SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Working Paper 12:

Mode effects on survey item measurement: A systematic review of the experimental evidence

Georgia D. Tomova, Richard J. Silverwood, Liam Wright

Centre for Longitudinal Studies, University College London

January 2026

Survey Futures is an Economic and Social Research Council (ESRC)-funded initiative (grant 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.

This paper is a product of the project "Assessing and disseminating methods for handling mode effects", led by Liam Wright (UCL), which forms part of Survey Futures Research Strand 6, "Reducing and evaluating mode effects".

Prior to citing this paper, please check whether a final version has been published in a journal. If so, please cite that version. In the meanwhile, the suggested form of citation for this working paper is:

Tomova GD, Silverwood RJ & Wright L (2025) 'Mode effects on survey item measurement: A systematic review of the experimental evidence, *Survey Futures Working Paper* no. 12. Colchester, UK: University of Essex. Available at https://surveyfutures.net/working-papers/.

Table of contents

Abstract

Survey data are increasingly collected using mixed-mode designs. However, the measurement of survey items may differ across modes, introducing 'mode effects', a type of systematic measurement error which can bias analyses of mixed-mode data. While the theoretical mechanisms giving rise to mode effects have been discussed in detail, the empirical evidence on their occurrence and size is fragmented. In addition, while many existing statistical approaches for handling mode effects require unrealistic assumptions, other more suitable approaches remain underutilised due to the need for external evidence on the magnitude of mode effects. To address this, we conducted a systematic review of the experimental literature on mode effects. We searched multiple bibliographic databases, grey literature sources, and implemented backwards and forwards citation screening. Studies eligible for inclusion were (quasi-)experimental, sampled from the general population (or age-, sex-, region-specific strata), and reported mode effect estimates on item measurement. We extracted comprehensive information relating to the study design, sampling, mode effect estimates, and reporting. Ninety experimental studies published between 1967 and 2024 met the inclusion criteria, which included 4,113 mode effect estimates for 3,545 unique variables in total. Mode effects were generally small, typically below 0.2 SD. However, larger mode effects were more commonly observed when modes differed by interviewer involvement or by question delivery (visual vs aural), as well as for sensitive items (e.g., sexual behaviour, social life), which aligns with pre-existing theory on the causes of mode effects. Generally, where mode effects occur, they are item-, mode-, and population-specific. Reporting quality varied substantially and insufficient details regarding randomisation compliance, non-response, and uncertainty of estimates were common. We collated all mode effect estimates into a free online database and provide a set of recommendations to improve the reporting of future studies.

Background

Survey data are increasingly collected using mixed-mode designs (Brown and Calderwood 2020; DeLeeuw 2018). This is driven by multiple factors, including declining response rates, decreasing coverage of established modes, the introduction of new technologies for data collection, and increasing interviewer costs. Although mixing modes can lessen these problems and make surveys more adaptable to a changing environment, it also creates challenges. In particular, the measurement of survey items can differ across modes. These differences are commonly referred to as 'mode effects' (Leeuw et al. 2008) (or 'mode measurement effects' (Klausch et al. 2013)), and are a form of systematic measurement error (Leeuw et al. 2008) that can bias analyses of mixed-mode survey data. Mode effects are distinct from mode selection, which is another reason for observing differences in responses between modes. Mode selection refers specifically to differences in *who* responds by each mode (rather than *how* they respond) (Burton and Jäckle 2020; Vannieuwenhuyze and Loosveldt 2013). Often, both mode effects and mode selection are present.

The proposed mechanisms contributing to mode effects involve a combination of factors related to the psychology and motivations of the respondent, the presence and role of the interviewer, the presentation of items in a mode, and the social and physical context in which the survey is completed. There is substantial evidence that social desirability (the tendency to respond in a way that makes one appear favourable) affects responses to socially sensitive items, particularly when modes differ based on whether an interviewer is present or not, or on their perceived physical distance (Berzelak and Vehovar 2024; Kreuter et al. 2008; Roberts et al. 2006; Tourangeau and Yan 2007). Generally, questions deemed private or otherwise sensitive by the respondent (e.g., related to illegal activity) may be less likely to be truthfully reported in interviewer-led settings (Tourangeau and Smith 1996). The presence of an

interviewer may also increase acquiescence, the tendency to provide agreeable responses (Liu et al. 2017).

Mode effects may also be introduced by the presentation of items. For example, the ordering of responses might nudge respondents towards choosing a particular answer. Respondents may be prompted to select an answer appearing at the start of a list when presented visually, or the end when presented aurally (termed primacy and recency effects, respectively (Krosnick and Alwin 1987)). Perceived response burden may result in so-called 'satisficing' behaviour (Krosnick 1991), in which respondents do not provide optimal or considered answers, for example by selecting a response before reading the question in full. The main risks for satisficing include questions that are long or complex, with open-ended answers, or otherwise perceived as requiring considerable effort. Satisficing behaviour may be higher in self-administered surveys as interviewers might attempt to prevent this by engaging with and motivating the respondent. Repeated response options (e.g., as in battery measures using the same Likert scales) can also induce low effort responding (e.g., 'straightlining' (Kim et al. 2019)), especially if an interviewer is not present.

Although the potential causes and consequences of mode effects have been extensively discussed, this is not sufficient to predict the size of mode effects that may occur in practice. Given the extent of bias in analyses of mixed-mode data is related to the strength of mode effects, empirical evidence on the frequency and size of mode effects is necessary. Many studies have sought to quantify mode effects in both observational and experimental settings. For example, mode effects have been examined in general-purpose longitudinal studies that have implemented mixed-mode data collection, e.g. the European Social Survey (Jäckle et al. 2010; Roberts et al. 2020), the UK Household Longitudinal Study (Al Baghal 2019; Nandi and Platt 2017), and the National Child Development Study (Goodman et al. 2022), and in studies that have focused on a specific domain, e.g., cognition (Domingue et al. 2023), sexual

identity (Dahlhamer et al. 2019), drug use (Miech et al. 2021), or alcohol consumption (ZuWallack et al. 2023). Some studies found evidence of sizeable mode effects, while others found that they were negligible. This suggests mode effects likely depend on the specific mode comparison and the specific item (e.g., is it sensitive), rather than representing a consistent systematic difference across all survey items. However, many studies examining mode effects are conducted in observational data, which has two key limitations: 1) in mixed-mode surveys where mode is not randomly allocated, any observed differences would be at least partly attributable to mode selection, and 2) differences arising from mode comparisons between sweeps or surveys may be attributable to population differences or changes in variables over time. Therefore, ideally, empirical evidence on the size of mode effects should come from experimental studies. Although selection issues are still possible due to differential non-response or non-compliance, the consequences are likely to be less severe than those arising from observational studies.

Based on the existing literature, a recent framework by d'Ardenne *et al.* (2025) includes a set of recommendations for reducing the risk of mode effects, for example by randomising scale directions and multiple-choice answer options, simplifying the language and granularity of questions, and, when conducting face-to-face interviews, asking sensitive questions in a self-completion element. However, the authors noted that some types of mode effects, interviewer effects in particular, may not be fully preventable. Methods for reducing the implications of mode effects *post hoc*, i.e., in the data analysis stage, are therefore important (Kolenikov and Kennedy 2014; Maslovskaya et al. 2020; Wright et al. 2024). A common and straightforward method is to 'control' for mode by including a mode indicator (or dummy) variable as a model covariate. Although the intention behind this is to remove the influence of mode and therefore *reduce* bias, in the presence of mode selection this practice can *introduce* a type of bias known as 'collider bias' (Cole et al. 2010). The reasons and conditions under which this

occurs have been described elsewhere (Tomova et al. 2025). Similarly, such bias may arise with alternative methods that also require no uncontrolled mode selection (e.g. multiple imputation).

An alternative approach is quantitative bias analysis (QBA), which includes a suite of methods for obtaining bias-adjusted estimates or determining whether bias is likely to be material in practice (Fox et al. 2021). This can take the form of a counterfactual simulation in which a single-mode dataset is simulated using information from real mixed-mode data and an assumed size of mode effect, with substantive analyses performed in the simulated dataset. Alternatively, QBA can be used to quantify the size of a mode effect required to explain an observed association (so called 'simple sensitivity analysis' (VanderWeele and Li 2019)). This approach can also be applied to summary statistics (e.g., regression coefficients), allowing researchers to evaluate the potential extent of bias in existing studies. However, to provide informative results, all QBA approaches require accurate information on the size of mode effects in practice, ideally sourced from experimental studies.

There is a need for empirical assessments of mode effect to inform both the design of future surveys and the appropriate analysis of existing mixed-mode survey data, using evidence from different mode comparisons, populations, and survey items. Although many experiments have been conducted, they are scattered across the literature, which makes it more difficult to find sufficiently relevant studies or utilise multiple estimates at the same time. Previous evidence synthesis studies typically only focus on a single reference mode, e.g., telephone (Ye et al. 2011), a specific cause of mode effects, e.g., social desirability (Richman et al. 1999), a specific variable, e.g., loneliness (Stegen et al. 2024), or an outcome other than item measurement, e.g., response rate (Edwards and Perkins 2024). The vast experimental literature on mode effects has not yet been systematically reviewed and synthesised in full. In this study, we therefore sought to:

1) systematically review and synthesise the literature on mode effects estimated from experimental or quasi-experimental studies conducted in the general population or age-, sex-, and/or region-specific strata of the population;

2) use the findings to produce a freely accessible and searchable database of mode effect estimates that can be used for the purpose of informing future survey design and analysis.

Methods

Study design

This systematic review sought to identify mode effect estimates for survey items across the health and social sciences, without restrictions on the specific variables of interest. The review was limited to studies with experimental (mode assigned randomly) or quasi-experimental (mode assigned 'as random') designs to minimise the influence of mode selection. It focussed on studies conducted in the general population, or age-, sex-, and/or geographical region-specific strata of the population, to provide as useful and generalisable results as possible, whilst being feasible to perform. Before the systematic review commenced, a protocol detailing the study design was developed, which informed a pilot search, screen, and extraction process. The protocol was updated using information gained from the pilot stages and was pre-registered on the Open Science Framework (Tomova et al. 2025b), adhering to guidance on the pre-registration of systematic reviews (Van Den Akker et al. 2020). Any deviations from the original protocol are reported and justified in **Supplementary Table 1**.

Search and screening strategy

The search was conducted in three stages to maximise the likelihood of identifying all relevant literature.

First, we searched the bibliographic databases Scopus, MEDLINE, Health and Psychosocial Instruments, APA PsycInfo, APA PsycExtra, and Web of Science Core Collection. The search was conducted on 10 March 2025 using search queries designed to locate relevant studies based on the inclusion criteria (**Table 1**). The searches were designed to be as similar as possible across the different databases within the constraints of their syntax. Results were restricted to English. The exact search queries used in each database are available in **Supplementary Table 2**.

Table 1. Inclusion and exclusion criteria used for determining inclusion of a study in the systematic review during the screening process.

Inclusion criteria	Exclusion criteria
▪ Seeks to estimate and report mode effect estimates on survey item measurement ▪ Uses an experimental (where the exposure to a survey mode is randomly assigned) or quasi-experimental (e.g. where the exposure to a survey mode is assigned as-if randomised [though not explicitly so], or where a change in data collection practices may have occurred such that more than one mode is used to collect data from otherwise similar groups of people) design ▪ Within the domain of health and social science survey design and methodology ▪ Longitudinal or cross-sectional study sampled from the general population or a sex-, age- and/or region-specific stratum of the population ▪ Published in English language ▪ Published any time since database inception	▪ No mode effect estimate is reported ▪ Mode effect estimates only reported for response rates ▪ Estimates a mode effect for an association rather than item measurement ▪ Population is defined by clinical or occupational characteristics (e.g. teachers, people with diabetes, psychology students) or other characteristics not limited to age, sex, and/or geographical region ▪ No full text available (e.g. conference abstract or abstract where a full text version cannot be identified)

Second, we searched Google Scholar to identify grey literature which may not have been indexed in the standard bibliographic databases, and which may instead be available on university or pre-print repositories. Due to limited search functionality, the Google Scholar search query was a simplified version of the other searches (see Supplementary Table 2). Based on recommendations, we searched the first 1,000 results of Google Scholar to locate potentially relevant grey literature (Haddaway et al. 2015).

Third, we implemented both a backwards and forwards citation screen (i.e. screened all references listed within each article and all current citations of the article) of all articles that were identified as relevant following full text screening after the first two stages, to identify further potentially relevant papers. We used Google Scholar to identify the citations of each article. At the citation screen stage, the reference lists of systematic reviews and meta-analyses located in the previous stages were also screened to identify other potentially relevant articles. The backwards and forwards citation screen was completed between 7-10 April 2025 and therefore included citations which had appeared on Google Scholar by then. Search results from all databases were de-duplicated using both Zotero (Corporation for Digital Scholarship 2025) and Rayyan (Ouzzani et al. 2016) to maximise the de-duplication success.

All studies identified in the first two search stages were screened for inclusion based on adherence to the inclusion and exclusion criteria (Table 1). First, the titles and, where relevant, abstracts were screened, and if deemed potentially relevant, the studies moved to the full-text screen stage. If articles were excluded at the full-text screening stage, the decisions for this were recorded. All studies identified in the backwards and forwards citation search as well as those found in systematic reviews and meta-analyses had their titles screened, after which potentially eligible studies were de-duplicated with those already included, and the full texts of any remaining studies were then screened. A single reviewer (GDT) screened all

studies, while a second reviewer (LW) screened a random sub-sample of 10% of studies for validation purposes.

Data extraction

A data extraction form was developed a priori to collect all necessary information from each included study (see **Table 2** for a summary of the extraction items, and **Supplementary File 1** for a copy of the data extraction form). The data extraction form was designed to capture information on the mode effect estimates and their associated uncertainty (e.g., standard errors), the modes compared, survey items and populations they were estimated for, as well as general information on the study design, sampling, and quality of reporting. The intention behind this was to provide enough details so that researchers can make an informed decision as to whether a mode effect estimate is reliable and useful for their own analyses, for example when assessing the existing evidence of mode effects in a particular context or when performing QBA. Since the focus was on item-level information, no implicit respondent behaviour measures (e.g. straightlining) were extracted. Where values for certain extraction items were not directly provided, they were manually derived where possible. Commonly, extraction items that required manual calculation included variable standard deviations, standardised mode effect sizes, and mode effect standard errors and corresponding 95% confidence intervals. Standardised effect sizes were calculated using Glass's delta (Kumar et al. 2022). For a full list of extraction items that were derived, and how they were derived, see **Supplementary Table 3.** Not all items were possible to manually derive based on the available information. For example, standard deviations for continuous variables are not straightforward to calculate without any measure of spread. Data that were not reported or not possible to derive were therefore recorded as being missing. Some extraction items could be extracted in multiple ways. Where a mode effect was presented using more than one type of effect measure (e.g. both a mean difference between modes as well as an odds ratio), then

both were extracted as separate entries. Where a comparison between more than two modes was made, each pairwise comparison was extracted separately. Where age- or sex-specific mode effect estimates were reported, these were extracted in addition to the overall mode effects. Where both unadjusted and adjusted estimates were presented, the unadjusted estimates were extracted to improve comparability between the studies and avoid extracting estimates that may have been adjusted *post hoc* (due to, e.g. p-hacking). However, where weighting was applied specifically to address selection problems (e.g. to generalise estimates to the target population), then weighted results were extracted so that results better reflect the population of interest. A single reviewer (GDT) completed data extraction for all studies, and a second reviewer (LW) completed data extraction on a random sub-sample of 10% of studies for validation purposes.

Table 2. Items extracted (or derived) from each study included in the systematic review.

Extraction category	Extraction items
General information	Year of publication, authors, journal (or repository), digital object identifier (DOI) or alternative unique identifier
Study design and sampling	Source population (i.e. sampled from an existing survey population or from the general population), population profile, survey name and sweep, country, sampling strategy, experimental design, study modes, response rate, post-response exclusions, final sample size, compliance
Variables and mode effects	Variable category and sub-category selected from a priori defined list, reference and alternate mode, item response rate per mode, outcome standard deviation per mode, item measure per mode, mode effect estimand, effect measure, mode effect estimate, standard error, confidence interval, p-value, standardised effect estimate, standard error and confidence interval
Quality of reporting and general appraisal	General quality of reporting, potential challenges related to selection or item non-response, and any general comments not captured elsewhere

Validation

Several steps were taken to improve the validity, completeness, and relevance of the findings. The search, screening, and data extraction design were all tested on a smaller initial sample of 300 search hits before conducting the formal systematic review stages. This pilot phase helped to validate, refine, and ensure the relevance of the search query and the data extraction form. As noted, a random sub-sample of 10% of the search hits were double-screened (both at abstract and full-text screen stage), and a random sub-sample of 10% of the included studies were double-extracted. The double-screening and double-extracting were conducted at the start so that any insight gained from them could be incorporated into the rest of the process. The two reviewers were blinded to each other's decisions until the process was complete. Following this, any discrepancies were noted and resolved through discussion.

Results synthesis

We produced descriptive statistics indicating the types of survey items (categories and sub-categories of topics) for which mode effects have been examined, the types and number of specific mode comparisons, as well as the typical distribution of mode effects observed for different mode contrasts, different mode characteristics, and variable categories. For synthesis purposes, we classified each reported mode into one of the following broader categories: paper, web, face-to-face, face-to-face (computer-assisted and self-completed, henceforth referred to as (A)CASI), telephone, mobile (app or web browser), and other (which included hybrid, ballot box, and randomised response). Following the framework on drivers of mode effects developed by d'Ardenne *et al.* (2025), we also classified each mode according to the following characteristics: physical presence or absence of an interviewer (regardless of whether answers were reported to them); written or aural delivery of questions and response options; computer-assisted or traditional survey; whether answers were reported directly to

the interviewer or not (and if yes, whether this was in-person or over the phone; if not, whether this was on paper or web); and, where an interviewer was involved, whether they were present in-person (with or without collecting responses directly) or directly collecting responses (with or without being in-person). Further details on which modes were classified into each category are available in **Supplementary Table 4**. We calculated the proportion of estimates in each comparison that exceeded 0.2 SD (a common 'rule of thumb' classification for a 'small' effect size for standardised effects (Cohen 2009; Sullivan and Feinn 2012). We did not calculate the proportion of medium (0.5 SD) or large (0.8 SD) effects since such effect sizes are very rare in the context of mode effects. To maintain comparability, only estimates for which a standardised effect size was reported or derived were used for producing the visual plots. Since the direction of any mode effect is dependent on the way survey items are coded, we used absolute effect sizes in the plots. Finally, we collated the reported mode effect estimates and all accompanying information into a freely accessible and searchable online database (https://cls-data.github.io/mode-effects-database/).

Reporting

We adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al. 2021) and the PRISMA-S extension for the reporting of the literature search (Rethlefsen et al. 2021) (see **Supplementary Tables 5** and **6** for checklists).

Results

Search and screen

The search returned 13,003 unique records (2,808 from the bibliographic databases, 925 from Google Scholar, and 9,270 from the citation screening). After screening all titles and abstracts, 313 articles were identified as potentially relevant. After full text screening, 90

studies met all inclusion and exclusion criteria and were included in the systematic review. For full details of the process, see the PRISMA flowchart (Page et al. 2021) in **Figure 1**. A list of all included studies is available in **Table 3**.

Table 3. A list of all studies meeting the inclusion and exclusion criteria and included in the systematic review.

Study	Source population category[a]	Survey name[b]	Study modes
Li et al. 2024	Cross-sectional survey	Youth Risk Behavior Survey (YRBS)	Paper, tablet
Feng and Huang 2024	Longitudinal survey	Labour Force Survey	CAWI, CAPI
O'Muircheartaigh et al. 2025	Longitudinal survey	National Social Life, Health, and Aging Project (NSHAP)	Face-to-face, web
Schumann and Lück 2023	Longitudinal survey	Generations and Gender Survey (GGS)	CAPI, CAWI
Smith et al. 2023	Longitudinal survey	Health and Retirement Study (HRS)	Face-to-face, telephone
Otsuka et al. 2023	General population (adolescents)	N/A	Paper, web
Domingue et al. 2023	Longitudinal survey	Health and Retirement Study (HRS)	Web, telephone
Hope et al. 2022	Cross-sectional survey	NatCen Social Research Omnibus survey	CAPI, CATI, web
Reisinger 2022	Longitudinal survey	National Longitudinal Survey of Youth 1997 cohort (NSLY 97)	Telephone, paper
Ofstedal et al. 2022	Longitudinal survey	Health and Retirement Study (HRS)	Web, telephone
Piccitto et al. 2022	Longitudinal survey	Generations and Gender Survey (GGS)	Web, face-to-face
Clarke and Bao 2022	Longitudinal survey	Understanding Society	Web, face-to-face
Goodman et al. 2022	Longitudinal survey	1958 National Child Development Study (NCDS)	Web, telephone
De Vitiis et al. 2021	Cross-sectional survey	Aspects of Daily Life (ADL)	Web, face-to-face
Adalı et al. 2022	Cross-sectional survey	Turkey Demographic and Health Survey	CAPI, PAPI
Kim and Couper 2021	Cross-sectional survey	National Survey of Smoking and Health	Web (smartphone), CATI
Miech et al. 2021	Cross-sectional survey	Monitoring the Future (MTF)	Paper, tablet
Roberts et al. 2020	Cross-sectional survey	European Social Survey (ESS)	CAPI, CATI
Patrick et al. 2021	Longitudinal survey	Monitoring the Future (panel)	Paper, web
Colasante et al. 2019	Cross-sectional survey	European School Survey Project on Alcohol and other Drugs (ESPAD) The Alcohol Misuse Prevention Study & Genesse Intermediatea	Paper, computer
Sakshaug et al. 2019	Longitudinal survey	School District Study	Mail, CATI
Dahlhamer et al. 2019	Cross-sectional survey	National Health Interview Survey	CAPI, ACASI
Fischer and Bayham 2019	Longitudinal survey	UC Berkeley Egocentric Network Survey (UCNets)	Face-to-face, web
Šmigelskas et al. 2019	Cross-sectional survey	Health Behavior in School-aged Children (HBSC) in Lithuania	Paper, web
Al Baghal 2019	Longitudinal survey	Understanding Society Innovation Panel	Web, CAPI
Sanchez Tome 2018	General population (adults)	N/A	Mail, CAWI, CATI
Fishbein et al. 2018	Large-scale educational assessment survey	Trends in International Mathematics and Science Study (TIMMS)	Paper, electronic
Baier 2018	General population (adolescents)	N/A	Paper, netbook
Jerrim et al. 2018	Large-scale educational assessment survey	Programme for International Student Assessment (PISA)	Paper, computer
Helppie-Mcfall and Hsu 2017	Longitudinal survey	Cognitive Economics (Cog-Econ) Study	Web, mail
Cea D'Ancona 2017	Cross-sectional survey	MEDIM	Face-to-face, paper
Cernat et al. 2016	Longitudinal survey	Health and Retirement Study (HRS)	CAPI, CATI
Nandi and Platt 2017	Longitudinal survey	Understanding Society Innovation Panel	Face-to-face, telephone
Liu and Wang 2016	Longitudinal survey	American National Election Studies (ANES)	Face-to-face, web
Holford and Pudney 2015	Longitudinal survey	Understanding Society Innovation Panel	Face-to-face, CASI, CATI
Wells et al. 2014	Cross-sectional survey	KnowledgePanel	Mobile app, web
Cea D'Ancona 2014	Cross-sectional survey	MEXEES	Face-to-face, paper
Schouten et al. 2013	General population (adults)	N/A	Face-to-face, telephone, web, paper
de Bruijne and Wijnant 2013	Longitudinal survey	CentERpanel	Mobile, computer, hybrid
Fleming et al. 2013	Longitudinal survey	Raising Healthy Children (RHC)	Web, telephone
Christensen et al. 2014	Cross-sectional survey	Danish Health and Morbidity Survey	CAPI, paper
Anglewicz et al. 2013	General population (adults)	N/A	Face-to-face, ballot box, randomised response
Vannieuwenhuyze and Loosveldt 2013	General population (adults)	N/A	Mail, face-to-face
Vannieuwenhuyze et al. 2012	General population (adults)	N/A	Mail, face-to-face
Caeyers et al. 2012	General population (adults)	N/A	CAPI, restricted CAPI, PAPI
Le and Vu 2012	General population (adults)	N/A	ACASI, paper, face-to-face
Lugtig et al. 2011	General population (adults)	N/A	CATI, WAPI

Källmén et al. 2011	General population (adults)	N/A	Mail, web
Sinadinovic et al. 2011	General population (adults)	N/A	Web, IVR
Wettergren et al. 2011	General population (adolescents and young adults)	N/A	Telephone, mail
Jäckle et al. 2010	Cross-sectional survey	European Social Survey (ESS)	Face-to-face, telephone
Eaton et al. 2010	Cross-sectional survey	Youth Risk Behavior Survey (YRBS)	Paper, web
Rosel et al. 2010	General population (children)	N/A	Face-to-face, SAQ
Beach et al. 2010	General population (older adults)	N/A	CAPI, ACASI, CATI, IVR
Lygidakis et al. 2010	General population (adolescents)	N/A	Paper, web
Erhart et al. 2009	General population (adolescents)	N/A	Mail, CATI
Langhaug 2009	General population (adolescents)	N/Aa	Paper SAQ, Audio-SAQ, face-to-face, ACASI
Harmon et al. 2009	General population (adults)	N/A	T-IAQ, T-ACASI
McMorris et al. 2009	Longitudinal survey	Raising Healthy Children (RHC)	Web, CASI
Van De Looij-Jansen and De Wilde 2008	Longitudinal survey	Youth Health Monitor Roterdam (YMR)	Paper, web
Villarroel et al. 2008	General population (adults)	N/A	T-IAQ, T-ACASI
Midanik and Greenfield 2008	Cross-sectional survey	US National Alcohol Survey	CATI, IVR
Béland and St-Pierre 2007	Cross-sectional survey	The Canadian Community Health Survey (CCHS)	CAPI, CATI
Mensch et al. 2008	Longitudinal survey	Malawi Diffusion and Ideational Change Project (MDICP)	Face-to-face, ACASI
Lucia et al. 2007	General population (adolescents)	N/A	Paper, web
Villarroel 2006	General population (adults)	N/A	CATI, T-ACASI
Jörngården et al. 2006	General population (adolescents and young adults)	N/A	Telephone, mail
Brener et al. 2006	General population (adolescents)	N/A	Paper, CASI
Turner et al. 2005	General population (adults)	N/A	T-IAQ, T-ACASI
Wang et al. 2005	General population (adolescents)	N/A	Paper, web
Mangunkusumo et al. 2005	General population (adolescents)	N/A	Paper, web
McCabe et al. 2005	General population (children)	N/A	Paper, web
Hewett et al. 2004	General population (adolescents)	N/A	Face-to-face, ACASI
Moskowitz 2004	General population (adolescents)	N/A	CATI, T-ACASI
Currivan 2004	Longitudinal survey	UMass Tobacco Study	CATI, T-ACASI
Chromy et al. 2002	Cross-sectional survey	National Household Survey of Drug Abuse (NHSDA)	Paper, CAI
Supple et al. 1999	General population (adolescents and young adults)	N/A	Paper, CASI
Bongers and Van Oers 1998	General population (adults)	N/A	Mail, face-to-face
Perkins and Sanson-Fisher 1998	General population (adults)	N/A	Mail, CATI
Rogers et al. 1998	Cross-sectional survey	National Household Survey of Drug Abuse (NHSDA)	Paper, face-to-face
Aquilino 1998	General population (adults)	N/A	Paper, face-to-face, telephone
Turner et al. 1998	General population (adolescents)	N/A	Paper, ACASI
Wright et al. 1998	General population (adolescents and young adults)	N/A	Paper, CASI
Aquilino 1997	General population (adults)	N/A	Paper, face-to-face, telephone
Tourangeau and Smith 1996	General population (adults)	N/A	CAPI, CASI, ACASI
McHorney et al. 1994	Cross-sectional survey	National Survey of Functional Health Status (NSFHS)	Mail, CATI
Aquilino 1994	General population (adults)	N/A	Paper, face-to-face, telephone
Dillman and Tarnai 2004	General population (adults)	N/A	Mail, telephone
Aneshensel et al. 1982	General population (adults)	N/A	Telephone, face-to-face
Hochstim 1967	General population (adults)	N/A	Mail, telephone, face-to-face

[a] To distinguish between standalone experiments and those embedded within existing surveys; [b]If embedded within an existing survey.

Studies

The included studies were published between 1967 and 2024, with most publications (n = 83, 92%) occurring from 1998 onwards. No quasi-experimental studies were identified that also met all other inclusion and exclusion criteria. Therefore, all estimates come from studies with random allocation of mode, though quasi-experimental methods (but not allocation), such as instrumental variable analyses, were employed in some. Most studies examined USA populations (n = 41, 46%), followed by the Netherlands (n = 6, 7%) and the UK (n = 6, 7%). Only 12 out of 90 studies examined populations outside of Europe and North America. Most experiments were conducted as part of existing surveys (n = 46, 51%), while the rest were standalone experiments conducted in the general population of adults (n = 23, 26%), adolescents (n = 12, 13%), adolescents and young adults (n = 4, 4%), children (n = 2, 2%), or older adults (n = 1, 1%), and two (2%) experiments were conducted as part of large-scale educational assessments. Among the studies reporting experiments embedded within existing surveys, almost all surveys were examined just once, though some surveys were explored by several studies, namely the Health and Retirement Survey (HRS) (n = 4, 4%) and Understanding Society Innovation Panel (n = 3, 3%). Almost all studies were published in academic journals (n = 84, 93%), but some were published in books (n = 3, 3%) and institutional or other online repositories (n = 3, 3%). See **Table 4** for further details.

Figure 1. PRISMA flow diagram depicting the identification and screening process of the systematic review. Adapted from Page et al. (2021).

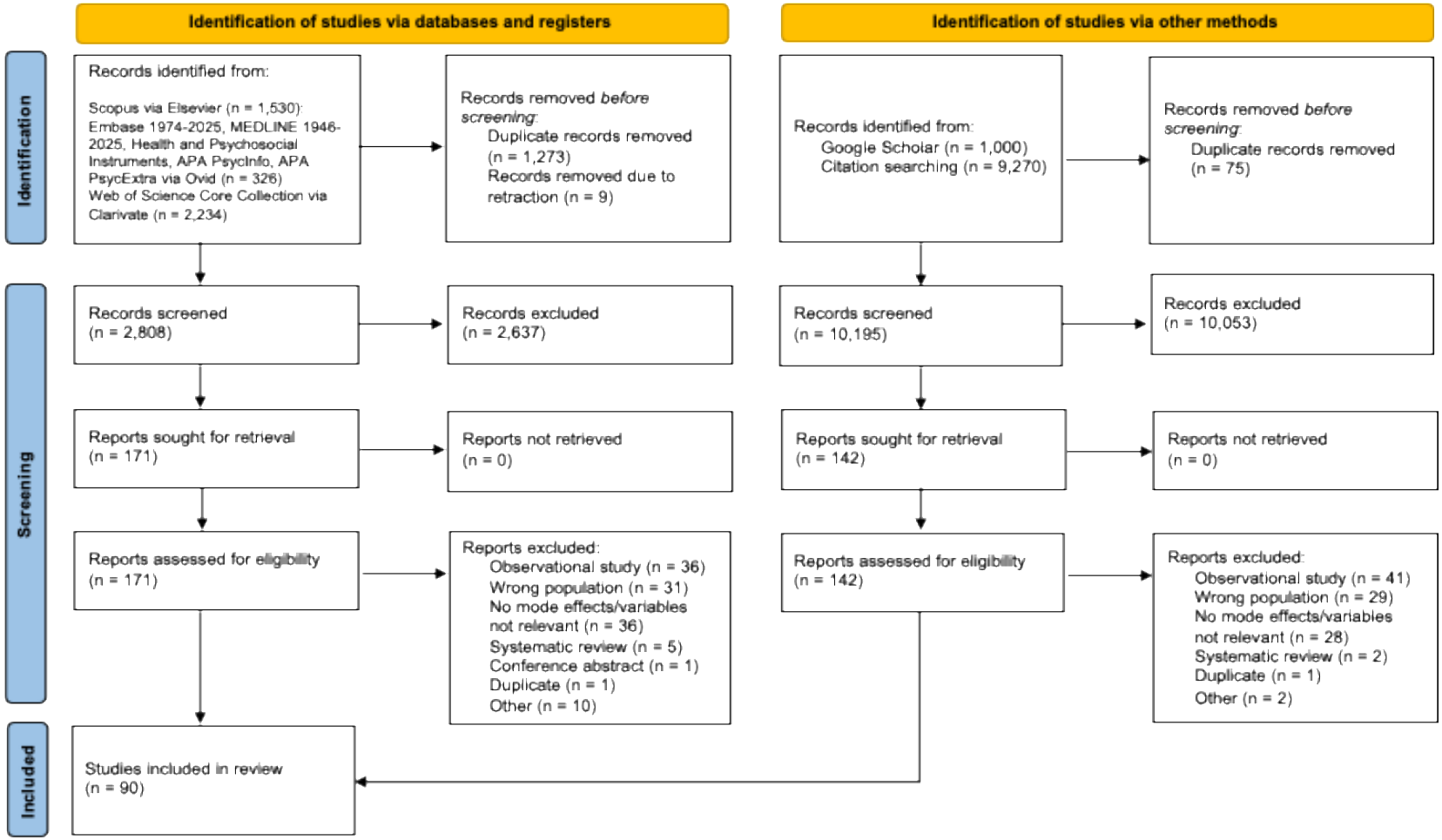


Table 4. The number and proportion of studies according to year of publication, population type, population country, and survey affiliation.

Year of publication	n (%)	Population source	n (%)
<1995	5 (6%)	Survey members	46 (51%)
1996-1999	9 (10%)	Longitudinal survey	25 (28%)
2000-2004	4 (4%)	Cross-sectional survey	21 (23%)
2005-2009	17 (19%)	General population	42 (47%)
2010-2014	20 (22%)	adults	23 (26%)
2015-2019	16 (18%)	adolescents	12 (13%)
2020-2025	19 (21%)	adolescents and young adults	4 (4%)
		children	2 (2%)
		older adults	1 (1%)
		Large-scale educational assessment survey	2 (2%)
Total	**90 (100%)**		**90 (100%)**

Population country	n (%)	Survey	n (%)
USA	41 (46%)	Health and Retirement Study (HRS)	4 (4%)
UK	6 (7%)	Understanding Society Innovation Panel	3 (3%)
the Netherlands	6 (7%)	European Social Survey (ESS)	2 (2%)
Sweden	4 (4%)	Generations and Gender Survey (GSS)	2 (2%)
Germany	3 (3%)	National Household Survey of Drug Abuse (NHSDA)	2 (2%)
Italy	3 (3%)	Raising Healthy Children (RHC)	2 (2%)
Spain	3 (3%)	Youth Risk Behaviour Survey (YRBS)	2 (2%)
Switzerland	3 (3%)	Monitoring the Future (MTF)	2 (2%)
Belgium	2 (2%)	1958 National Child Development Study (NCDS)	1 (1%)
Australia	1 (1%)	American National Election Studies (ANES)	1 (1%)
Botswana	1 (1%)	Aspects of Daily Life (ADL)	1 (1%)
Canada	1 (1%)	CentERpanel	1 (1%)
China	1 (1%)	Cognitive Economics (Cog-Econ) Study	1 (1%)
Denmark	1 (1%)	Danish Health and Morbidity Survey	1 (1%)
Hungary	1 (1%)	European School Survey Project on Alcohol and other Drugs (ESPAD)	1 (1%)
Japan	1 (1%)	Health Behavior in School-aged Children (HBSC)	1 (1%)
Kenya	1 (1%)	KnowledgePanel	1 (1%)
Lithuania	1 (1%)	Labour Force Survey	1 (1%)
Malawi	1 (1%)	Malawi Diffusion and Ideational Change Project (MDICP)	1 (1%)
South Korea	1 (1%)	MEDIM	1 (1%)
Taiwan	1 (1%)	MEXEES	1 (1%)
Tanzania	1 (1%)	NatCen Social Research Omnibus Survey	1 (1%)
Turkey	1 (1%)	National Health Interview Survey	1 (1%)
Vietnam	1 (1%)	National Longitudinal Survey of Youth 1997 cohort (NSLY 97)	1 (1%)
Zimbabwe	1 (1%)	National Social Life, Health, and Aging Project (NSHAP)	1 (1%)
International	3 (3%)	National Survey of Functional Health Status (NSFHS)	1 (1%)
		National Survey of Smoking and Health (NSSH)	1 (1%)
		Programme for International Student Assessment (PISA)	1 (1%)
		The Alcohol Misuse Prevention Study & Genesse Intermediate School District Study	1 (1%)
		The Canadian Community Health Survey (CCHS)	1 (1%)
		Trends in International Mathematics and Science Study (TIMMS)	1 (1%)
		Turkey Demographic and Health Survey	1 (1%)
		UC Berkeley Egocentric Network Survey (UCNets)	1 (1%)
		UMass Tobacco Study	1 (1%)
		Understanding Society	1 (1%)
		US National Alcohol Survey	1 (1%)
		Youth Health Monitor Rotterdam (YMR)	1 (1%)
		N/A (not affiliated with a survey)	42 (47%)
Total	**90 (100%)**		**90 (100%)**

Mode comparisons

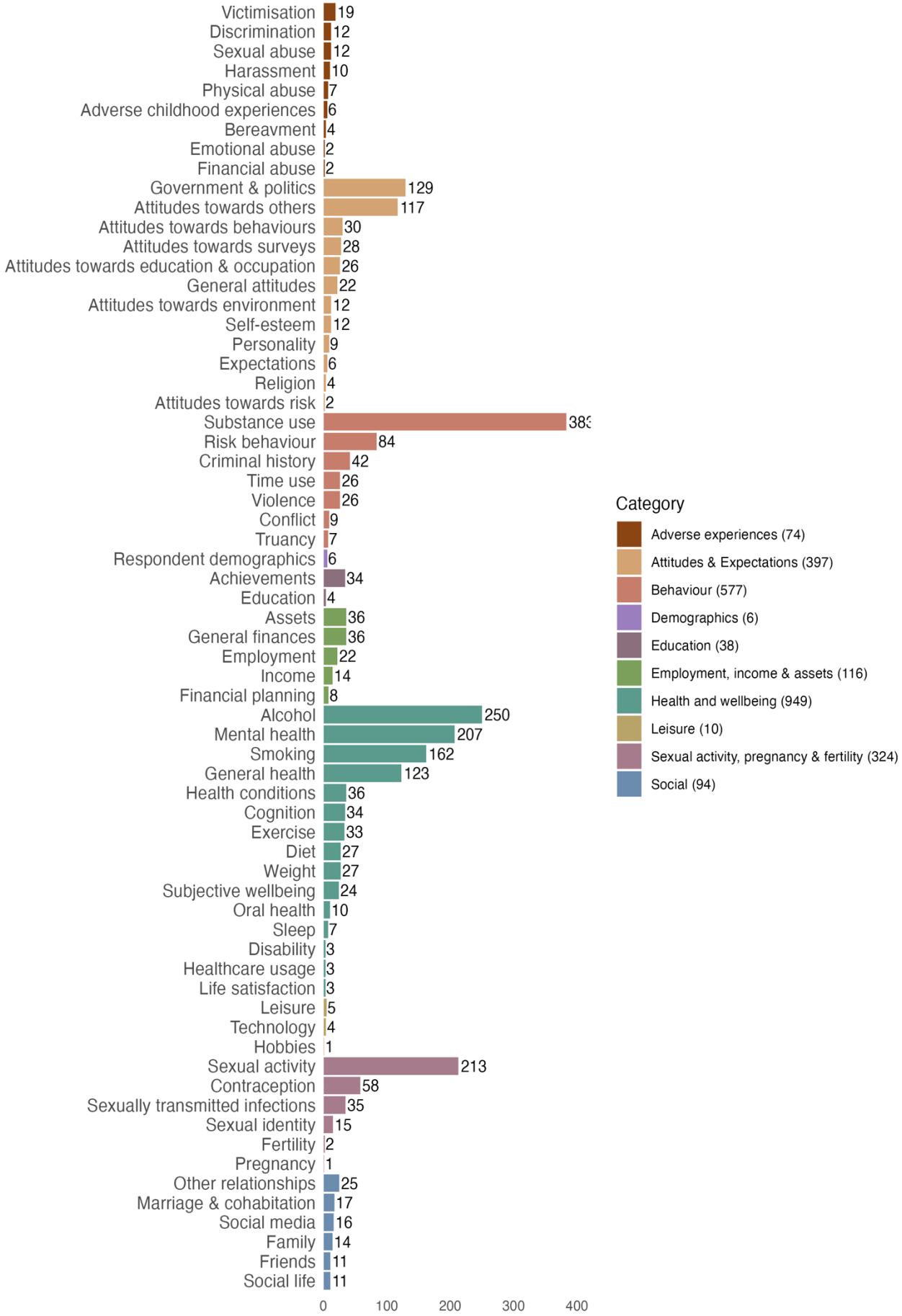
There were 128 mode comparisons in total (corresponding to 50 unique mode comparisons) examined across the 90 studies (min=1, median=1, max=9 mode comparisons per study). The most common comparisons were between paper and web modes (n = 16, 13%) and between

face-to-face and paper modes (n = 16, 13%), followed by face-to-face and telephone (n = 14, 11%), paper and telephone (n = 13, 10%), and face-to-face and web (n = 9, 22%). It was not uncommon for similar types of modes to be compared to each other, for example comparing telephone to telephone audio computer-assisted self-interview (T-ACASI) (n = 3, 2%), computer-assisted telephone interview (CATI) to T-ACASI (n = 2, 2%), or CATI to interactive voice response (IVR) (n = 2, 2%). Several comparisons (n = 6, 5%) included contrasting standard modes to non-standard ones such as randomised response, ballot box response, or hybrid modes (e.g., those including elements of both mobile and web). For a complete list of mode comparisons see **Supplementary Table 7.**

Mode effects

In total, 4,113 mode effect estimates were identified across the 90 studies. Of these, 3,545 represented unique items, with the remainder being alternative estimates for the same item in the same study, e.g. reporting both an odds ratio and a mean difference, or both an intention-to-treat (ITT) and a complier average causal effect (CACE). Variables were most commonly classified under the category of health and wellbeing (n = 949 unique items, 26.8% of total unique items), followed by behaviour (n = 577, 16.3%), attitudes & expectations (n = 397, 11.2%), and sexual activity, pregnancy & fertility (n = 324, 9.1%). The breakdown and distribution across all topic sub-categories is available in **Figure 2**. Where a single type of effect measure was used, most mode effects were almost exclusively reported as mean differences (n = 3,166, 77.0% of total mode effect estimates), though some were reported as odds ratios (n = 350, 8.5%), risk differences (n = 42, 1.0%), prevalence ratios (n = 34, 0.8%), or median differences (n = 12, 0.3%). 509 items (14.3% of unique items) had estimates reported using more than one effect measure. This was always in addition to a mean difference and was either an odds ratio (n = 499, 98.0% of those with more than 1 effect measure) or log odds (n = 10, 2.0%). Only 55 (1.3%) items had estimates reported for two

Figure 2. Number of items within each topic category examined across all studies in the systematic review.



comparisons for unique items (i.e. excluding those reported using multiple effect measures or estimands), 2,859 (80.6%) had standardised effect sizes that were either reported or successfully derived.

Figure 3 shows distribution of absolute standardised mode effects across the eight most common mode comparisons, and the proportion of estimates in each that exceed 0.2 SD. For some comparisons, mode effects were found to be relatively small overall, for example between paper and web (only 9.8% exceeded 0.2 SD), between face-to-face and telephone (8.7%), and between paper and (A)CASI (7.7%). However, larger mode effects were observed for other types of comparisons, such as between telephone and paper modes (40.4% exceeded the threshold), between standard face-to-face and (A)CASI (39.4%), and between face-to-face and web (37.3%). **Figure 4** shows mode effect distributions stratified by variable categories for the eight most common mode comparisons. Where mode effects were observed in the overall comparisons (Figure 3), they also tended to be observed across most or all categories when stratified. However, for some mode comparisons that had smaller mode effects on average, certain variable categories did exhibit more substantial mode effects. For example, in the face-to-face and telephone comparison, 20.5% of behavioural variables exceeded the threshold, whereas under 10% of variables in the other categories did. In the paper and (A)CASI comparison, health and wellbeing variables were subject to larger mode effects (17.1% exceeding threshold) more commonly than others (under 5%). In the paper and web comparison, mode effects were observed for education variables (35.5%) and health and wellbeing variables (13.5%) but were rare in the other categories (under 10%). Not all variable categories were available for all mode comparisons.

Figure 3. Mode effect distributions for the eight most common categories of mode comparisons. All estimates are standardised and represent absolute values. Dashed lines represent a threshold of 0.2 SD, indicating the presence of (at least) a "small" effect size. Only estimates for which standardised effect sizes were available (80.6% of all estimates) are included.

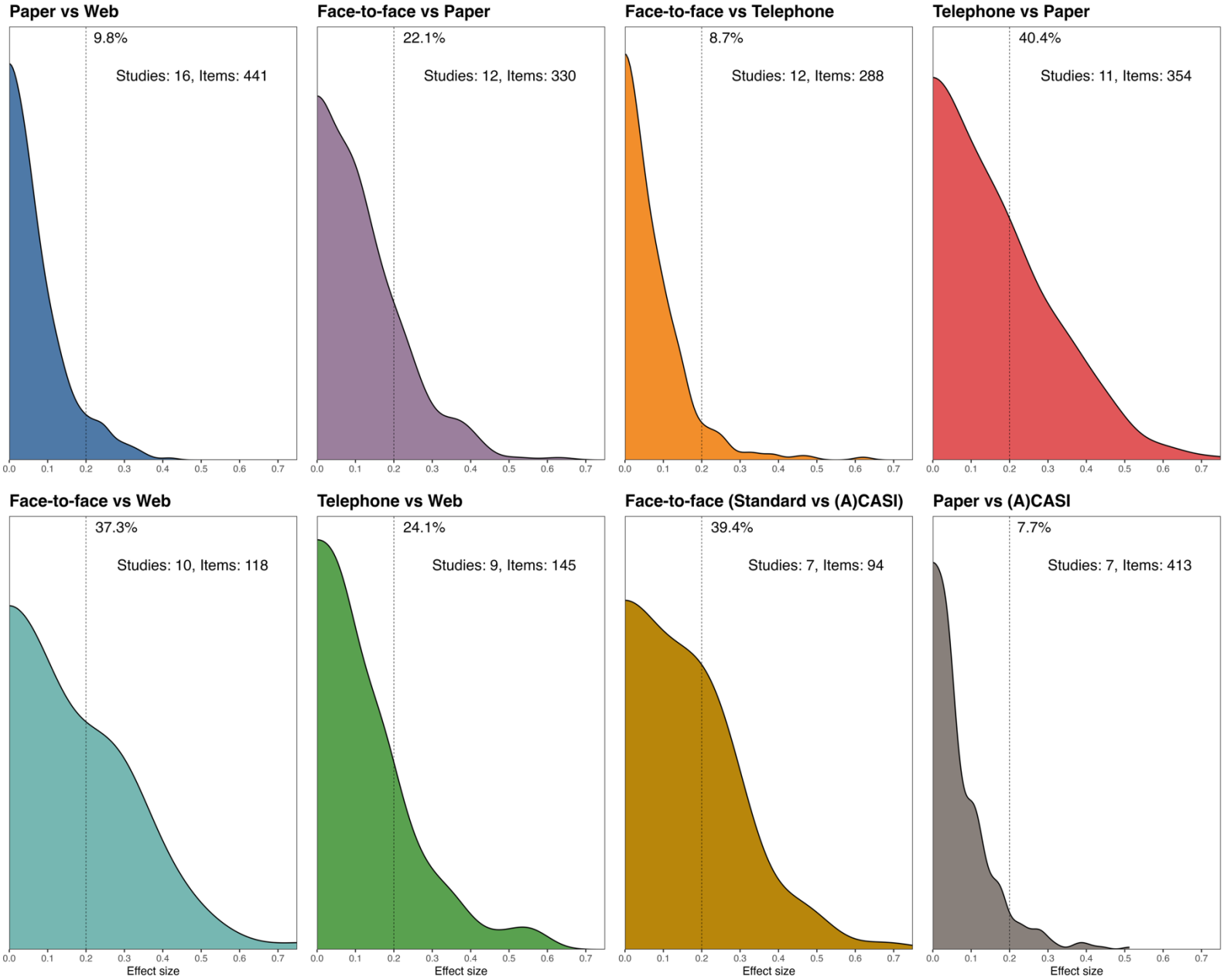


Figure 4. Mode effect distributions for the eight most common categories of mode comparisons, broken down by item category. All estimates are standardised and represent absolute values. Dashed lines represent a threshold of 0.2 SD, indicating the presence of (at least) a "small" effect size. Only estimates for which standardised effect sizes were available (80.6% of all estimates) are included.

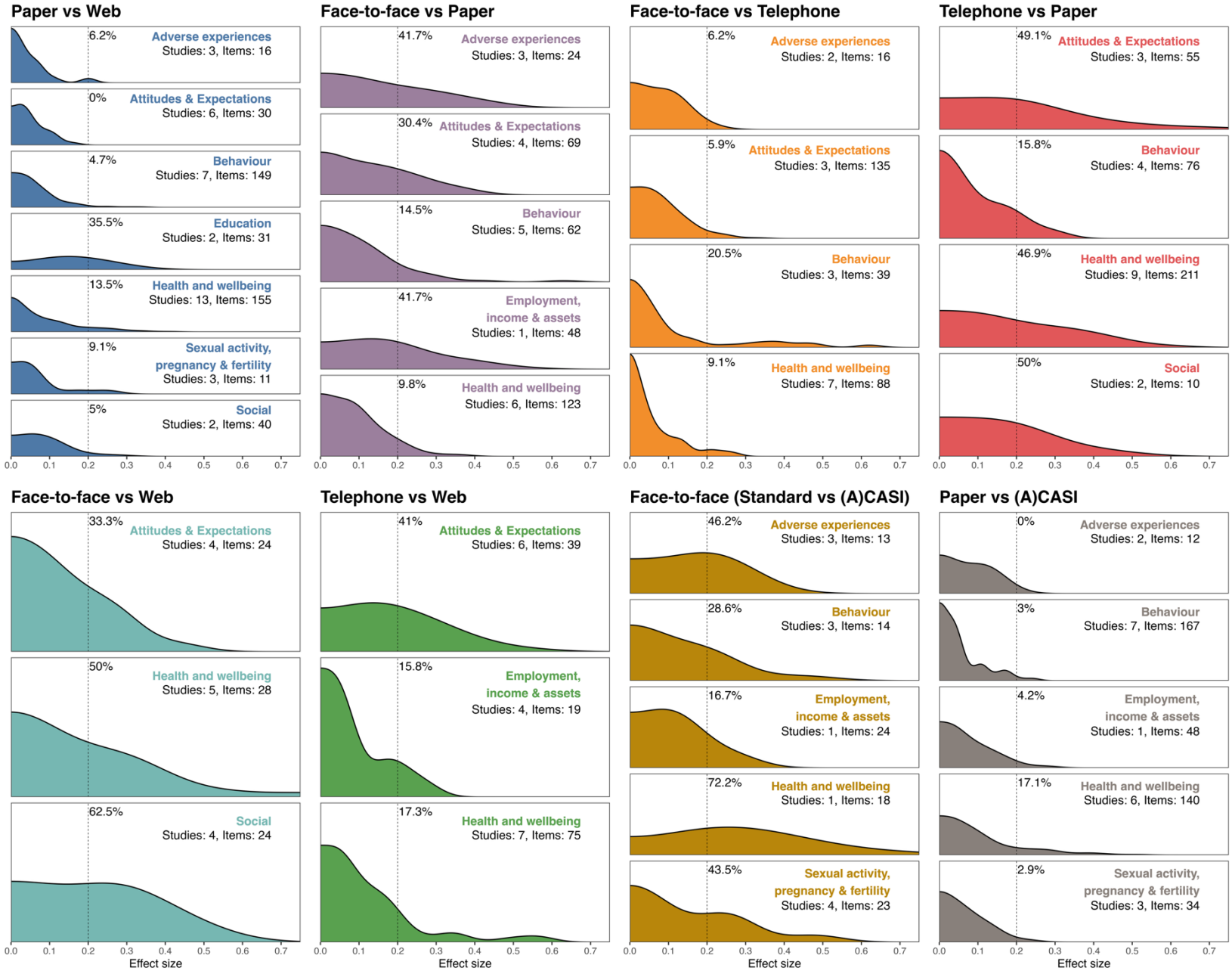


Figure 5 shows the distributions of absolute standardised mode effects across six out of the seven examined groups classified according to specific characteristics of the mode design, and **Figure 6** shows the distributions stratified by item category. One of the groups (among those who responded directly to an interviewer, whether they did so over the phone or in-person) was excluded from analysis due to having too few items (n = 262) to be presented reliably. Figure 5 suggests that larger mode effects were most common when comparing modes that differ based on whether the answers were reported directly to an interviewer or not (28.7% of items exceeded 0.2 SD), whether the questions were delivered visually or aurally (27.4%), or, when an interviewer was involved in any way, whether they were simply present or the answers were directly reported to them (22.7%). Mode effects were typically small when comparing paper or electronic self-completion surveys (9.5%). The proportion of items exceeding 0.2 SD was 17.9% when comparing modes that differed by presence of an interviewer and 13.9% when comparing computer-assisted and computer-unassisted modes. The results were more nuanced when comparisons were made by item category. Some mode effects were particularly small in certain categories, e.g., 0% of the attitudinal variable estimates and only 2% of the adverse experience estimates exceeded 0.2 SD when comparing self-reported paper or electronic modes. Items relating to sexual activity, pregnancy & fertility generally exhibited larger mode effects more frequently, with more than 40% of the estimates exceeding 0.2 SD when comparing modes according to interviewer presence, written vs aural delivery, reporting answers directly to an interviewer, and whether the interviewer was only present or collecting the responses. Similarly, items relating to one's social life appeared to more frequently exhibit mode effects, with 27.9% of estimates exceeding 0.2 SD when comparing modes according to interviewer presence, 51.2% for written vs aural questions, and 48.8% for reporting answers to an interviewer or not, however only 4.8% when comparing paper and electronic self-reported modes.

Figure 5. Mode effect distributions for the six most common categories of mode comparisons based on common mode characteristics of interest. All estimates are standardised and represent absolute values. Dashed lines represent a threshold of 0.2 SD, indicating the presence of (at least) a "small" effect size. Only estimates for which standardised effect sizes were available (80.6% of all estimates) are included.

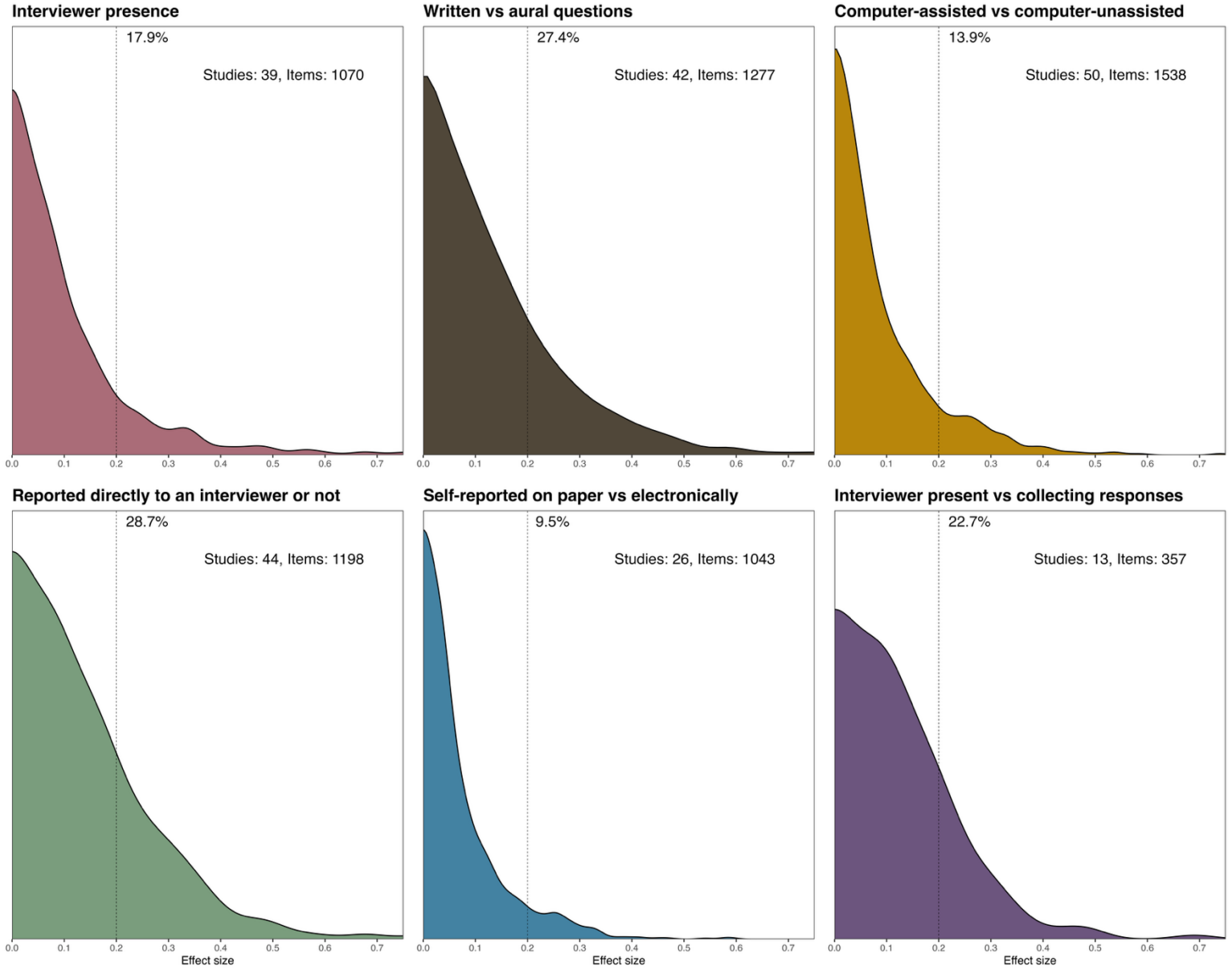
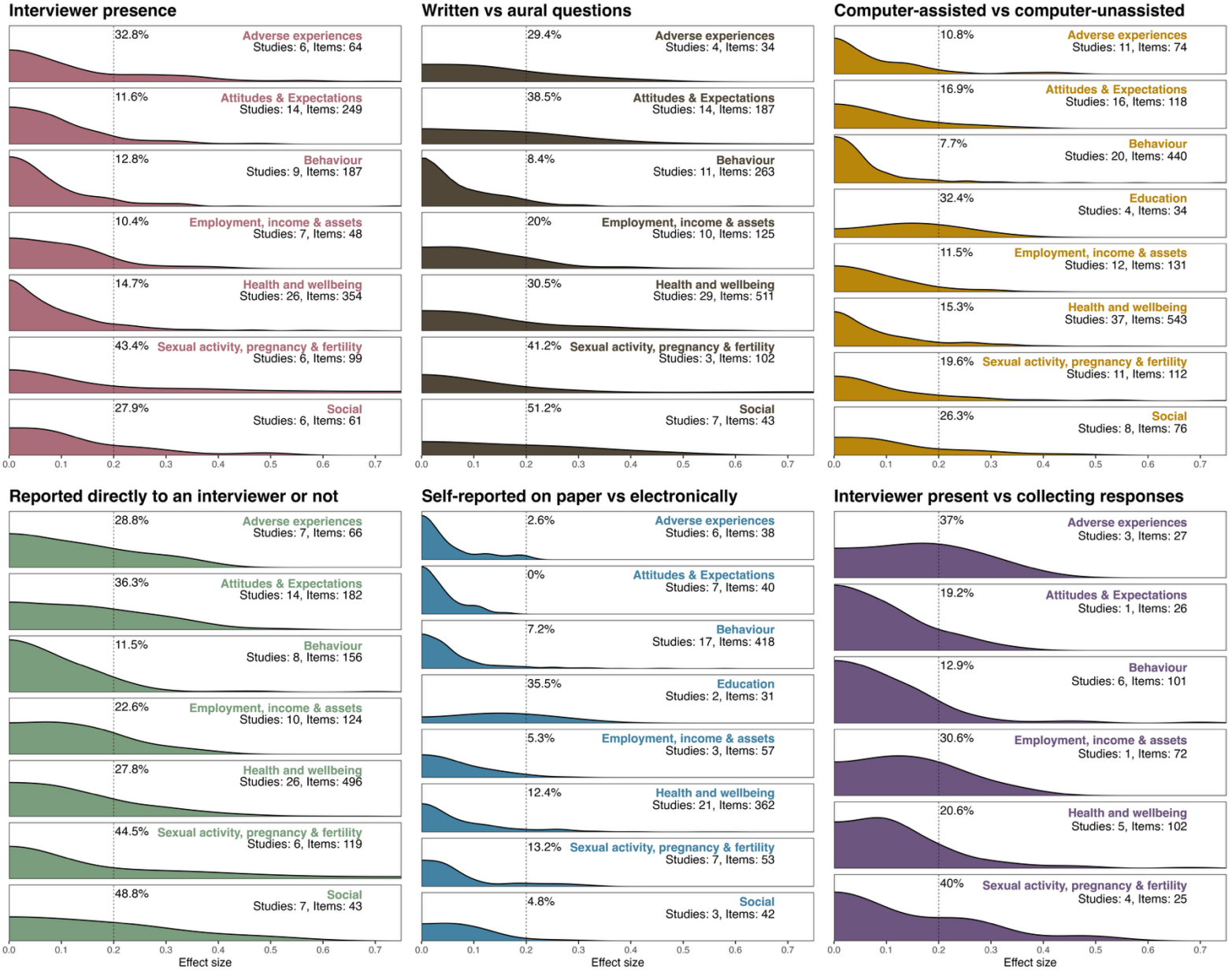


Figure 6. Mode effect distributions for the six most common categories of mode comparisons based on common mode characteristics of interest, broken down by item category. All estimates are standardised and represent absolute values. Dashed lines represent a threshold of 0.2 SD, indicating the presence of (at least) a "small" effect size. Only estimates for which standardised effect sizes were available (80.6% of all estimates) are included.



Database of results

All 4,113 extracted mode effect estimates alongside all other extracted information were collated to produce a searchable online database (https://cls-data.github.io/mode-effects-database/ and in **Supplementary File 2**).

Reporting quality

There were substantial differences in the range and depth of reporting across the studies. While standardised effect sizes were directly reported for only 83 (2.3%) of the unique items, they were possible to derive for the majority (n = 2,776, 78.3%) of others from the information provided, but it was not possible to extract or derive the effect sizes for 686 (19.4%) items, so they had to be excluded from the summary distributions. In some studies, it was not straightforward to ascertain all relevant elements of the mode, for example whether paper questionnaires were completed privately or in the presence of an interviewer, or whether specific items were self-completed as part of an otherwise face-to-face survey. It was also not possible to assess the degree of publication bias based on statistical significance since the majority of mode effects (n = 3,055, 86.2%) did not report p-values. Three studies (3.3%) provided statements referring to, explicitly or otherwise, only presenting statistically significant results. Six (6.7%) studies reported only the overall sample size of the study, but not the sample sizes in each arm. In 24 (26.7%) studies there were inconsistencies in the sample sizes reported throughout or it was not clear what the final analytical sample size was after any exclusions. Two (2.2%) studies did not report the sample size at all. The majority of studies (n = 53, 58.9%) did not report or discuss the extent of compliance to the randomly allocated mode. Almost half of all studies (n = 40, 44.4%) also did not discuss whether any post-randomisation processes such as differential non-response, mode switching, or other forms of non-compliance could have introduced issues with selection. Of the 50 studies that

acknowledged or discussed this, only just over half (n = 28, 56%) reported exploring or addressing this in some way, for example through weighting, adjustment, or the estimation of an alternative estimand. Thirty-five (38.9%) studies did not report or discuss whether item non-response differed across experimental arms or across items themselves. Eighteen (20%) studies did not discuss this explicitly, but it was indirectly implied by other information e.g. unexplained variation in analytic item sample sizes. Of the studies that reported this more explicitly (n = 37, 41.1%), 22 (59.5%) discussed that any existing differential item non-response is likely negligible, whereas 15 (40.5%) acknowledged it may have impacted the study results. Only 6 (6.7%) studies clearly defined one or more target estimands of interest, with 2 studies estimating the CACE (Feng and Huang 2024; Reisinger 2022), 2 studies estimating the ITT (McMorris et al. 2009; Ofstedal et al. 2022), and 2 studies estimating both the ITT and the CACE (Goodman et al. 2022; Smith et al. 2023).

Discussion

Overview of findings

Overall, we found that mode effects were 1) more likely to occur when modes differ in whether answers are provided directly to an interviewer or not, or whether the questions were presented visually or aurally, 2) highly driven by the item category (topic), and 3) predominantly very small, commonly below 0.2 SD. The smallest mode differences were observed when comparing paper and web or face-to-face and telephone modes, while larger mode differences were observed when comparing face-to-face and web, telephone and paper, or other interviewer-led and self-completion face-to-face modes. Sensitive items such as those relating to one's sexual activity, behaviour & pregnancy or social life tended to be subject to mode effects more commonly than others. However, there was also little

replication across studies, and almost none of the results referred to the same item compared between the same modes and in the same (or sufficiently similar) population. Overall, where mode effects exist, they are likely to be mode-, item-, and population-specific.

Most effect sizes were below 0.2 SD, suggesting mode effects are generally small in size. However, the degree to which a mode effect of 0.2 SD or above is likely to materially bias analyses depends on many factors. For example, where mixed-mode data inform the basis of major decisions (e.g., assessing the unemployment rate or other descriptive statistics), even small degrees of bias can be important, but less impact can be expected for causal effect estimation or where only the direction of effect may be of interest. There were only 37 (0.01%) mode effects that were particularly large (above 1 SD). Almost all came from a single study (Anglewicz et al. 2013), which examined responses to sensitive sexual activity questions reported face-to-face, self-completed using the ballot box method, or reported via randomised response (i.e. where respondents use a random mechanism to decide whether to answer truthfully or not, protecting their privacy (Warner 1965)). The large mode effects observed (up to 5 SD) all involved comparisons to a randomised response, which could by itself fully explain these large differences. Where other large mode effects occurred, they tended to be outliers within single studies, making it difficult to establish whether they are true, flukes, or data errors.

Very little previous evidence synthesis has been conducted in the context of mode effects, and where it has been conducted, it has been for limited types of survey items or mode comparisons. This being the first systematic review of its kind, it is not possible to directly compare it to similar existing literature. However, our findings are broadly consistent with recent recommendations on mitigating against mode effects (d'Ardenne et al. 2025). One of the main risks discussed in this framework were mode effects driven by the presence or absence of an interviewer. Indeed, our findings suggest that mode effects tend to be most

common between settings with different interviewer involvement, e.g., face-to-face and web, telephone and paper. This is important, since d'Ardenne *et al.* explained that, although implications may be *reduced*, it may not always be possible to completely prevent interviewer effects. The authors discussed that socially desirable or sensitive questions pose higher risk of mode effects, and we similarly observed that items concerning sexual activity, behaviour & pregnancy, social life, and health & wellbeing exhibited mode effects more frequently. d'Ardenne *et al.* also considered the risks of satisficing and presentation effects, however the scope of our data extraction did not include aspects such as question length, complexity, and presentation, or respondent behaviours like straightlining and satisficing.

Challenges for evidence synthesis on mode effects

It is possible that the existing literature on mode effects suffers some publication bias. Researchers may (reasonably) focus on items most likely to exhibit mode effects due to pre-existing beliefs. Indeed, Figure 3 suggests that most mode effects were examined for items relating to substance use, alcohol, mental health, and sexual activity, for which mode effects are more likely to be expected. Researchers may also (less reasonably) engage in selective reporting. In our review, three studies referred to only presenting statistically significant results for at least some of their findings (Fischer and Bayham 2019; Jäckle et al. 2010b; Sinadinovic et al. 2011), although this practice will likely have been more common as authors are rarely explicit about such decisions. P-values were not reported for most estimates (86.2%), making it difficult to empirically assess whether results were likely impacted by selective reporting or p-hacking.

Some results may have been specific to their domain and setting. For example, education-related items were commonly observed to exhibit mode effects, particularly between paper and electronic modes. However, these results mostly come from two large-scale educational

assessment studies (Fishbein et al. 2018; Jerrim et al. 2018) (see Fig. 4 and Fig. 6). These differ in nature from other types of studies such as social surveys and may therefore not necessarily transport reliably to other settings.

Insufficient reporting was common, especially regarding potential post-randomisation issues, such as differential non-compliance and non-response, but also for more basic elements of study reporting, e.g., sample size, response rate, uncertainty of estimates. This has implications for both evidence synthesis and the utility of the studies themselves in informing future survey design and analyses. The poor reporting may be partially explained by the lack of dedicated reporting criteria for survey data, e.g., the Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA) (Seidenberg et al. 2023) was only published in 2023.

Strengths and limitations

To our knowledge, this is the first systematic review of its kind, examining all mode comparisons for all health and social science survey items from experimental studies. In addition to synthesising mode effects in a variety of settings, over four thousand mode effect estimates were collated into an easily accessible online database (https://cls-data.github.io/mode-effects-database/). We employed an extensive systematic search strategy designed to capture as many relevant publications as possible. The addition of alternative sources such as grey literature and backwards and forwards citation screening helped to capture many studies that were not included in the standard databases. We also extracted a detailed set of auxiliary information alongside the mode effect estimates, aiding users to judge the relevance and validity of specific results.

However, due to the extensive scope of the work and with only 10% of studies double-screened and extracted, errors are possible. Although we exclusively focussed on

experimental studies to avoid the problems of mode selection in observational data, some selection is nevertheless possible due to differential non-response or non-compliance. A recent systematic review of experimental studies suggests that the odds of responding are 76% (95% CI: 34%, 132%) higher with mail compared to web mode (Edwards and Perkins 2024). There was some evidence that this may have occurred in some studies but due to the lack of sufficient reporting, it is not possible to gauge the extent to which the extracted mode effects may be affected by mode selection. Poor reporting may have also impacted the accuracy of the extracted information. Importantly, we were only able to synthesise mode effect estimates that had corresponding standardised effect sizes. Although these were available for most estimates (80.6%), the rest had to be excluded from the summary results as it is not possible to meaningfully synthesise unstandardised effects across different items. The precision of mode effect estimates varied across studies but was not accounted for in our synthesis. The classification of items and modes into categories was conducted manually and may not necessarily match the authors' original interpretation. The risk of such mismatch is greater when the quality of reporting is lower, due to difficulties in determining all relevant mode characteristics. The variety of populations examined also makes it challenging to produce reliable summary findings as different populations may not be comparable. Findings from one population may not necessarily transport to another – for instance, many concepts, such as what constitutes a socially desirable response, may differ between populations (Tellis and Chandrasekaran 2010).

Future research and recommendations

Based on the findings of this systematic review (both the substantive results and the observed quality of reporting), we suggest a set of actionable recommendations for researchers to consider.

Data from mixed-mode surveys where modes differ substantially based on the presence and/or involvement of an interviewer or based on whether questions are delivered aurally or visually, should be treated with caution, especially if sensitive or socially desirable responses may be anticipated. What constitutes 'appropriate' handling of any resulting mode effects may differ between studies, particularly depending on the expected degree of mode selection. A number of approaches appropriate for different scenarios have been described in detail elsewhere (Maslovskaya et al. 2020; Wright et al. 2024). Should researchers wish to conduct a QBA to examine the potential impact of mode effects on their substantive conclusions, the database produced as part of this systematic review can provide necessary external information.

In future research, to make studies examining mode effects more usable and relevant to others, we encourage researchers to prioritise the quality of reporting. In particular, we recommend that authors report sufficient information on the study design (including sampling strategy and randomisation), target estimand, randomisation compliance, unit and item non-response, starting and final sample size in each arm (including justifications for any exclusions), and an overview of how non-response or non-compliance were handled. We encourage the reporting of all conducted analyses, regardless of whether they were statistically significant, as well as confidence intervals (or standard errors) for all reported estimates, rather than binary indicators of statistical significance. A number of reporting guidelines exist to aid authors in the reporting of their studies, namely PRICSSA for complex sample surveys (Seidenberg et al. 2023) and the related CONSORT (CONsolidated Standards Of Reporting Trials) for randomised trials (Hopewell et al. 2025). First introduced in clinical journals, CONSORT was found to substantially improve the quality of reporting in randomised trials (Moher et al. 2001).

Where pooled effect estimates are of interest, given the range of mode comparisons, items, and populations, they should be produced from comparable (or transportable) effects, with sufficient evidence available for different populations. Future research should also prioritise replication to verify the reliability of existing evidence.

Conclusion

In this systematic review examining the experimental evidence of mode effects on item measurement, we found that mode effects were more likely to occur when comparing modes where answers were either directly reported to an interviewer or not, or whether questions were presented visually as opposed to aurally. The occurrence and size of mode effects varied by item category, but most mode effects were relatively small (below 0.2 SD). Future studies should prioritise appropriate reporting of all relevant study aspects and we provide a set of recommendations to support this.

Statements

Funding

This research was supported by UKRI-ESRC strategic research grant ES/X014150/1 for "Survey data collection methods collaboration: securing the future of social surveys", known as Survey Futures. The UCL Centre for Longitudinal Studies is also supported by the ESRC [grant number ES/W013142/1]. The funder had no role in the study conceptualisation, analysis, writing, or decision to publish.

Pre-registration

This review was pre-registered on the Open Science Framework (registration and protocol DOI: 10.17605/OSF.IO/BS5DW).

Conflicts of interest

None.

Acknowledgements

None.

Authorship (CRediT)

GDT: Conceptualisation, Data curation, Investigation, Formal analysis, Visualisation, Validation, Writing – original draft; RJS: Conceptualisation, Writing – review & editing, Funding acquisition; LW: Conceptualisation, Validation, Writing – review & editing, Funding acquisition.

References

Adalı, T., Türkyılmaz, A. S., and Lepkowski, J. M. (2022), "Evaluating the Demographic and Health Surveys Mode Switch From PAPI to CAPI: An Experiment From Turkey," *Social Science Computer Review*, 40, 1393–1415. https://doi.org/10.1177/08944393211009566.

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

Aneshensel, C. S., Frerichs, R. R., Clark, V. A., and Yokopenic, P. A. (1982), "Measuring Depression in the Community: A Comparison of Telephone and Personal Interviews," *Public Opinion Quarterly*, 46, 110. https://doi.org/10.1086/268703.

Anglewicz, P., Gourvenec, D., Halldorsdottir, I., O'Kane, C., Koketso, O., Gorgens, M., and Kasper, T. (2013), "The Effect of Interview Method on Self-Reported Sexual Behavior and Perceptions of Community Norms in Botswana," *AIDS and Behavior*, 17, 674–687. https://doi.org/10.1007/s10461-012-0224-z.

Aquilino, W. S. (1994), "Interview mode effects in surveys of drug and alcohol use: A field experiment," *Public Opinion Quarterly*, 58, 210–240. https://doi.org/10.1086/269419.

Aquilino, W. S. (1997), "Privacy Effects on Self-Reported Drug Use: Interactions With Survey Mode and Respondent," *The validity of self-reported drug use: Improving the accuracy of survey estimates*, US Department of Health and Human Services, National Institutes of Health …, 167, 383.

Aquilino, W. S. (1998), "Effects of interview mode on measuring depression in younger adults," *Journal of Official Statistics*, Statistics Sweden (SCB), 14, 15.

d'Ardenne, J., Bull, R., Das, A., Perera, Z., and Sexton, O. (2025), "Survey Practice Guide 2: How to mitigate against measurement effects when surveys move online."

Baier, D. (2018), "Computer-assisted versus paper-and-pencil self-report delinquency surveys: Results of an experimental study," *European Journal of Criminology*, 15, 385–402. https://doi.org/10.1177/1477370817743482.

Beach, S. R., Schulz, R., Degenholtz, H. B., Castle, N. G., Rosen, J., Fox, A. R., and Morycz, R. K. (2010), "Using audio computer-assisted self-interviewing and interactive voice response to measure elder mistreatment in older adults: Feasibility and effects on prevalence estimates," *Journal of Official Statistics*, 26, 507–533.

Béland, Y., and St-Pierre, M. (2007), "Mode Effects in the Canadian Community Health Survey: A Comparison of CATI and CAPI," in *Advances in Telephone Survey Methodology*, John Wiley & Sons, Ltd, pp. 297–314. https://doi.org/10.1002/9780470173404.ch14.

Berzelak, N., and Vehovar, V. (2024), "Mode effects on socially desirable responding in web surveys compared to face-to-face and telephone surveys," *Advances in Methodology and Statistics*, 15, 21–43. https://doi.org/10.51936/lrkv4884.

Bongers, I., and Van Oers, J. (1998), "Mode effects on self-reported alcohol use and problem drinking: mail questionnaires and personal interviewing compared.," *Journal of studies on alcohol*, Rutgers University Piscataway, NJ, 59, 280–285.

Brener, N. D., Eaton, D. K., Kann, L., Grunbaum, J. A., Gross, L. A., Kyle, T. M., and Ross, J. G. (2006), "The association of survey setting and mode with self-reported health risk behaviors among high school students," *Public Opinion Quarterly*, 70, 354–374. https://doi.org/10.1093/poq/nfl003.

Brown, M., and Calderwood, L. (2020), "Mixing modes in longitudinal surveys: an overview. CLS Working Paper 2020/3.," London: UCL Centre for Longitudinal Studies.

de Bruijne, M., and Wijnant, A. (2013), "Comparing Survey Results Obtained via Mobile Devices and Computers: An Experiment With a Mobile Web Survey on a Heterogeneous Group of Mobile Devices Versus a Computer-Assisted Web Survey," *Social Science Computer Review*, 31, 482–504. https://doi.org/10.1177/0894439313483976.

Burton, J., and Jäckle, A. (2020), "Mode effects," Understanding Society Working Paper Series.

Caeyers, B., Chalmers, N., and De Weerdt, J. (2012), "Improving consumption measurement and other survey data through CAPI: Evidence from a randomized experiment," *Journal of Development Economics*, 98, 19–33. https://doi.org/10.1016/j.jdeveco.2011.12.001.

Cea D'Ancona, M. (2014), "Measuring xenophobia: Social desirability and survey mode effects," *MIGRATION STUDIES*, 2, 255–280. https://doi.org/10.1093/migration/mnt014.

Cea D'Ancona, M. Á. (2017), "Measuring multiple discrimination through a survey-based methodology," *Social Science Research*, 67, 239–251. https://doi.org/10.1016/j.ssresearch.2017.04.006.

Cernat, A., Couper, M. P., and Ofstedal, M. B. (2016), "Estimation of mode effects in the health and retirement study using measurement models," *Journal of Survey Statistics and Methodology*, 4, 501–524. https://doi.org/10.1093/jssam/smw021.

Christensen, A. I., Ekholm, O., Glümer, C., and Juel, K. (2014), "Effect of survey mode on response patterns: comparison of face-to-face and self-administered modes in health surveys," *The European Journal of Public Health*, Oxford University Press, 24, 327–332.

Chromy, J., Davis, T., Packer, L., and Gfroerer, J. (2002), "Comparison of 1999 CAI and PAPI Data," *Redesigning an ongoing national household survey: Methodological issues*, Department of Health and Human Services, Substance Abuse and Mental Health …, 135.

Clarke, P. S., and Bao, Y. (2022), "ESTIMATING MODE EFFECTS FROM A SEQUENTIAL MIXED-MODE EXPERIMENT USING STRUCTURAL MOMENT MODELS," *Annals of Applied Statistics*, 16, 1563–1585. https://doi.org/10.1214/21-AOAS1557.

Cohen, J. (2009), *Statistical power analysis for the behavioral sciences*, New York, NY: Psychology Press.

Colasante, E., Benedetti, E., Fortunato, L., Scalese, M., Potente, R., Cutilli, A., and Molinaro, S. (2019), "Paper-and-pencil versus computerized administration mode: Comparison of data quality and risk behavior prevalence estimates in the European school Survey Project on Alcohol and other Drugs (ESPAD)," *PLoS ONE*, 14. https://doi.org/10.1371/journal.pone.0225140.

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., and Poole, C. (2010), "Illustrating bias due to conditioning on a collider," *International Journal of Epidemiology*, Oxford University Press (OUP), 39, 417–420. https://doi.org/10.1093/ije/dyp334.

Corporation for Digital Scholarship (2025), "Zotero."

Currivan, D. B. (2004), "Does Telephone Audio Computer-Assisted Self-Interviewing Improve the Accuracy of Prevalence Estimates of Youth Smoking?: Evidence from the UMass Tobacco Study," *Public Opinion Quarterly*, 68, 542–564. https://doi.org/10.1093/poq/nfh039.

Dahlhamer, J. M., Galinsky, A. M., and Joestl, S. S. (2019), "Asking about Sexual Identity on the National Health Interview Survey: Does Mode Matter?," *Journal of Official Statistics*, 35, 807–833. https://doi.org/10.2478/jos-2019-0034.

De Vitiis, C., Guandalini, A., Inglese, F., and Terribili, M. (2021), "Assessing and Adjusting Bias Due to Mixed-Mode in Aspect of Daily Life Survey," *JOURNAL OF OFFICIAL STATISTICS*, 37, 461–480. https://doi.org/10.2478/JOS-2021-0020.

DeLeeuw, E. D. (2018), "Mixed-Mode: Past, Present, and Future," *Survey Research Methods*, European Survey Research Association, Vol 12, 75-89 Pages. https://doi.org/10.18148/SRM/2018.V12I2.7402.

Dillman, D. A., and Tarnai, J. (2004), "Mode effects of cognitively designed recall questions: A comparison of answers to telephone and mail surveys," *Measurement errors in surveys*, Wiley Online Library, 73–93.

Domingue, B. W., McCammon, R. J., West, B. T., Langa, K. M., Weir, D. R., and Faul, J. (2023), "The Mode Effect of Web-Based Surveying on the 2018 U.S. Health and Retirement Study Measure of Cognitive Functioning," *The Journals of Gerontology: Series B*, (A. Gamaldo, ed.), 78, 1466–1473. https://doi.org/10.1093/geronb/gbad068.

Eaton, D. K., Brener, N. D., Kann, L., Denniston, M. M., McManus, T., Kyle, T. M., Roberts, A. M., Flint, K. H., and Ross, J. G. (2010), "Comparison of Paper-and-Pencil Versus Web Administration of the Youth Risk Behavior Survey (YRBS): Risk Behavior Prevalence Estimates," *Evaluation Review*, 34, 137–153. https://doi.org/10.1177/0193841X10362491.

Edwards, P., and Perkins, C. (2024), "Response is increased using postal rather than electronic questionnaires – new results from an updated Cochrane Systematic Review," *BMC Medical Research Methodology*, 24, 209. https://doi.org/10.1186/s12874-024-02332-0.

Erhart, M., Wetzel, R. M., Krügel, A., and Ravens-Sieberer, U. (2009), "Effects of phone versus mail survey methods on the measurement of health-related quality of life and emotional and behavioural problems in adolescents," *BMC Public Health*, 9, 491. https://doi.org/10.1186/1471-2458-9-491.

Feng, S., and Huang, F. (2024), "Does survey mode matter? An experimental evaluation of data quality in China," *China Economic Review*, 88, 102271. https://doi.org/10.1016/j.chieco.2024.102271.

Fischer, C. S., and Bayham, L. (2019), "Mode and Interviewer Effects in Egocentric Network Research," *Field Methods*, 31, 195–213. https://doi.org/10.1177/1525822X19861321.

Fishbein, B., Martin, M. O., Mullis, I. V. S., and Foy, P. (2018), "The TIMSS 2019 Item Equivalence Study: examining mode effects for computer-based assessment and implications for measuring trends," *Large-scale Assessments in Education*, 6, 11. https://doi.org/10.1186/s40536-018-0064-z.

Fleming, C. B., Marchesini, G., Elgin, J., Haggerty, K. P., Woodward, D., Abbott, R. D., and Catalano, R. F. (2013), "Use of Web and Phone Survey Modes to Gather Data from Adults about Their Young Adult Children: An Evaluation Based on a Randomized Design," *Field Methods*, 25, 388–404. https://doi.org/10.1177/1525822X12466888.

Fox, M. P., MacLehose, R. F., and Lash, T. L. (2021), *Applying Quantitative Bias Analysis to Epidemiologic Data*, Statistics for Biology and Health, Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-82673-4.

Goodman, A., Brown, M., Silverwood, R. J., Sakshaug, J. W., Calderwood, L., Williams, J., and Ploubidis, G. B. (2022), "The Impact of Using the Web in a Mixed-Mode Follow-up of a Longitudinal Birth Cohort Study: Evidence from the National Child Development Study," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185, 822–850. https://doi.org/10.1111/rssa.12786.

Haddaway, N. R., Collins, A. M., Coughlin, D., and Kirk, S. (2015), "The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching," *PLOS ONE*, (K. B. Wray, ed.), 10, e0138237. https://doi.org/10.1371/journal.pone.0138237.

Harmon, T., Turner, C. F., Rogers, S. M., Eggleston, E., Roman, A. M., Villarroel, M. A., Chromy, J. R., Ganapathi, L., and Li, S. (2009), "Impact of T-ACASI on Survey Measurements of Subjective Phenomena," *Public Opinion Quarterly*, 73, 255–280. https://doi.org/10.1093/poq/nfp020.

Helppie-Mcfall, B., and Hsu, J. W. (2017), "A test of web and mail mode effects in a financially sensitive survey of older Americans," *Journal of Economic and Social Measurement*, 42, 151–169. https://doi.org/10.3233/JEM-170444.

Hewett, P., Mensch, B., and Erulkar, A. (2004), "Consistency in the reporting of sexual behaviour by adolescent girls in Kenya: a comparison of interviewing methods," *SEXUALLY TRANSMITTED INFECTIONS*, 80, 43–48. https://doi.org/10.1136/sti.2004.013250.

Hochstim, J. R. (1967), "A Critical Comparison of Three Strategies of Collecting Data from Households," *Journal of the American Statistical Association*, 62, 976–989. https://doi.org/10.1080/01621459.1967.10500909.

Holford, A. J., and Pudney, S. (2015), *Survey Design and the Determinants of Subjective Wellbeing: An Experimental Analysis*, IZA Discussion Papers, Bonn: Institute for the Study of Labor (IZA).

Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., and Jäckle, A. (2022), "The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration," *Survey Research Methods*, 16, 207–226. https://doi.org/10.18148/srm/2022.v16i2.7771.

Hopewell, S., Chan, A.-W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., Tunn, R., Aggarwal, R., Berkwits, M., Berlin, J. A., Bhandari, N., Butcher, N. J., Campbell, M. K., Chidebe, R. C. W., Elbourne, D., Farmer, A., Fergusson, D. A., Golub, R. M., Goodman, S. N., Hoffmann, T. C., Ioannidis, J. P. A., Kahan, B. C., Knowles, R. L., Lamb, S. E., Lewis, S., Loder, E., Offringa, M., Ravaud, P., Richards, D. P., Rockhold, F. W., Schriger, D. L., Siegfried, N. L., Staniszewska, S., Taylor, R. S., Thabane, L., Torgerson, D., Vohra, S., White, I. R., and Boutron, I. (2025), "CONSORT 2025 statement: updated guideline for reporting randomised trials," *BMJ*, 389, e081123. https://doi.org/10.1136/bmj-2024-081123.

Jäckle, A., Roberts, C., and Lynn, P. (2010), "Assessing the Effect of Data Collection Mode on Measurement," *International Statistical Review*, 78, 3–20. https://doi.org/10.1111/j.1751-5823.2010.00102.x.

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., and McKeown, C. (2018), "PISA 2015: how big is the 'mode effect' and what has been done about it?," *Oxford Review of Education*, 44, 476–493. https://doi.org/10.1080/03054985.2018.1430025.

Jörngården, A., Wettergen, L., and Von Essen, L. (2006), "Measuring health-related quality of life in adolescents and young adults: Swedish normative data for the SF-36 and the HADS, and the influence of age, gender, and method of administration," *Health and Quality of Life Outcomes*, 4, 91. https://doi.org/10.1186/1477-7525-4-91.

Källmén, H., Sinadinovic, K., Berman, A. H., and Wennberg, P. (2011), "Risky Drinking of Alcohol in Sweden: A Randomized Population Survey Comparing Web- and Paper-based Self-reports," *Nordic Studies on Alcohol and Drugs*, 28, 123–130. https://doi.org/10.2478/v10199-011-0013-4.

Kim, S., and Couper, M. P. (2021), “Feasibility and Quality of a National RDD Smartphone Web Survey: Comparison With a Cell Phone CATI Survey,” *Social Science Computer Review*, 39, 1218–1236. https://doi.org/10.1177/0894439320964135.

Kim, Y., Dykema, J., Stevenson, J., Black, P., and Moberg, D. P. (2019), “Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys,” *Social Science Computer Review*, 37, 214–233. https://doi.org/10.1177/0894439317752406.

Klausch, T., Hox, J. J., and Schouten, B. (2013), “Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions,” *Sociological Methods & Research*, 42, 227–263. https://doi.org/10.1177/0049124113500480.

Kolenikov, S., and Kennedy, C. (2014), “Evaluating three approaches to statistically adjust for mode effects,” *Journal of Survey Statistics and Methodology*, American Association for Public Opinion Research, 2, 126–158.

Kreuter, F., Presser, S., and Tourangeau, R. (2008), “Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity,” *Public opinion quarterly*, Oxford University Press, 72, 847–865.

Krosnick, J. A. (1991), “Response strategies for coping with the cognitive demands of attitude measures in surveys,” *Applied Cognitive Psychology*, 5, 213–236. https://doi.org/10.1002/acp.2350050305.

Krosnick, J. A., and Alwin, D. F. (1987), “An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement,” *The Public Opinion Quarterly*, [Oxford University Press, American Association for Public Opinion Research], 51, 201–219.

Kumar, L. M., Stephen, J., George, R., Harikrishna, G., and Anisha, P. (2022), “Use of Effect Size in medical research: A brief primer on its why and how,” *Kerala Journal of Psychiatry*, 35, 322–322. https://doi.org/10.30834/KJP.35.1.2022.322.

Langhaug, L. F. (2009), “How you ask the question really matters: A randomized comparison of four questionnaire delivery modes to assess validity and reliability of self-reported socially censured data in rural Zimbabwean youth,” *Doctoral thesis, UCL (University College London).*, Doctoral, UCL (University College London).

Le, and Vu (2012), “Audio Computer-Assisted Self Interview Compared to Traditional Interview in an HIV-Related Behavioral Survey in Vietna,” *MEDICC Review*, 14, 26. https://doi.org/10.37757/MR2012V14.N4.7.

Leeuw, E. D. de, Hox, J. J., Dillman, D. A., and European Association of Methodology (eds.) (2008), *International handbook of survey methodology*, EAM book series, New York ; London: Lawrence Erlbaum Associates.

Li, J., Rico, A., Brener, N., Roberts, A., Mpofu, J., and Underwood, M. (2024), “Comparison of Paper-and-Pencil Versus Tablet Administration of the 2021 National Youth Risk Behavior Survey (YRBS),” *Journal of Adolescent Health*, 74, 814–819. https://doi.org/10.1016/j.jadohealth.2023.10.032.

Liu, M., Conrad, F. G., and Lee, S. (2017), “Comparing acquiescent and extreme response styles in face-to-face and web surveys,” *Quality & Quantity*, 51, 941–958. https://doi.org/10.1007/s11135-016-0320-7.

Liu, M., and Wang, Y. (2016), “Comparison of Face-to-Face and Web Surveys on the Topic of Homosexual Rights.”

Lucia, S., Herrmann, L., and Killias, M. (2007), “How important are interview methods and questionnaire designs in research on self-reported juvenile delinquency? An experimental comparison of Internet vs paper-and-pencil questionnaires and different definitions of the reference period,” *Journal of Experimental Criminology*, 3, 39–64. https://doi.org/10.1007/s11292-007-9025-1.

Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., and Greven, A. (2011), "Estimating Nonresponse Bias and Mode Effects in a Mixed-Mode Survey," *International Journal of Market Research*, 53, 669–686. https://doi.org/10.2501/IJMR-53-5-669-686.

Lygidakis, C., Rigon, S., Cambiaso, S., Bottoli, E., Cuozzo, F., Bonetti, S., Della Bella, C., and Marzo, C. (2010), "A web-based versus paper questionnaire on alcohol and tobacco in adolescents.," *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, 16, 925–930. https://doi.org/10.1089/tmj.2010.0062.

Mangunkusumo, R. T., Moorman, P. W., Van Den Berg-De Ruiter, A. E., Van Der Lei, J., De Koning, H. J., and Raat, H. (2005), "Internet-administered adolescent health questionnaires compared with a paper version in a randomized study," 36, 70.e1-70.e6. https://doi.org/10.1016/j.jadohealth.2004.02.020.

Maslovskaya, O., Calderwood, L., Ploubidis, G., and Nicolaas, G. (2020), "GenPopWeb2: Adjustments for Mode Effects."

McCabe, S. E., Boyd, C. J., Young, A., Crawford, S., and Pope, D. (2005), "Mode effects for collecting alcohol and tobacco data among 3rd and 4th grade students: a randomized pilot study of Web-form versus paper-form surveys," *Addictive Behaviors*, Elsevier, 30, 663–671.

McHorney, C. A., Kosinski, M., and Ware, J. E. (1994), "Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: Results from a national survey," *Medical Care*, 32, 551–567. https://doi.org/10.1097/00005650-199406000-00002.

McMorris, B. J., Petrie, R. S., Catalano, R. F., Fleming, C. B., Haggerty, K. P., and Abbott, R. D. (2009), "Use of Web and In-Person Survey Modes to Gather Data From Young Adults on Sex and Drug Use: An Evaluation of Cost, Time, and Survey Error Based on a Randomized Mixed-Mode Design," *Evaluation Review*, 33, 138–158. https://doi.org/10.1177/0193841X08326463.

Mensch, B. S., Hewett, P. C., Gregory, R., and Helleringer, S. (2008), "Sexual Behavior and STI/HIV Status Among Adolescents in Rural Malawi: An Evaluation of the Effect of Interview Mode on Reporting," *Studies in Family Planning*, 39, 321–334. https://doi.org/10.1111/j.1728-4465.2008.00178.x.

Midanik, L. T., and Greenfield, T. K. (2008), "Interactive Voice Response Versus Computer-Assisted Telephone Interviewing (CATI) Surveys and Sensitive Questions: The 2005 National Alcohol Survey," *Journal of Studies on Alcohol and Drugs*, 69, 580–588. https://doi.org/10.15288/jsad.2008.69.580.

Miech, R. A., Couper, M. P., Heeringa, S. G., and Patrick, M. E. (2021), "The impact of survey mode on US national estimates of adolescent drug prevalence: results from a randomized controlled study," *Addiction*, 116, 1144–1151. https://doi.org/10.1111/add.15249.

Moher, D., Jones, A., Lepage, L., and For The Consort Group (2001), "Use of the CONSORT Statement and Quality of Reports of Randomized Trials: A Comparative Before-and-After Evaluation," *JAMA*, 285, 1992. https://doi.org/10.1001/jama.285.15.1992.

Moskowitz, J. M. (2004), "Assessment of cigarette smoking and smoking susceptibility among youth: Telephone computer-assisted self-interviews versus computer-assisted telephone interview," *Public Opinion Quarterly*, 68, 565–587. https://doi.org/10.1093/poq/nfh040.

Nandi, A., and Platt, L. (2017), "Are there differences in responses to social identity questions in face-to-face versus telephone interviews? Results of an experiment on a longitudinal survey," *International Journal of Social Research Methodology*, 20, 151–166. https://doi.org/10.1080/13645579.2016.1165495.

Ofstedal, M. B., Kézdi, G., and Couper, M. P. (2022), "Data quality and response distributions in a mixed-mode survey," *Longitudinal and Life Course Studies*, 13, 621–646. https://doi.org/10.1332/175795921X16494126913909.

O'Muircheartaigh, C., Schumm, L. P., English, N., and Curtis, B. (2025), "Disentangling Selection into Mode from Mode Effects," *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, 80, S8–S16. https://doi.org/10.1093/geronb/gbae140.

Otsuka, Y., Kinjo, A., Kaneita, Y., Itani, O., Kuwabara, Y., Minobe, R., Maesato, H., Higuchi, S., Kanda, H., Yoshimoto, H., Jike, M., Kasuga, H., Ito, T., and Osaki, Y. (2023), "Comparison of the responses of cross-sectional web- and paper-based surveys on lifestyle behaviors of Japanese adolescents," *Preventive Medicine Reports*, 36, 102462. https://doi.org/10.1016/j.pmedr.2023.102462.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016), "Rayyan—a web and mobile app for systematic reviews," *Systematic Reviews*, 5, 210. https://doi.org/10.1186/s13643-016-0384-4.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021), "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, n71. https://doi.org/10.1136/bmj.n71.

Patrick, M. E., Couper, M. P., Parks, M. J., Laetz, V., and Schulenberg, J. E. (2021), "Comparison of a web-push survey research protocol with a mailed paper and pencil protocol in the Monitoring the Future panel survey," *Addiction*, 116, 191–199. https://doi.org/10.1111/add.15158.

Perkins, J. J., and Sanson-Fisher, R. W. (1998), "An examination of self- and telephone-administered modes of administration for the Australian SF-36," *Journal of Clinical Epidemiology*, 51, 969–973. https://doi.org/10.1016/S0895-4356(98)00088-2.

Piccitto, G., Liefbroer, A. C., and Emery, T. (2022), "Does the Survey Mode Affect the Association Between Subjective Well-being and its Determinants? An Experimental Comparison Between Face-to-Face and Web Mode," *Journal of Happiness Studies*, 23, 3441–3461. https://doi.org/10.1007/s10902-022-00553-y.

Reisinger, J. (2022), "Subjective well-being and social desirability," *Journal of Public Economics*, 214. https://doi.org/10.1016/j.jpubeco.2022.104745.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., PRISMA-S Group, Blunt, H., Brigham, T., Chang, S., Clark, J., Conway, A., Couban, R., De Kock, S., Farrah, K., Fehrmann, P., Foster, M., Fowler, S. A., Glanville, J., Harris, E., Hoffecker, L., Isojarvi, J., Kaunelis, D., Ket, H., Levay, P., Lyon, J., McGowan, J., Murad, M. H., Nicholson, J., Pannabecker, V., Paynter, R., Pinotti, R., Ross-White, A., Sampson, M., Shields, T., Stevens, A., Sutton, A., Weinfurter, E., Wright, K., and Young, S. (2021), "PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews," *Systematic Reviews*, 10, 39. https://doi.org/10.1186/s13643-020-01542-z.

Richman, W. L., Kiesler, S., Weisband, S., and Drasgow, F. (1999), "A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews.," *Journal of applied psychology*, American Psychological Association, 84, 754.

Roberts, C., Jäckle, A., and Lynn, P. (2006), "Causes of Mode Effects: Separating out Interviewer and Stimulus Effects in Comparisons of Face-to-Face and Telephone Surveys."

Roberts, C., Sarrasin, O., and Ernst Stähli, M. (2020), "Investigating the Relative Impact of Different Sources of Measurement Non-Equivalence in Comparative Surveys," *Survey Research Methods*, Survey Research Methods, 399-415 Pages. https://doi.org/10.18148/SRM/2020.V14I4.7416.

Rogers, S. M., Miller, H. G., and Turner, C. F. (1998), "Effects of interview mode on bias in survey measurements of drug use: Do respondent characteristics make a difference?," *Substance Use and Misuse*, 33, 2179–2200. https://doi.org/10.3109/10826089809069820.

Rosel, E., Tsakos, G., Bernabé, E., Sheiham, A., and Bravo, M. (2010), "Assessing the level of agreement between the self- and interview- administered Child-OIDP," *Community Dentistry and Oral Epidemiology*, 38, 340–347. https://doi.org/10.1111/j.1600-0528.2010.00533.x.

Sakshaug, J. W., Cernat, A., and Raghunathan, T. E. (2019), "Do Sequential Mixed-Mode Surveys Decrease Nonresponse Bias, Measurement Error Bias, and Total Bias? An Experimental Study," *Journal of Survey Statistics and Methodology*, 7, 545–571. https://doi.org/10.1093/jssam/smy024.

Sanchez Tome, R. (2018), "The impact of mode of data collection on measures of subjective wellbeing," University of Lausanne.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., and Klausch, T. (2013), "Disentangling mode-specific selection and measurement bias in social surveys," *Social Science Research*, 42, 1555–1570. https://doi.org/10.1016/j.ssresearch.2013.07.005.

Schumann, A., and Lück, D. (2023), "Better to ask online when it concerns intimate relationships? Survey mode differences in the assessment of relationship quality," *Demographic Research*, 48, 609–640. https://doi.org/10.4054/DEMRES.2023.48.22.

Seidenberg, A. B., Moser, R. P., and West, B. T. (2023), "Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA)," *Journal of Survey Statistics and Methodology*, 11, 743–757. https://doi.org/10.1093/jssam/smac040.

Sinadinovic, K., Wennberg, P., and Berman, A. H. (2011), "Population screening of risky alcohol and drug use via Internet and Interactive Voice Response (IVR): A feasibility and psychometric study in a random sample," *Drug and Alcohol Dependence*, 114, 55–60. https://doi.org/10.1016/j.drugalcdep.2010.09.004.

Šmigelskas, K., Lukoševičiute, J., Vaičlunas, T., Mozuraityte, K., Iva-Navičiute, U., Milevičiute, I., and Žemaitaityte, M. (2019), "Measurement of health and social behaviors in schoolchildren: Randomized study comparing paper versus electronic mode," *Zdravstveno Varstvo*, 58, 1–10. https://doi.org/10.2478/sjph-2019-0001.

Smith, J. R., Gibbons, L. E., Crane, P. K., Mungas, D. M., Glymour, M. M., Manly, J. J., Zahodne, L. B., Rose Mayeda, E., Jones, R. N., and Gross, A. L. (2023), "Shifting of Cognitive Assessments Between Face-to-Face and Telephone Administration: Measurement Considerations," *The Journals of Gerontology: Series B*, (A. Gamaldo, ed.), 78, 191–200. https://doi.org/10.1093/geronb/gbac135.

Stegen, H., Duppen, D., Savieri, P., Stas, L., Pan, H., Aartsen, M., Callewaert, H., Dierckx, E., and De Donder, L. (2024), "Loneliness prevalence of community-dwelling older adults and the impact of the mode of measurement, data collection, and country: A systematic review and meta-analysis," *International Psychogeriatrics*, 36, 747–761. https://doi.org/10.1017/S1041610224000425.

Sullivan, G. M., and Feinn, R. (2012), "Using Effect Size—or Why the P Value Is Not Enough," *Journal of Graduate Medical Education*, 4, 279–282. https://doi.org/10.4300/JGME-D-12-00156.1.

Supple, A. J., Aquilino, W. S., and Wright, D. L. (1999), "Collecting sensitive self-report data with laptop computers: Impact on the response tendencies of adolescents in a home interview," *Journal of Research on Adolescence*, 9, 467–488. https://doi.org/10.1207/s15327795jra0904_5.

Tellis, G. J., and Chandrasekaran, D. (2010), "Extent and impact of response biases in cross-national survey research," *International Journal of Research in Marketing*, 27, 329–341. https://doi.org/10.1016/j.ijresmar.2010.08.003.

Tomova, G. D., Silverwood, R. J., Tennant, P. W., and Wright, L. (2025a), "How can the use of different modes of survey data collection introduce bias? A simple introduction to mode effects using directed acyclic graphs (DAGs)," arXiv. https://doi.org/10.48550/ARXIV.2510.00900.

Tomova, G. D., Silverwood, R. J., and Wright, L. (2025b), "A systematic review of the (quasi-)experimental evidence of survey mode effects on item measurement," OSF Registries. https://doi.org/10.17605/OSF.IO/BS5DW.

Tourangeau, R., and Smith, T. W. (1996), "Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context," *Public Opinion Quarterly*, 60, 275. https://doi.org/10.1086/297751.

Tourangeau, R., and Yan, T. (2007), "Sensitive questions in surveys.," *Psychological Bulletin*, 133, 859–883. https://doi.org/10.1037/0033-2909.133.5.859.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., and Sonenstein, F. L. (1998), "Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology," *Science*, American Association for the Advancement of Science, 280, 867–873.

Turner, C. F., Villarroel, M. A., Rogers, S. M., Eggleston, E., Ganapathi, L., Roman, A. M., and Al-Tayyib, A. (2005), "Reducing bias in telephone survey estimates of the prevalence of drug use: a randomized trial of telephone audio-CASI," *Addiction*, 100, 1432–1444. https://doi.org/10.1111/j.1360-0443.2005.01196.x.

Van De Looij-Jansen, P. M., and De Wilde, E. J. (2008), "Comparison of Web-Based versus Paper-and-Pencil Self-Administered Questionnaire: Effects on Health Indicators in Dutch Adolescents," *Health Services Research*, 43, 1708–1721. https://doi.org/10.1111/j.1475-6773.2008.00860.x.

Van Den Akker, O., Peters, G.-J. Y., Bakker, C., Carlsson, R., Coles, N. A., Corker, K. S., Feldman, G., Mellor, D. T., Moreau, D., Nordström, T., Pfeiffer, N., Pickering, J. S., Riegelman, A., Topor, M., Van Veggel, N., and Yeung, S. K. (2020), "Increasing the Transparency of Systematic Reviews: Presenting a Generalized Registration Form," MetaArXiv. https://doi.org/10.31222/osf.io/3nbea.

VanderWeele, T. J., and Li, Y. (2019), "Simple Sensitivity Analysis for Differential Measurement Error," *American Journal of Epidemiology*, 188, 1823–1829. https://doi.org/10.1093/aje/kwz133.

Vannieuwenhuyze, J. T. A., and Loosveldt, G. (2013), "Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects," *Sociological Methods & Research*, 42, 82–104. https://doi.org/10.1177/0049124112464868.

Vannieuwenhuyze, J. T. A., Loosveldt, G., and Molenberghs, G. (2012), "A Method to Evaluate Mode Effects on the Mean and Variance of a Continuous Variable in Mixed-Mode Surveys," *International Statistical Review*, 80, 306–322. https://doi.org/10.1111/j.1751-5823.2011.00167.x.

Villarroel, M. A. (2006), "Same-Gender Sex in the United States: Impact of T-Acasi on Prevalence Estimates," *Public Opinion Quarterly*, 70, 166–196. https://doi.org/10.1093/poq/nfj023.

Villarroel, M. A., Turner, C. F., Rogers, S. M., Roman, A. M., Cooley, P. C., Steinberg, A. B., Eggleston, E., and Chromy, J. R. (2008), "T-ACASI Reduces Bias in STD Measurements: The National STD and Behavior Measurement Experiment," *Sexually Transmitted Diseases*, 35, 499–506. https://doi.org/10.1097/OLQ.0b013e318165925a.

Wang, Y.-C., Lee, C.-M., Lew-Ting, C.-Y., Hsiao, C. K., Chen, D.-R., and Chen, W. J. (2005), "Survey of substance use among high school students in Taipei: Web-based questionnaire versus paper-and-pencil questionnaire," *Journal of Adolescent Health*, 37, 289–295. https://doi.org/10.1016/j.jadohealth.2005.03.017.

Warner, S. L. (1965), "Randomized response: a survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 60, 63–66.

Wells, T., Bailey, J. T., and Link, M. W. (2014), "Comparison of Smartphone and Online Computer Survey Administration," *Social Science Computer Review*, 32, 238–255. https://doi.org/10.1177/0894439313505829.

Wettergren, L., Mattsson, E., and von Essen, L. (2011), "Mode of administration only has a small effect on data quality and self-reported health status and emotional distress among Swedish adolescents and young adults," *Journal of Clinical Nursing*, 20, 1568–1577. https://doi.org/10.1111/j.1365-2702.2010.03481.x.

Wright, D. L., Aquilino, W. S., and Supple, A. J. (1998), "A comparison of computer-assisted and paper-and-pencil self-administered questionnaires in a survey on smoking, alcohol, and drug use," *Public Opinion Quarterly*, 62, 331–353. https://doi.org/10.1086/297849.

Wright, L., Ploubidis, G., and Silverwood, R. (2024), "Handling mode effects in the CLS cohort studies: User guide."

Ye, C., Fulton, J., and Tourangeau, R. (2011), "More positive or more extreme? A meta-analysis of mode differences in response choice," *Public Opinion Quarterly*, Oxford University Press, 75, 349–365.

ZuWallack, R., Jans, M., Brassell, T., Bailly, K., Dayton, J., Martinez, P., Patterson, D., Greenfield, T. K., and Karriker-Jaffe, K. J. (2023), "Estimating Web Survey Mode and Panel Effects in a Nationwide Survey of Alcohol Use," *Journal of Survey Statistics and Methodology*, 11, 1089–1109. https://doi.org/10.1093/jssam/smac028.

Supplementary materials

Supplementary File 1
Download: https://osf.io/nj256/files/29eu7

Supplementary File 2
Download: https://osf.io/nj256/files/am9je

Supplementary Table 1. Deviations from the pre-registered protocol and accompanying justifications.

Location	Original text	Amendment/deviation	Reason
Title	A Systematic Review Of The (Quasi-)Experimental Evidence Of Survey Mode Effects On Item Measurement	A systematic review of the experimental evidence of survey mode effects on item measurement	“Quasi-“ was removed from the title to better reflect the content of the review since no studies with quasi-experimental designs (according to treatment allocation) met all inclusion criteria.
Exclusion criteria	Sample from a population defined by clinical or occupational characteristics (e.g. teachers, people with diabetes, psychology students)	Sample from a population defined by clinical or occupational characteristics (e.g. teachers, people with diabetes, psychology students) or other characteristics not limited to age, sex, and geographical region	To avoid any potential misinterpretation and improve clarity of the exclusion criteria regarding population.
Google Scholar search query	(“mode” AND (“effect” OR “difference” OR "differ by")) AND (“survey” OR “cohort” OR “study” OR “data collect*”) AND (“mixed-mode” OR “interview” OR “face-to-face” OR “f2f” OR “ftf” OR “web” OR “online” OR “internet” OR “mobile” OR “mail” OR “phone” OR “telephone” OR “video” OR “paper” OR “paper-and-pencil” OR “paper-pencil” OR “VMI” OR “PAPI” OR “CASI” OR “CASQ” OR “SAQ” OR “CAPI” OR “ACASI” OR “computer-assisted” OR “self-administered”) AND (“experiment” OR “experimental” OR “randomly assigned” OR	(“mode” AND (“effect” OR “difference” OR "differ by")) AND (“survey” OR “cohort” OR “study” OR “data collect*” OR "questionnaire") AND (“mixed-mode” OR “interview” OR “face-to-face” OR “f2f” OR “ftf” OR “web” OR “online” OR “internet” OR “mobile” OR “mail” OR “phone” OR “telephone” OR “video” OR “paper” OR “paper-and-pencil” OR “paper-pencil” OR “VMI” OR “PAPI” OR “CASI” OR “CASQ” OR “SAQ” OR “CAPI” OR “ACASI” OR “computer-assisted” OR “self-administered”) AND (“experiment” OR “experimental” OR “randomly assigned” OR “randomised” OR “quasi”)	The term "re-interview" was removed and the term "questionnaire" included, in line with the search queries for other databases. The discrepancy was previously introduced in error based on an older preliminary version of the search.

	"randomised" OR "quasi" OR "re-interview")		
Other search strategies	Once all articles have been screened and the appropriate articles identified, we will review citations by implementing both the ascendancy and descendancy approaches to identify other potentially relevant sources that may have been missed in the formal search process.	Once all articles have been screened and the appropriate articles identified, we will review citations by implementing both the ascendancy and descendancy approaches to identify other potentially relevant sources that may have been missed in the formal search process. Additionally, if any systematic reviews or meta-analyses are identified during the screening stage, their reference lists will be screened for potentially relevant articles.	During the screening process, we found systematic reviews to be useful sources for identifying additional studies for screening, which may not have been picked up using our normal search. Although we implemented this amendment, none of the studies included in our review came from the screened reference lists of systematic reviews.
Screening stages	[*No information on citation and systematic review screen*]	CITATION SCREEN The following steps will be applied to all articles included after the full-text screen. The citation screen will be conducted by a single reviewer with no double-screening. 1. Title and abstract screen The titles of all citations, based on both the ascendancy and descendancy approaches, will be reviewed. This will include the list of references provided within each article, as well as all citations of an article identified by Google Scholar. Where an article has more than 500 citations, the Google Scholar-specific search will be conducted within the citations to narrow	The procedure for citation screening was missed in the original pre-registration, and systematic review screening was not originally planned at the time of initial protocol pre-registration.

down the results. Otherwise, the titles of all citations will be screened. Where a title is identified as potentially relevant, its abstract will be located online and screened, unless the article has already been previously included. Where an abstract or full-text version cannot be located anywhere online, the study will be excluded.

2. Deduplication
The first step of the citation screen is designed to avoid duplicates by only screening abstracts of articles that have not been previously included, which will be assessed manually. However, if any studies have accidentally been included twice, any duplicates will be identified and resolved using Rayyan.

3. Full-text screen
Any study which has reached the full-text screen stage will be screened in the same way as described in the general screen stage above.

SYSTEMATIC REVIEW SCREEN
This screen will be applied to any systematic reviews or meta-analyses identified during the general screening stage. The process will be the same as the one outlined for the citation screen

		above, with the exception that only the ascendancy approach will be used, i.e. we will screen the reference lists of each systematic review but not other studies that have cited the review. The systematic review screen will be conducted by a single reviewer with no double-screening.	
Entities to extract	Author-reported risk of bias tools	*[deleted]*	Found not to be applicable for the types of studies examined.
Entities to extract	Missing data	*[deleted]*	Missing data information was instead recorded in other quality of reporting sections to maintain simplicity of the data extraction form and resulting database
Planned data transformations	Where possible, we will aim to derive and report all mode effect estimates in terms of both absolute and standardised effect sizes. For binary variables, where possible, we will report absolute and relative risk differences.	Mode effects from binary variables were extracted as reported in each study, without additional transformation (e.g. from absolute to relative).	To maintain feasibility given the amount of data extraction, we focussed on providing standardised effect sizes for all possible variables, but did not additionally derive relative risk differences for binary variables, unless reported in the studies.
Publication bias analyses	Although we will not be estimating any summary effects, our findings may still be impacted by publication bias due to selective reporting or decisions (not) to publish. Where a study reports considering multiple variables but only provides estimates for a selection of these, and especially without justification, we will record this as	No p-curve analysis was conducted.	The majority of studies did not report p-values. It was also a common practice for studies to report all estimates, but only report p-values for the significant ones, and no p-values for the non-significant.

	part of our risk of bias data extraction. If possible, we will also conduct a p-curve analysis to identify potential publication bias related to the statistical significance of the results.		
Synthesis data management and sharing	The findings from the systematic review will be combined into a freely available searchable database, in the form of an R Shiny app. Links to any relevant sources of information will be provided (e.g. journal article links, cohort profile links). Any R scripts used in the process will be made available on GitHub.	Html page instead of R Shiny app. No cohort profile links were provided.	An html page provided faster loading. Due to the number of studies and surveys, we deemed it sufficient to provide a link to the study itself, but not to each published cohort profile. However, we expect the studies to contain a reference to the cohort profile, where applicable.

Supplementary Table 2. Query strings used to search each database. The strings were designed to be as similar as possible between databases, given the expected syntax.

Interface	Database	Search query
Elsevier	Scopus	(TITLE-ABS-KEY (mode W/4 (effect OR difference OR “differ by”)) AND TITLE-ABS-KEY (survey* OR stud* OR (data W/2 collect*) OR cohort* OR questionnaire*) AND TITLE-ABS-KEY ("mixed-mode" OR interview* OR "face-to-face" OR f2f OR ftf OR web OR online OR internet OR mobile OR mail* OR phone OR telephone OR video OR paper OR "paper-and-pencil" OR “paper-pencil” OR {VMI} OR {PAPI} OR {CASI} OR {CSAQ} OR {SAQ} OR {CAPI} OR {ACASI} OR (comput* W/2 assist* W/2 interview*) OR “self-administ*”) AND TITLE-ABS-KEY (experiment* OR quasi* OR random*) AND (LIMIT-TO (SUBJAREA , "SOCI") OR LIMIT-TO (SUBJAREA , "MATH") OR LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MEDI") OR LIMIT-TO (SUBJAREA , "DECI") OR LIMIT-TO (SUBJAREA , "ECON") OR LIMIT-TO (SUBJAREA , "PSYC") OR LIMIT-TO (SUBJAREA , "MULT") OR LIMIT-TO (SUBJAREA , "HEAL") OR LIMIT-TO (SUBJAREA , "NURS")) AND (LIMIT-TO (LANGUAGE , "English"))
Ovid	Embase 1974, MEDLINE 1946, Health and Psychosocial Instruments, PsycINFO, PsycEXTRA	1 (mode adj4 (effect or difference or "differ by")).ti,ot,ab,kf,sh,ox. 2 (survey* or stud* or cohort* or (data adj2 collect*) or questionnaire*).ti,ot,ab,kf,sh,ox. 3 ("mixed-mode" or interview* or "face-to-face" or f2f or ftf or web or online or internet or mobile or mail* or phone or telephone or video or paper or "paper-and-pencil" or "paper-pencil" or VMI or PAPI or CASI or CSAQ or SAQ or CAPI or ACASI or (comput* and assist* and interview*) or "self-administ*").ti,ot,ab,kf,sh,ox. 4 (experiment* or quasi* or random*).ti,ot,ab,kf,sh,ox. 5 1 and 2 and 3 and 4

		6 remove duplicates from 5
Clarivate	Web of Science Core Collection	1: ((TI=(mode NEAR/4 (effect OR difference OR "differ by"))) OR AB=(mode NEAR/4 (effect OR difference OR "differ by"))) OR KP=(mode NEAR/4 (effect OR difference OR "differ by")) 2: ((TI=(survey* OR stud* OR (data NEAR/2 collect*) OR cohort* OR questionnaire*)) OR AB=(survey* OR stud* OR (data NEAR/2 collect*) OR cohort* OR questionnaire*)) OR KP=(survey* OR stud* OR (data NEAR/2 collect*) OR cohort* OR questionnaire*) 3: ((TI=("mixed-mode" or interview* or "face-to-face" or f2f or ftf or web or online or internet or mobile or mail* or phone or telephone or video or paper or "paper-and-pencil" or "paper-pencil" or VMI or PAPI or CASI or CSAQ or SAQ or CAPI or ACASI or (comput* NEAR/2 assist* NEAR/2 interview*) or "self-administ*")) OR AB=("mixed-mode" or interview* or "face-to-face" or f2f or ftf or web or online or internet or mobile or mail* or phone or telephone or video or paper or "paper-and-pencil" or "paper-pencil" or VMI or PAPI or CASI or CSAQ or SAQ or CAPI or ACASI or (comput* NEAR/2 assist* NEAR/2 interview*) or "self-administ*")) OR KP=("mixed-mode" or interview* or "face-to-face" or f2f or ftf or web or online or internet or mobile or mail* or phone or telephone or video or paper or "paper-and-pencil" or "paper-pencil" or VMI or PAPI or CASI or CSAQ or SAQ or CAPI or ACASI or (comput* NEAR/2 assist* NEAR/2 interview*) or "self-administ*") 4: ((TI=(experiment* or quasi* or random*)) OR AB=(experiment* or quasi* or random*)) OR KP=(experiment* or quasi* or random*) 5: #4 AND #3 AND #2 AND #1 6: #4 AND #3 AND #2 AND #1 and English (Languages)
Google Scholar		(“mode” AND (“effect” OR “difference” OR "differ by”)) AND (“survey” OR “cohort” OR “study” OR “data collect*” OR "questionnaire") AND (“mixed-mode” OR “interview” OR “face-to-face” OR “f2f” OR “ftf” OR “web” OR “online” OR “internet” OR “mobile” OR

	“mail” OR “phone” OR “telephone” OR “video” OR “paper” OR “paper-and-pencil” OR “paper-pencil” OR “VMI” OR “PAPI” OR “CASI” OR “CASQ” OR “SAQ” OR “CAPI” OR “ACASI” OR “computer-assisted” OR “self-administered”) AND (“experiment” OR “experimental” OR “randomly assigned” OR “randomised” OR “quasi”)

Supplementary Table 3. Data extraction items which were manually derived from the available information.

Derived item	Calculation
Mode effect	$\overline{X}_A$, mean measure in mode A $\overline{X}_B$, mean measure in mode B $ME = \overline{X}_A - \overline{X}_B$
Mode effect standard error	n_A, sample size in mode A n_B, sample size in mode B s_A, outcome SD in mode A s_B, outcome SD in mode B $SE = \sqrt{\left(\frac{s_A^2}{n_A}\right) + \left(\frac{s_B^2}{n_B}\right)}$
Mode effect 95% confidence interval	$95\%\ \mathrm{CI} = ME \pm 1.96 \cdot SE$
Standardised effect size (Glass's delta)	$\overline{X}_A$, mean measure in mode A $\overline{X}_B$, mean measure in mode B s_A, outcome SD in mode A (reference mode) $\Delta = \frac{(\overline{X}_A - \overline{X}_B)}{s_A}$
Glass's delta standard error	n_A, sample size in mode A (reference mode) n_B, sample size in mode B $SE_\Delta = \sqrt{\frac{n_A + n_B}{n_A n_B} + + \frac{\Delta^2}{2(n_A - 1)}}$

Standard deviation of a proportion variable	p, proportion (0-1)	$SD = \sqrt{p(1-p)}$

Supplementary Table 4. Classification of modes according to different characteristics.

Interviewer physical presence		Question delivery		Computer-assisted survey		Reporting of answers		Reporting to an interviewer		Self-administered		Type of interviewer involvement	
0 (not present)	1 (present)	0 (aural)	1 (written)	0 (not computer-assisted)	1 (computer-assisted)	0 (not reported to an interviewer)	1 (reported to an interviewer)	0 (in-person)	1 (over the phone)	0 (on paper)	1 (on web)	0 (present)	1 (collecting responses)
Web Telephone CATI Paper (mailed) Computer** Mobile Hybrid IVR T-ACASI Randomised response Ballot box	Face-to-face CAPI CASI ACASI PAPI Paper (self-administered)* Paper (self-administered, with audio soundtrack) Tablet** Computer** Netbook**	Telephone CATI CAPI Face-to-face PAPI ACASI IVR T-ACASI Paper (self-administered, with audio soundtrack)	Paper (self-administered)* Paper (mailed) Web CASI Mobile Computer Tablet Netbook Hybrid Ballot box	Face-to-face Telephone PAPI Paper (self-administered)* Paper (mailed) Paper (self-administered, with audio soundtrack) IVR Randomised response Ballot box	CAPI CATI CASI ACASI T-ACASI Web Mobile Computer Hybrid Netbook	Paper (self-administered)* Paper (mailed) Paper (self-administered, with audio soundtrack) CASI ACASI IVR Web Computer Tablet Netbook Hybrid Randomised response Ballot box	Face-to-face Telephone CAPI CATI PAPI T-ACAS	Face-to-face CAPI PAPI	Telephone CATI T-ACASI	Paper (self-administered)* Paper (mailed) Paper (self-administered, with audio soundtrack)	Web CASI ACASI Computer Mobile Hybrid Tablet Netbook	Paper (self-administered)* Paper (self-administered, with audio soundtrack) CASI ACASI Computer**	Face-to-face Telephone CAPI PAPI CATI T-ACASI

*evaluated on a study-by-study basis; **a self-administered paper questionnaire in the presence of an interviewer; for self-administered paper questionnaire with no interviewer present, see “Paper (mailed)”

Supplementary Table 5. PRISMA checklist.

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	p.8
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	p.9
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Table 1
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	p. 9-10 “Search and screening strategy”
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Supplementary Table 2
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	p. 9 and p. 14 “Validation”
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	p.12-13
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	p.12-13, Table 2, Supplementary File 1
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	p.12-13, Table 2, Supplementary File 1
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	NA – rather than using a tool, qualitative information on limitations and risk of bias was extracted for each study and made available in the associated online database
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	p. 12 “Where a mode effect was presented using more than one type of effect measure (e.g. both a mean difference between modes as well as an odds ratio), then both were extracted as separate entries.”
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	p.14-15

Section and Topic	Item #	Checklist item	Location where item is reported
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	p.12-13, p.14-15
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	p.14-15
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	p.14-15
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	NA – no pooled estimates produced
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	NA – no pooled estimates produced
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	NA – no pooled estimates produced
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	NA – systematic review not focussed on a single outcome/domain
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	p.15-16, Figure 1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	NA
Study characteristics	17	Cite each included study and present its characteristics.	Table 3, Supplementary File 2
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Qualitative risk of bias available in Supplementary File 2 and associated online database
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	NA – due to different nature of review, but equivalent information is available in Supplementary File 2 and Figures 3-6
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	NA
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Figures 3-6
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	NA – not enough replication to examine heterogeneity per outcome
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	NA – no sensitivity analyses
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	p. 30
Certainty of	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	p. 33-34 (narrative)

Section and Topic	Item #	Checklist item	Location where item is reported
evidence			
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	p. 31-33
	23b	Discuss any limitations of the evidence included in the review.	p. 33-34
	23c	Discuss any limitations of the review processes used.	p. 34-35
	23d	Discuss implications of the results for practice, policy, and future research.	p. 35-36
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	p.38, p.9
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	p. 38
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Supplementary Table 1
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	p. 38
Competing interests	26	Declare any competing interests of review authors.	p. 38
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	p. 30 “Database of results”

Supplementary Table 6. PRISMA-S checklist.

Section/topic	#	Checklist item	Location(s) Reported
INFORMATION SOURCES AND METHODS			
Database name	1	Name each individual database searched, stating the platform for each.	p. 10
Multi-database searching	2	If databases were searched simultaneously on a single platform, state the name of the platform, listing all of the databases searched.	Supplementary Table 2
Study registries	3	List any study registries searched.	NA
Online resources and browsing	4	Describe any online or print source purposefully searched or browsed (e.g., tables of contents, print conference proceedings, web sites), and how this was done.	NA
Citation searching	5	Indicate whether cited references or citing references were examined, and describe any methods used for locating cited/citing references (e.g., browsing reference lists, using a citation index, setting up email alerts for references citing included studies).	p. 11
Contacts	6	Indicate whether additional studies or data were sought by contacting authors, experts, manufacturers, or others.	NA
Other methods	7	Describe any additional information sources or search methods used.	NA
SEARCH STRATEGIES			
Full search strategies	8	Include the search strategies for each database and information source, copied and pasted exactly as run.	Supplementary Table 2
Limits and restrictions	9	Specify that no limits were used, or describe any limits or restrictions applied to a search (e.g., date or time period, language, study design) and provide justification for their use.	Supplementary Table 2 and Pre-registered protocol
Search filters	10	Indicate whether published search filters were used (as originally designed or modified), and if so, cite the filter(s) used.	NA
Prior work	11	Indicate when search strategies from other literature reviews were adapted or reused for a substantive part or all of the search, citing the previous review(s).	NA
Updates	12	Report the methods used to update the search(es) (e.g., rerunning searches, email alerts).	p. 9, p. 14 (pilot stage)
Dates of searches	13	For each search strategy, provide the date when the last search occurred.	p. 10-11
PEER REVIEW			
Peer review	14	Describe any search peer review process.	Not conducted, but search validated in pilot stage – p.14
MANAGING RECORDS			
Total Records	15	Document the total number of records identified from each database and other information sources.	Figure 1
Deduplication	16	Describe the processes and any software used to deduplicate records from multiple database searches and other information sources.	p. 11

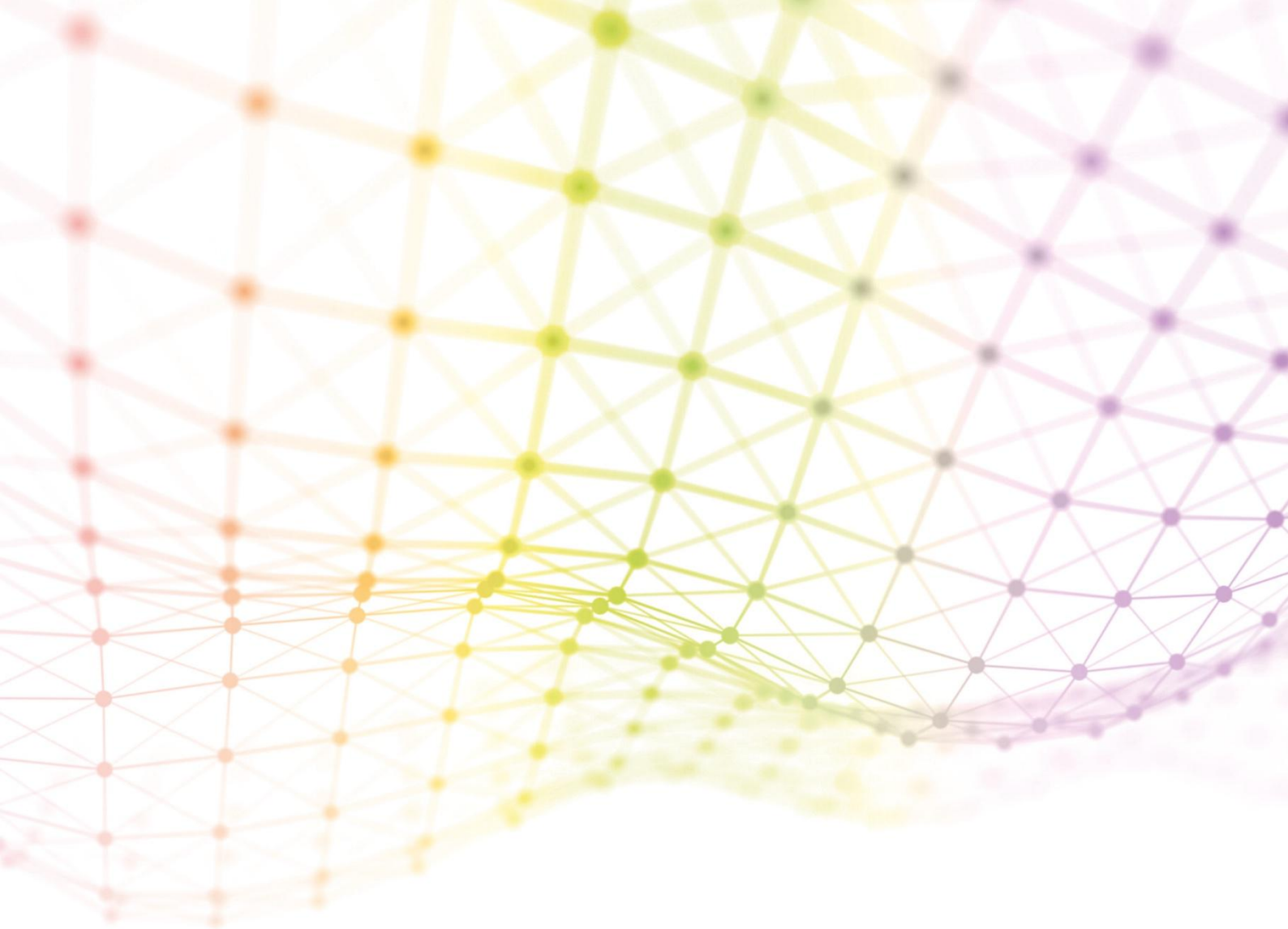
PRISMA-S: An Extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews

Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, Koffel JB, PRISMA-S Group.
Last updated February 27, 2020.

Supplementary Table 7. Types of mode effect comparisons examined across all studies, including the number of studies and the total number of survey items relating to each mode comparison.

Mode category comparison	N studies	N items
Paper vs Web	16	711
Paper (self-administered) vs Web	9	559
Paper (mailed) vs Web	4	41
Paper (self-administered) vs Computer	3	111
Face-to-face vs Paper	16	435
Face-to-face vs Paper (self-administered)	8	214
Face-to-face vs Paper (mailed)	7	139
Face-to-face vs Paper (self-administered, with audio)	1	82
Face-to-face vs Telephone	15	384
Face-to-face vs Telephone	9	297
CAPI vs CATI	4	71
Face-to-face vs CATI	1	10
CAPI vs IVR	1	6
Paper vs Telephone	13	367
Paper (mailed) vs CATI	5	105
Telephone vs Paper (mailed)	4	221
Telephone vs Paper (self-administered)	4	41
Face-to-face vs Web	11	168
Face-to-face vs Web	6	108
CAPI vs Web	4	55
PAPI vs Web	1	5
Telephone vs Web	9	240
Telephone vs Web	4	137
CATI vs Web	4	96
IVR vs Web	1	7
Face-to-face vs (A)CASI	9	194
Face-to-face vs ACASI	4	122
CAPI vs ACASI	3	36
CAPI vs CASI	1	26
Face-to-face vs CASI	1	10
Telephone vs Telephone	8	192
Telephone vs T-ACASI	3	126
CATI vs T-ACASI	3	26
CATI vs IVR	2	40
(A)CASI vs Paper	10	564
CASI vs Paper (self-administered)	5	251
ACASI vs Paper (self-administered)	4	265
ACASI vs Paper (self-administered, with audio)	1	48
Mobile vs Paper	3	278
Tablet vs Paper (self-administered)	2	268
Notebook vs Paper (self-administered)	1	10
Face-to-face vs Face-to-face	4	32
CAPI vs PAPI	2	20
CAPI (no consistency checks) vs CAPI	1	6
CAPI (no consistency checks) vs PAPI	1	6
Face-to-face vs Other	3	271
Face-to-face vs Ballot Box	1	189
Face-to-face vs Randomised response	1	54
CAPI vs SAQ (Paper or Web)	1	28
Mobile vs Web	2	29
Mobile vs Computer	1	23
Mobile vs Web	1	6
Paper vs Paper	2	98
Paper (self-administered) vs Paper (self-administered, with audio)	1	72
Paper (self-administered) vs Paper (mailed)	1	26
(A)CASI vs (A)CASI	1	9
ACASI vs CASI	1	9
(A)CASI vs Telephone	2	12
ACASI vs CATI	1	6
ACASI vs IVR	1	6
(A)CASI vs Web	1	29
CASI vs Web	1	29

Mobile vs Other	1	23
Mobile vs Hybrid	1	23
Other vs Other	1	54
Ballot box vs Randomised response	1	54
Web vs Other	1	23
Computer vs Hybrid	1	23



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