illustrates these different approaches.

Figure 1: Grid formats and their alternatives on mobiles (taken from Vehovar *et al*, 2023)




*Note: the main difference between the horizontal scrolling (carousel grid) option and the paging option lies in the navigation mechanism. In horizontal scrolling participants automatically advance to the next grid item when they select an answer. They do not need to select the 'Next page' button. Horizontal swiping can be used to move between grid items, all within the same page. The paging option involves on question per page and clicking 'Next page' to navigate between questions.*

Vehovar *et al*. (2023) note that paging, compared to alternatives, increases the number of breakoffs, increases administration time, and increases self-rated respondent burden. The significant increase in administration time for paging compared to layouts with multiple items per page has also been noted by Mason and Huff (2019). Practitioners have noted that in user testing of paging designs, participants mistakenly believe they are being asked the same question twice. Based on this we recommend that paging alternatives to grids are avoided.

Vehovar *et al* (2023) note that traditional grids, although quicker to complete than paging designs, had longer response times than both scrolling and unfolding formats. Grids were also more prone to straight-lining on both PCs and mobiles. This indicates grid formats should also be avoided regardless of the size of the device.

Horizontal scrolling alternatives (carousel grids) also have issues, most notably higher levels of item non-response, which was more than double that observed in other layouts (Vehovar *et al,* 2023). It is possible that higher levels of item non-response are caused by respondents not noticing that the question changes via auto-advance after they have inputted an answer, leading them to press 'Next Page' without realising there are more questions to answer in the battery.

Therefore, vertically displaying multiple options per page (either using a static design with scrolling or a dynamic design with unfolding) are the two alternative grid formats that should be prioritised by questionnaire designers. These formats should be used consistently for all devices or screen size, to prevent device effects.

## 3.3 Accessible question formats

Another consideration when developing web questionnaires is accessibility. Accessible web formats ensure that online content can be used by disabled people who have specific forms of impairment. For example, questions should be usable for those who rely on keyboard only navigation (due to impaired dexterity) and those who use screen magnification/ screen reader technology (due to impaired vision). Inclusive design is endorsed by the UK Government Statistical Service (GSS) and is part of the Respondent Centred Design Framework developed by Wilson & Dickinson (2022).

At the time of writing (2025) publicly funded digital services (including UK Government sponsored surveys) must meet level AA Web Content Accessibility Guidelines (WCAG 2.2) as a minimum standard and must also always include an accessibility statement[1]. Whereas most common question formats will pass WCAG 2.2 AA standards there are exceptions. Questions that require a 'dragging' method of data entry (e.g. a slider scale operated by a mouse or touchscreen) or 'drag and drop data' entry (e.g. ranking tasks where respondents have to move response options into a preferred order via dragging them using a mouse or touchscreen) would fail to meet the AA rating. This is because these formats, without modifications, are not compatible for people with dexterity or visual impairments who rely on either keyboard entry or screen readers.  WCAG 2.2

[1] Making your service accessible: an introduction - Service Manual - GOV.UK

guidelines also provide information relevant to questionnaire accessibility (e.g. colour contrast guidance, rules on time-outs and so on).

## 3.4 Use of 'Don't Know' and 'Refusal' Options

Item non-response (i.e. missing data, 'don't know' responses and item refusal) can be higher in self-administered modes compared to CAPI and CATI (e.g. Klíma *et al.* 2023).

In CAPI and CATI it is standard for interviewers to be able to input 'don't know' and/or 'refuse' answers using spontaneous codes. By this we mean interviewers can code don't knows and/or refusals as answers, but participants are not explicitly reminded these options are available at each question. For online surveys spontaneous codes are not possible. Four alternative approaches are available online:

1. DK/ Refusals are not available at all, and skipping questions is not permitted.
2. DK/Refusals are always displayed, for respondents to select as they see fit.
3. DK/ Refusals are not included, but skipping questions is permitted.
4. DK/ Refusals are only displayed to respondents who attempt to skip a question (referred to as 'hidden' code subsequently).

The first option is rarely used in practice for social surveys. Forcing respondents to answer all questions (i.e. not offering 'don't know' 'refuse' or the option to skip question) raises ethical concerns. There are also concerns about data quality for option one as respondents who legitimately do not know an answer to a question will have to either enter potentially incorrect data or to terminate the survey. There is some evidence to suggest forcing respondents to answer questions like this can lead to higher survey break-off rates (Kmetty and Stefkovics, 2022). Therefore, option one is not recommended.

In contrast always offering 'don't know' or 'refuse' to online respondents (option two) will increase the number of respondents selecting these options. It is important to note that this increase in 'don't know' responses is primarily thought to be due to satisficing, rather than genuine and valid 'don't know' responses. The selection of 'don't know' or 'refuse' in longitudinal studies increases if these options are explicitly offered as a result of a mode change (e.g. Lipps *et al.* 2023). This is more likely to occur among younger people and those with lower levels of education (Lipps *et al.* 2023). While a rise in 'don't know' responses due to satisficing is generally undesirable, it is not necessarily as problematic as the alternative i.e. suppressing genuine 'don't know' responses.

Some authors conclude that option three (not offering 'don't know' or 'refuse' and allowing skips) is the optimal presentation (e.g. Kmetty and Stefkovics, 2022). The trade-off with this approach is that it results in more data cleaning to determine if a missing value was due to routing (the question wasn't asked) or skipping (the question was asked but not answered). This option can also lead to accidental skips, for example if people move onto a new page without noticing they have not answered a question. This has implications for how questions are displayed. Participants may be more likely to accidentally omit questions if multiple items are displayed on a single page or if auto advance is used (a feature where the survey automatically moves to the next question after a response is entered).

Hidden ‘don’t know’/’refusal’ options (option four) are the final possibility. The advantages of hidden codes are that they reduce the volume of ‘don’t know’ and ‘refuse’ responses, whilst negating the risk of questions being accidentally skipped, and reducing the volume of data cleaning required. However, this format is more burdensome for respondents compared to other options. This is because multiple clicks are required to decline a question rather than just one. Additionally, there is a risk that respondents may not realise that these options are available. If respondents do not find out about these hidden options, it effectively becomes like option one (forced answer), which can lead to similar issues of respondent frustration. There is also some evidence that hidden codes are a less accessible format. Participants who use screen magnifiers may not always be able to easily see that new response codes have been added to a response list when they try to skip a question. Participants who use screen readers will also have a question read out to them twice every time they attempt to skip a question.

All this indicates that the optimal format depends on whether data commissioners wish to prioritise respondent experience and accessibility, minimise the volume of don’t knows/ refusal answers, or optimise data processing. A summary of the pros and cons of different approaches is shown in the table below. We recommend that survey practitioners always highlight these trade-offs with survey commissioners when making decisions on how to present don’t know and refusal options.

Table 3.4: Advantages and disadvantages of different approaches to handling DK/REF in CAWI.

| Option | Advantages | Disadvantages |
|---|---|---|
| **1. No DK/REF.** No skips | • No item level non-response.<br>• Less labour required for data cleaning compared for option 3. | • Associated with higher break-off rates.<br>• Ethical issues if right of refusal is removed.<br>• DK/ REF options are sometimes an interesting or valid response.<br>• Inaccurate data collected in case of legitimate DK responses. |
| **2. DK/REF always displayed** | • Low respondent burden.<br>• Less labour required for data cleaning compared for option 3.<br>• No risk of accidental skips compared to option 3. | • Higher level of DK and REF. compared to all other options<br>• Some of these DK and REF are a result of satisficing rather than legitimate response. |
| **3. No DK/REF.** Skips permitted | • Low respondent burden.<br>• Lower level of item non-response compared to option 2. | • Risk of accidental skips.<br>• No way of knowing which skips are accidental versus which are deliberate.<br>• More data cleaning required (i.e. to ascertain if missing data is due to questions being off-route or skipped). This has cost implications. |
| **4. DK/REF initially hidden**, but will display after a skip is attempted | • Fewer DK and REF compared to option 2.<br>• Less labour required for data cleaning compared to option 3.<br>• No risk of accidental skips compared to option 3. | • Higher respondent burden compared to option 2 and option 3 as multiple clicks required to decline a question.<br>• Respondents may not realise these options are available.<br>• Less accessible format (e.g. problems for groups using magnifiers and screen readers) |

Dillman suggests that whatever mechanism for skips/don't know/refusals is used, it should be consistent between modes to prevent measurement effects. Allowing skips (without a 'don't know' or 'refuse' option) is one format that can be implemented across all modes including paper. Therefore, it should be considered if a web survey has paper as a secondary mode. Hidden codes (i.e. option four) are more akin to what would happen in CAPI or CATI and may be most appropriate for web surveys that have CAPI or CATI as a secondary mode.

When don't know and refusal options are presented to respondents (as is the case for option 2 and option 4) it is beneficial if they are formatted in a way that is visually distinctive compared to the main response options. This can be done using a dividing line or by using a different colour. This visual distinction helps respondents clearly differentiate between options that are part of a scale, and options that are not. Visual distinctiveness of scale options is important so as there is no visual confusion over where the scale mid-point falls (Tourangeau et al, 2004).

## 3.5 Use of automated prompts in CAWI

CAWI questionnaires can include a range of automated prompts or check messages. These include:

- **Hard checks**: Where participants are unable to progress through the questionnaire until they have corrected an error (e.g. an out-of-range numeric answer).
- **Soft checks**: Where a potential error is flagged to participants, but they are allowed to close the warning without taking further action.
- **Speeder checks**: Which are triggered where a respondent appears to be progressing through a questionnaire too fast.

When introducing any form of automated prompts to CAWI the wording and formatting is crucial. Instructions should be clear and concise, informing respondents about how to correct any problem identified e.g. 'Please enter a number between 0 and 10' or 'Select up to three options only' rather than default messages such as 'Answer out of range'. Check messages need to appear within the respondent's field of vision, especially on small screen devices (e.g. Kunz and Fuchs, 2019).

Soft checks have been effectively used in CAWI to reduce non-differentiation and straight-lining, behaviours where respondents select the same response option across multiple questions without fully considering each one (e.g. Fischer & Bayham, 2019; Sun et al., 2022). They have also been used to reduce rounding error (e.g. Schnell et al. 2022). Soft check prompts address these behaviours by encouraging more thoughtful and accurate responses.

Additionally, speeder checks, which flag respondents who complete the survey unusually quickly, can be employed to further reduce straight-lining. For example, a speeder prompt might state, "We noticed that you completed these questions very quickly; please ensure you have taken time to fully read them before continuing." (Sun et al., 2022).

Selective use of automated prompts can improve data quality in CAWI surveys. However, it is important to carefully consider the overall context under which each individual prompt is implemented. Although experiments on individual prompts demonstrate their efficacy, high volumes of check messages within a single survey could contribute to overall survey length and could potentially lead to respondent frustration or drop-off. For this reason, we recommend automated prompts are used sparingly.

## 3.6 Use of CASI in web-CAPI surveys

As discussed in section 2.1 some questions may be prone to interviewer effects (e.g. questions that have socially desirable answers, sensitive questions, rating scales). One mitigation raised in the MERF is that CASI (Computer-Assisted Self-Interviewing) should be considered for all of these questions in a mixed-mode survey that includes both web and CAPI modes. If this practice is introduced, then it will mean a higher proportion of survey questions will need to be administered via CASI in the future. This approach may require interviewers administering multiple 'short' CASI sections during an interview to maintain the same question order and flow as in the web mode. Consideration needs to be given to the practicalities of this approach, such as whether CAPI interviewers can offer respondents their laptops more frequently during the interview rather than having a single standalone CASI block.

In order to minimise the risk of presentation effects, the visual formats used in CASI should mirror those used in web surveys (Al Baghal, 2019). Some authors (e.g. Lipps and Pekari, 2021) suggest that multi-mode approaches should use the same software programme and programming team, as far as practicable, to ensure consistency of presentation.

It has also been suggested that interviewers should have the flexibility to allow respondents to answer *any* questions privately, as this could be a mitigation if a question transpires to be sensitive even if this was not predicted by researchers (Valet et al., 2019). Flexibility of self-completion could be beneficial for questions that start having high refusal rates in a pilot or dress rehearsal The degree of flexibility should be determined during the piloting phase to avoid changing from CAPI to CASI part way through fieldwork.

## 3.7 Pre-Testing

It is generally recommended that new survey questions undergo some form of pre-testing with members of the public prior to a survey being launched. Longstanding interviewer administered questions that have transitioned to an online mode are no exception to this.

It is important to ascertain that respondents are able to understand questions and provide accurate answers without an interviewer being present to assist. In addition, it is important that the survey interface is checked from a respondent's perspective. Respondents need to be able to navigate the web instrument and conduct data entry without assistance, on their own devices. Qualitative 'cogability testing', a method that combines both cognitive interviewing and user-testing, (Wilson and Dickinson, 2022) is recommended as a pre-testing method for web surveys. Cogability testing quotas should always include people who vary in terms of their digital literacy skills and their preferred device type (smartphone, tablet, laptop etc)

Finally, we recommend that questions fielded in multiple modes should also be pre-tested in multiple modes. For example, for web-CATI surveys, pre-testing should occur in both web and telephone modes. This is because ease of use can be confounded by

mode; it should not be assumed that if a question works well in one mode, it will work well in another (Kim and Couper, 2021; Wilson and Dickinson, 2022).

# References


Adalı T.; Türkyılmaz A.S.; Lepkowski J.M. (2021). Evaluating the Demographic and Health Surveys Mode Switch From PAPI to CAPI: An Experiment From Turkey. *Social Science Computer Review*, 40 (6), 1393-1415. https://doi.org/10.1177/08944393211009566

Agraib, L. M., Alkhatib, B., Al Hourani, H., & Al-Shami, I. (2023). Are online and face-to-face questionnaires equally valid and reliable methods of assessing preconception care? *Quality & Quantity*, 57(5563–5576). http://dx.doi.org/10.1007/s11135-023-01614-x

Al Baghal, T. (2019). The Effect of Online and Mixed-Mode Measurement of Cognitive Ability. Social Science Computer Review, 37(1), 89-103. https://doi.org/10.1177/0894439317746328

Antoun, C., Conrad, F.G., Couper, M. P., & West, B. T. (2019) Simultaneous Estimation of Multiple Sources of Error in a Smartphone-Based Survey. *Journal of Survey Statistics and Methodology*, 7, 93-117. http://dx.doi.org/10.1093/jssam/smy002

Bauer, I., Kunz, T., & Gummer, T. (2023). Plain language in web questionnaires: effects on data quality and questionnaire evaluation. *International Journal of Social Research Methodology*, *28*(1), 57–69. https://doi.org/10.1080/13645579.2023.2294880

Campanelli, P., Nicolaas, G., Jäckle, A., Lynn, P., Hope, S., Blake, M., & Gray, M. (2011). A classification of question characteristics relevant to measurement (error) and consequently important for mixed mode questionnaire design. *Royal Statistical Society Presentation, October*, *11*.

Clement S. L., Severin-Nielsen M. K. & Shamshiri-Petersen, D. (2020). Device effects on survey response quality. A comparison of smartphone, tablet and PC responses on a cross sectional probability sample. *Survey Methods: Insights from the Field, Special issue: 'Advancements in Online and Mobile Survey Methods'* https://surveyinsights.org/?p=13585

Clement S.L., Severin-Nielsen M. K. & Shamshiri-Petersen, D. (2023). Satisficing Behaviour in Web Surveys. Results from a Comparison of Web and Paper Mode across Four National Survey Experiments. *Survey Methods: Insights from the Field* https://surveyinsights.org/?p=16640

Dillman, D., Hao, F. Morgan, M. (2017) Improving the Effectiveness of Online Data Collection by Mixing Survey Modes, *The Sage Handbook of Online Research Methods (Second Edition)* Eds Fielding N.G, Lee, R.M & Blank, G, Chapter 13

Emery, T., Cabaco, S., Fadel, L., Lugtig, P., Toepoel, V., Schumann, A., Luck, D., & Bujard, M. (2023). Breakoffs in an hour-long, online survey. *Survey Practice*, 16 (1). https://doi.org/10.29115/SP-2023-0008

Ferguson, Martine, Amy M. Lando, Fanfan Wu, and Linda Verrill. (2022), Transitioning the FDA Food Safety and Nutrition Survey from RDD to ABS. *Survey Practice,* 15 (1). https://doi.org/10.29115/SP-2022-0003

Fischer C.S.; Bayham L. (2019) Mode and Interviewer Effects in Egocentric Network Research. Field Studies, 31(3). https:/doi.org/10.1177/1525822X19861321

Hohne, J. K., and T. Lenzner (2018). New insights on the cognitive processing of agree/ disagree and item-specific questions. *Journal of Survey Statistics and Methodology* 6:401–17. https://doi.org/10.1093/jssam/smx028

Kim, S., & Couper, M. P. (2023). A National RDD Smartphone Web Survey: Comparison With a Large-Scale CAPI Survey. Social Science Computer Review, 0(0). https://doi.org/10.1177/08944393231222675

Klíma O.; Lakomý M.; Volevach E. (2023). Impacts of cultural factors and mode of administration on item nonresponse for political questions in the European context. International Journal of Social Research Methodology, 27(4). Https:/doi.org/10.1080/13645579.2023.2175921.

Krosnick, J. (1991). 'Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys', *Applied Cognitive Psychology*, 5, pp. 213–236. https://psycnet.apa.org/doi/10.1002/acp.2350050305

Kunz T.; Fuchs M. (2019). Dynamic Instructions in Check-All-That-Apply Questions. Social Science and Computer Review, 37(1). https:/doi.org/10.1177/0894439317748890

Lipps, O., & Monsch, G.-A. (2022). Effects of Question Characteristics on Item Nonresponse in Telephone and Web Survey Modes. Field Methods, 34(4), 318-333. https://doi.org/10.1177/1525822X221115838

Lipps, O. & Pekari, N. (2021). Sequentially mixing modes in an election survey. Survey Methods: Insights from the Field. Retrieved from https://surveyinsights.org/?p=15281

Lipps O.; Voorpostel M.; Monsch G.-A. (2023). Effects of Changing Modes on Item Nonresponse in Panel Surveys. Journal of Official Statistics, 39(2). Https:/doi.org/10.2478/jos-2023-0007.

Mason R.; Huff K. (2019). The effect of format and device on the performance and usability of web-based questionnaires. International Journal of Social Research Methodology, 22(3). Https:/doi.org/10.1080/13645579.2018.1542150.

Ofcom (2024). Adults Media Use and Attitudes Report, available here: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/adults/adults-media-use-and-attitudes-2024/adults-media-use-and-attitudes-report-2024.pdf?v=321395

Peytchev, Andy. (2009). Survey Breakoff. Public Opinion Quarterly. 73 (1) http://dx.doi.org/10.1093/poq/nfp014

Schouten, B., van den Braken, J., Buelens, B., Giesen, D., Luiten, A. & Meertens, V. (2020) Mode-Specific Measurement Effects in *Mixed-Mode Specific Surveys, Design and Analysis*, Chapter 3 CRC Press

Schnell, R., Redlich, S., & Göritz, A. S. (2022). Conditional Pop-up Reminders Reduce Incidence of Rounding in Web Surveys. Field Methods, 34(4), 334-345. https://doi.org/10.1177/1525822X221115829

Schork J.; Riillo C.A.F.; Neumayr J. (2021). "Survey Mode Effects on Objective and Subjective Questions: Evidence from the Labour Force Survey." Journal of Official Statistics, 37(1). https:/doi.org/10.2478/jos-2021-0009.

Soszynski, M. and Bliss, R. (2023). "Demographic and Measurement Differences between Text-to-Web and Phone Survey Respondents." Survey Practice 16 (1). https://doi.org/10.29115/SP-2023-0012.

Stefkovics, Á. (2022). Are Scale Direction Effects the Same in Different Survey Modes? Comparison of a Face-to-Face, a Telephone, and an Online Survey Experiment. Field Methods, 34(3), 206-222. https://doi.org/10.1177/1525822X221105940

Sun, H., Caporaso, A., Cantor, D., Davis, T., & Blake, K. (2023). The Effects of Prompt Interventions on Web Survey Response Rate and Data Quality Measures. Field Methods, 35(2), 100-116. https://doi.org/10.1177/1525822X211072358

Toepoel, V. (2016). Programming the survey. In Programming the Survey (pp. 136-159). SAGE Publications Ltd, https://doi.org/10.4135/9781473967243

Tourangeau, R., Couper, M. & Conrad. (2004). Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. Public Opinion Quarterly 68, 368-393 http://dx.doi.org/10.1093/poq/nfh035

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

Valet P, Adriaans J, Liebig S. (2019) Comparing survey data and administrative records on gross earnings: nonreporting, misreporting, interviewer presence and earnings inequality. Quality & Quantity, 53(1):471-491, DOI 10.1007/s11135-018-0764-z

Vehovar, V., Couper, M. P., & Čehovin, G. (2023). Alternative Layouts for Grid Questions in PC and Mobile Web Surveys: An Experimental Evaluation Using Response Quality Indicators and Survey Estimates. Social Science Computer Review, 41(6), 2122-2144. https://doi.org/10.1177/08944393221132644

Verbree A.-R.; Toepoel V.; Perada D. (2020). The Effect of Seriousness and Device Use on Data Quality. Social Science Computer Review, 38(6) https:/doi.org/10.1177/0894439319841027

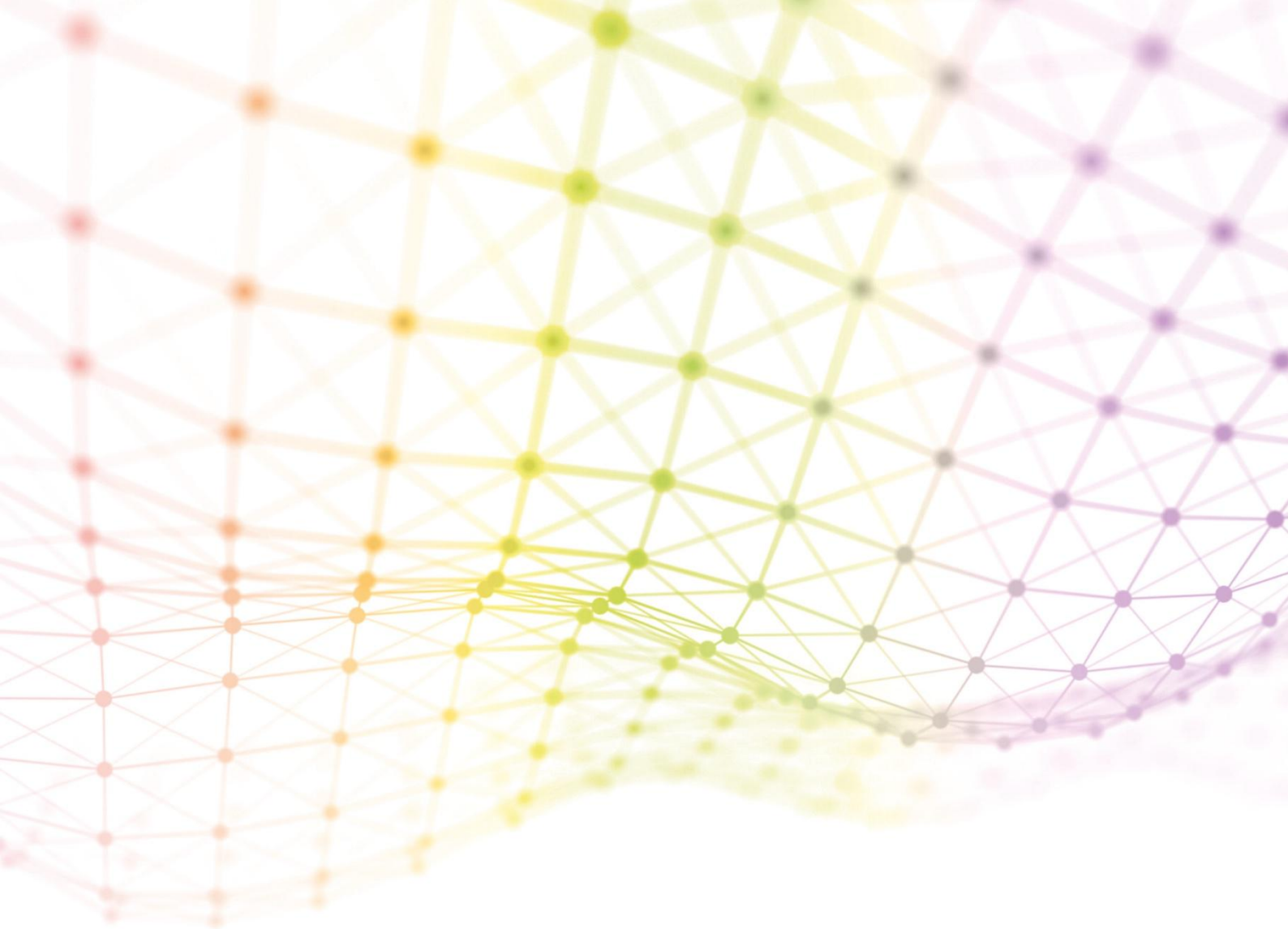
Wenz, A., Keusch, F., & Bach, R. L. (2024). Measuring Smartphone Use: Survey Versus Digital Behavioral Data. Social Science Computer Review, 0(0). https://doi.org/10.1177/08944393231224540

Willis, Gordon & Lessler, Judith. (1999). "Question Appraisal System QAS-99" Available at: https://www.researchgate.net/publication/267938670_Question_Appraisal_System_QAS-99_By

Wilson, L & Dickinson, E (2021). Respondent Centred Surveys: Stop, Listen and then Design. Sage Publishing.

# Appendix A: MERF checklist

| Category of risk | Risk codes | | Mitigation |
|---|---|---|---|
| **A. Interviewer effects** | A1. Socially desirable responses | ☐ | |
| | A2. Sensitive question | ☐ | |
| | A3: Rating scale | ☐ | |
| | A4: Knowledge or skills test | ☐ | |
| **B. Respondent satisficing** | B1: Long or complex question | ☐ | |
| | B2: Looped question | ☐ | |
| | B3: Calculation required | ☐ | |
| | B4: Open question | ☐ | |
| **C. Presentation effects** | C1: Long lists | ☐ | |
| | C2: Check all that apply | ☐ | |
| | C3: Interviewer observation/hidden code | ☐ | |
| | C4: Interviewer check | ☐ | |
| | C5: Visual aid | ☐ | |
| | C6: Battery of questions/ Grids | ☐ | |



University of Essex

University of Southampton

UKRI Economic and Social Research Council

NCRM National Centre for Research Methods

Office for National Statistics

UCL

Warwick The University of Warwick

Manchester 1824 The University of Manchester

City University of London Est 1894

National Centre for Social Research

LSE The London School of Economics and Political Science

Ulster University

Unil UNIL | Université de Lausanne

Ipsos

verian Formerly Kantar Public

www.surveyfutures.net