Surveys in the Age of AI: A Scoping Review from Population Frame to Statistical Inference

Muhammad Nadim Hanif*

Abstract

Artificial intelligence (AI) can change both the production of survey data and the behaviour that surveys seek to measure. This scoping review maps AI applications across 16 stages of the survey lifecycle, from specifying objectives and constructs to dissemination, evaluation and redesign, and examines their implications for stage-specific errors and total survey error (TSE). A scoping review approach was used to synthesize heterogeneous evidence from survey methodology, statistics, economics, marketing, computer science and official statistics. Eligible literature examined AI applications at one or more lifecycle stages, their effects on survey-error components, or the benefits, limitations and trade-offs of human–AI survey practices. The evidence shows applications of AI and related computational methods in frame construction, sample optimization, questionnaire assessment, interviewing, fieldwork monitoring, coding, imputation, weighting, estimation, inference, validation and disclosure-protected dissemination. Reported gains include improved prediction, classification, precision, response quality, error detection and operational efficiency. However, the evidence remains predominantly stage-specific; within the literature identified for this review we did not locate any explicit reduction in TSE across the complete lifecycle, and AI agentic capabilities remain largely unevaluated. AI may reduce, transfer, conceal or create representation, measurement, processing and inferential errors, particularly when synthetic respondents or opaque systems are treated as substitutes for human observations or accountable methodological judgement. There is need for a review that systematically maps AI applications across the complete survey lifecycle with implications for different stages of survey life cycle. This review contributes a lifecycle-based synthesis and proposes a human-supervised, risk-proportionate approach in which AI performs bounded, auditable tasks while human researchers retain responsibility for consequential design and inferential decisions. The proposal specifies adaptable responsibilities and approval arrangements.

JEL classifications: C81, C83, C18, D83, D84, O33

Keywords: Survey Methodology; Artificial Intelligence; Perception; Expectation; Measurement Error; Synthetic Respondents; Household Survey; Firm Survey.

1. Introduction

A survey is a structured process for collecting standardized information from defined population units through a census or sample, then processing and analysing it to produce knowledge about the target population (Groves et al., 2009). Surveys make otherwise difficult-to-observe economic objects measurable. Administrative records reveal earnings, transactions and employment, but generally not the beliefs, intentions, perceptions and expectations underlying them. Households may postpone purchases because they fear unemployment, expect lower prices or face credit constraints; similar behaviour can therefore require different policy responses. Manski (2004) establishes the case for directly eliciting subjective expectations, while Stantcheva (2023) explains how carefully designed surveys reveal mechanisms and create identifying variation unavailable in ordinary observational data.

The credibility of survey findings depends on the coherence of the entire process. Objectives and constructs establish what is measured; population and frame definitions determine who can be represented; and sampling governs selection and precision. Questionnaires translate

* Member of the Board of Advanced Studies and Research, Institute of Business Management, Karachi. AI tools assisted with literature searching, language editing, and cross-checking in-text citations against the reference list reviewed. Email: muhammadnadeemhanif@yahoo.com.

concepts into observable responses, while consent, pretesting, piloting, training and field protocols support appropriate and consistent collection. Processing, weighting, estimation and analysis connect those responses to population quantities and substantive conclusions. Validation, dissemination and evaluation assess credibility, communicate uncertainty and inform subsequent rounds. Weakness at any stage can compromise the contribution of otherwise sound procedures elsewhere in the lifecycle. Ethical safeguards protect participants and support truthful disclosure; pretesting examines comprehension, while piloting tests procedures under realistic conditions. Training and monitoring help maintain the intended protocol. Dissemination must preserve the qualifications needed for appropriate interpretation, and evaluation should prevent weaknesses from recurring in later rounds.

These connections explain why survey quality cannot be inferred from sample size, response rates or questionnaire design alone. A strong questionnaire cannot correct an incomplete frame, nor can accurate fieldwork compensate for inappropriate weighting. Representation errors arise through population definition, coverage, selection, nonresponse and adjustment; measurement errors arise when questions, response processes, interviewing or processing fail to capture intended constructs. Together, their effects on the bias and variance of estimates are considered within the total survey error (TSE) framework. Errors may amplify, conceal or partly offset one another, requiring assessment across the lifecycle and against intended uses (Groves et al., 2009; Biemer, 2010; Groves & Lyberg, 2010).

Artificial intelligence (AI) now enters this process through several routes. Statistical learning supports frame construction, response prediction, coding and imputation; conversational agents administer questions and probe narrative answers; and large language models assist with instrument development, translation, pretesting, text analysis and statistical narratives. Buskirk and Kirchner (2020) and Kern et al. (2019) discuss applications across multiple survey operations. Researchers use AI to design and operate surveys, while respondents may use it to interpret questions, retrieve information or formulate answers. AI also changes the household and firm behaviour that surveys seek to measure, making it both a methodological tool and an object of measurement.

The relevant question is whether these applications improve the connection between the target population, intended constructs and resulting inference. Faster processing, better prediction or more fluent responses need not reduce survey error. AI outputs require documentation, task-specific validation and methodological judgement; apparent gains may otherwise conceal limitations or shift errors between stages. Grimmer and Stewart (2013) emphasize validation in automated text analysis, while Ziems et al. (2024) examine the opportunities and limitations of language models in computational social science. Their cautions are relevant to assessing AI-assisted survey procedures without equating computational performance with substantive validity. Heavy reliance on AI can create errors alongside operational gains. Synthetic respondents generate answers from learned textual patterns rather than directly reporting the contemporaneous experiences, incentives or beliefs of the target population. A large synthetic sample may therefore estimate a model's response distribution precisely while misrepresenting people. Its use for pretesting or sensitivity analysis must be distinguished from replacing human observations. Adaptive interviewing raises a related concern: explanations and probes may improve comprehension while changing question interpretation or influencing answers. Greater disclosure,

detail or consistency does not establish validity, because mode, perceived privacy, social presence and sensitivity affect reporting (Kreuter et al., 2008; Lucas et al., 2014).

Other risks extend across stages. Unequal digital representation may disadvantage populations with limited connectivity, literacy or minority-language presence. Opaque models and undocumented updates can undermine reproducibility and comparability between survey waves, while dependence on external providers introduces uncertainties about access and data handling. Transmitting sensitive responses to external systems can also compromise confidentiality. These risks require attention to how AI enters the survey process, whose information it represents and whether its outputs remain appropriate for the intended population and use. Efficiency cannot compensate for measuring an unsuitable construct or systematically excluding relevant population groups.

A hybrid human–AI approach should therefore establish the population, reference period, construct and elicitation protocol before selecting tools. Humans should retain authority over objectives, eligibility, consent, sampling, consequential classifications, inferential assumptions and publication. AI can undertake bounded, auditable tasks, including identifying frame deficiencies, reviewing questions, predicting contact difficulty, transcribing interviews, proposing codes, detecting anomalies and drafting summaries. Outputs should be validated and consequential decisions approved by responsible researchers. The proposed allocation is risk-proportionate: AI may assist throughout, but methodological responsibility remains human. Agentic workflows should be assessed against these requirements rather than presumed capable of coordinating the lifecycle reliably.

This study reviews AI applications across the survey lifecycle, focusing on documented contributions to survey processes, stage-specific errors and their possible implications for TSE. The reviewed literature reports gains in sample optimization, disclosure protection, fabrication detection, conversational collection, coding, imputation, weighting, estimation and targeted validation. These findings require careful interpretation: meeting precision constraints with fewer observations demonstrates design efficiency, for example, rather than necessarily reducing sampling error. Within the literature identified, no study explicitly demonstrated reduced TSE across the complete lifecycle. Early work by Bollenbacher et al. (2026) nevertheless envisages agentic survey operations evaluated within a TSE framework, providing a direction for investigation rather than evidence of an achieved lifecycle-wide reduction.

The review integrates evidence dispersed across survey methodology, statistics, economics, marketing, computer science and official statistics within a common framework of sixteen stages. It distinguishes implemented applications and demonstrated benefits from prospective capabilities, methodological risks and areas with limited evidence. This organization allows operational improvements to be considered alongside their implications for representation, measurement and inference, without assuming that success in one task establishes overall survey quality. For each stage, the review considers what the procedure is intended to achieve, the consequences of inadequate treatment and the evidence concerning AI assistance. It thereby connects technical applications with the methodological purposes they must serve, rather than treating automation itself as a contribution to knowledge.

The study also links overreliance risks to the stages at which they arise and examines their consequences for TSE and statistical inference. Building on this synthesis, it proposes a human-supervised complementarity principle for hybrid survey practice: delegate routine, well-specified

tasks to AI while retaining human responsibility for complex and consequential judgements under risk-proportionate oversight. This is a methodological proposal requiring evaluation, rather than a validated claim that hybrid systems necessarily outperform alternative arrangements. The proposed approach aims to preserve validity, transparency, accountability and the representation of actual populations while making appropriate use of computational assistance in survey research. Its value depends on the quality of resulting evidence. An operational extension assigns oversight functions and approval points across organizational settings.

Section 2 reviews related literature, and Section 3 describes the methodology. Section 4 examines AI applications across all sixteen lifecycle stages, identifying their purposes, methodological significance, documented benefits and error implications. Section 5 considers overreliance risks; Section 6(a) develops the hybrid approach; Section 6(b) specifies associated operational arrangements. Section 7 identifies future research priorities, and Section 8 concludes.

2. Review of Literature

Previous reviews of survey research have generally concentrated on particular modes, methodological stages or sources of survey error. Couper (2000), for example, reviewed alternative forms of web surveying and examined their implications for sampling, coverage, participation and implementation. The review established that web surveys comprise distinct designs whose inferential properties depend partly on how respondents are recruited and selected. De Leeuw (2005) extended the mode-related literature by synthesizing evidence on mixed-mode surveys, particularly their implications for coverage, nonresponse, measurement and mode effects. For the questionnaire development stage, Beatty and Willis (2007) reviewed cognitive-interviewing practices and studied design choices. Collectively, these reviews provide important methodological foundations for web and mixed-mode data collection, questionnaire development and pretesting.

A broader strand of the literature is organized around TSE. Biemer (2010) examined the application of the TSE framework to survey design, implementation and evaluation, emphasizing the need to balance data quality and survey costs. Groves and Lyberg (2010) reviewed the historical development of TSE and its role in connecting errors of representation and measurement across survey design, data collection, processing and estimation. Amaya et al. (2020) subsequently adapted TSE into a more general total error framework for classifying errors in big-data sources. These contributions established the conceptual basis for evaluating quality across multiple stages, although they were not designed to review the use of AI throughout the survey lifecycle.

The emergence of machine learning produced reviews with more direct relevance to AI-assisted survey research. Kern et al. (2019) discussed tree-based machine-learning methods, reviewed their existing and prospective applications in survey management, and demonstrated their use in predicting nonresponse. Buskirk and Kirchner (2020) provided a wider survey-process overview covering machine-learning applications to sampling-frame construction, sample design, questionnaire development, recruitment, data collection, coding, processing, weighting, adjustment, analysis and estimation. Their synthesis demonstrates that machine-learning methods can support several interconnected stages of survey research.

More recent work has examined generative AI in relation to synthetic survey respondents. Sarstedt et al. (2024) reviewed studies comparing responses generated by large language models with those obtained from human participants and developed guidance for using such silicon samples in consumer and marketing research. The review identified particularly relevant applications in questionnaire pretesting and piloting. Sen et al. (2026) subsequently proposed the

total simulated survey error framework, which identifies measurement and representation errors arising across the production of LLM-generated survey responses.

This emerging body of research is anchored by institutional frameworks, most notably the American Association for Public Opinion Research (AAPOR) task force report on responsible AI integration (Rothschild et al., 2026). This report establishes consensus guidelines for navigating data governance, privacy budgets, and human-on-the-loop requirements across survey organizational workflows. By synthesizing industry standards, the AAPOR framework bridges the gap between theoretical machine learning capabilities and practical survey administration. It underscores that while automated workflows offer operational scaling, they must remain tethered to strictly auditable, human-supervised protocols to safeguard public data integrity and institutional trust.

A complementary literature exists on overreliance upon automated systems. Parasuraman and Riley (1997) distinguished the use and misuse of automation, establishing that the consequences of automation depend partly on how people rely on it. Goddard et al. (2012), in a systematic review with a healthcare focus, examined automation bias and insufficient monitoring of automated output, identifying factors such as trust, workload and task complexity, alongside possible mitigation through training, accountability and system design. These studies provide a conceptual basis for examining overreliance in surveys, but they neither studied automation or AI applications at different stages of the survey lifecycle nor estimated their impact on stage-specific errors or TSE.

Similarly, the allocation of work between humans and machines has established theoretical foundations. Parasuraman et al. (2000) distinguished automation of information acquisition, information analysis, decision and action selection, and action implementation. Their framework evaluates automation using human-performance consequences, system reliability and the costs of decision or action outcomes. Their contribution supports differentiated task allocation and attention to consequences. Broader evidence, however, cautions against assuming that human involvement necessarily improves machine performance. In a systematic review and meta-analysis of 106 experiments, Vaccaro et al. (2024) found that human–AI combinations outperformed humans alone on average. Survey research separately provides an empirical example of selective human–machine task allocation. Schonlau and Couper (2016) combined automated categorization of easier open-ended responses with human coding of the remaining responses, using expected accuracy to select the allocation threshold. This study demonstrates a hybrid workflow at the coding stage. It does not establish that the same allocation rule applies to other survey stages to minimize TSE. For survey practice, a human–AI hybrid approach should be tested against relevant alternatives, with an explicit focus on minimizing TSE and improving the quality of statistical inference.

Taken together, the literature provides reviews of survey modes, questionnaire evaluation, machine-learning applications across several survey operations, and LLM-generated respondents. Prior institutional reviews, such as Rothschild et al. (2026), offer critical operational guidelines, but there remains a need for an explicit, academic scoping review that systematically maps AI applications across the complete survey lifecycle while evaluating their implications for stage-specific errors and TSE.

Moreover, in the literature reviewed above, we find general research on automation bias and survey-specific evidence on risks in using machines at selected survey stages. No review was identified that connects overreliance risks across the survey lifecycle. This review also addresses

this gap by examining the implications of overreliance on AI for representation error, measurement error, TSE and statistical inference.

Similarly, the existing literature on task allocation between humans and machines provides precedents for hybrid practice but does not specify human responsibilities and oversight across the survey lifecycle. This study fills this gap by proposing a human–AI hybrid approach to surveys.

3. Methodology

This study adopts a scoping review approach[2] to examine the use of AI across the survey lifecycle. This approach accommodates heterogeneous evidence from survey methodology, economics, marketing, statistics, computer science, and official statistics, considering different publication types, including peer-reviewed journal articles, conference papers, working papers, books, and authoritative institutional reports.

Studies were identified with assistance from Chat GPT-6 during September 1-22, 2026, using combinations of relevant terms including AI, Artificial Intelligence, Surveys, Machine Learning, and various stages of survey lifecycle. Only those publications are included in which authors explored AI use(s)/application(s) at any stage of survey and/or reviewed such work.
Drawing on the survey-methodology literature, including Groves et al. (2009), the survey lifecycle establishes the review's conceptual scope and provides the framework for classifying the literature into sixteen stages (as detailed in section 4).

Studies were included when authors used an AI application directly related to at least one lifecycle stage, provided evidence concerning a stage-specific component of survey error or total survey error, or investigated the benefits, limitations, and trade-offs associated with human- and AI-supported survey practices. Reported improvements associated with AI applications included lower bias, variance, root mean squared error, classification loss, miscoding, disclosure exposure or undetected fabrication, and better response quality.

The selected literature was synthesized thematically. For each stage, the review identifies relevant AI applications, their demonstrated contribution to error reduction, associated methodological and ethical limitations, and possible safeguards. After highlighting the potential risks of overreliance on AI applications in surveys, at different stages of the survey lifecycle, this review culminates in proposing a human–AI hybrid framework intended to balance operational efficiency with relevance, validity, transparency, accountability, and human oversight.

4. AI applications at different stages of the survey lifecycle

Figure 1 illustrates the sixteen-stage survey lifecycle used to organize the review. It links the stages discussed in Section 4 and shows how evaluation informs subsequent design and survey rounds. Total survey error assessment and accountable human oversight apply across the lifecycle; the arrows provide an organizing sequence rather than a rigid chronology.

[2] For different review methods, see Whittemore & Knafl, (2005) and Snyder (2019).

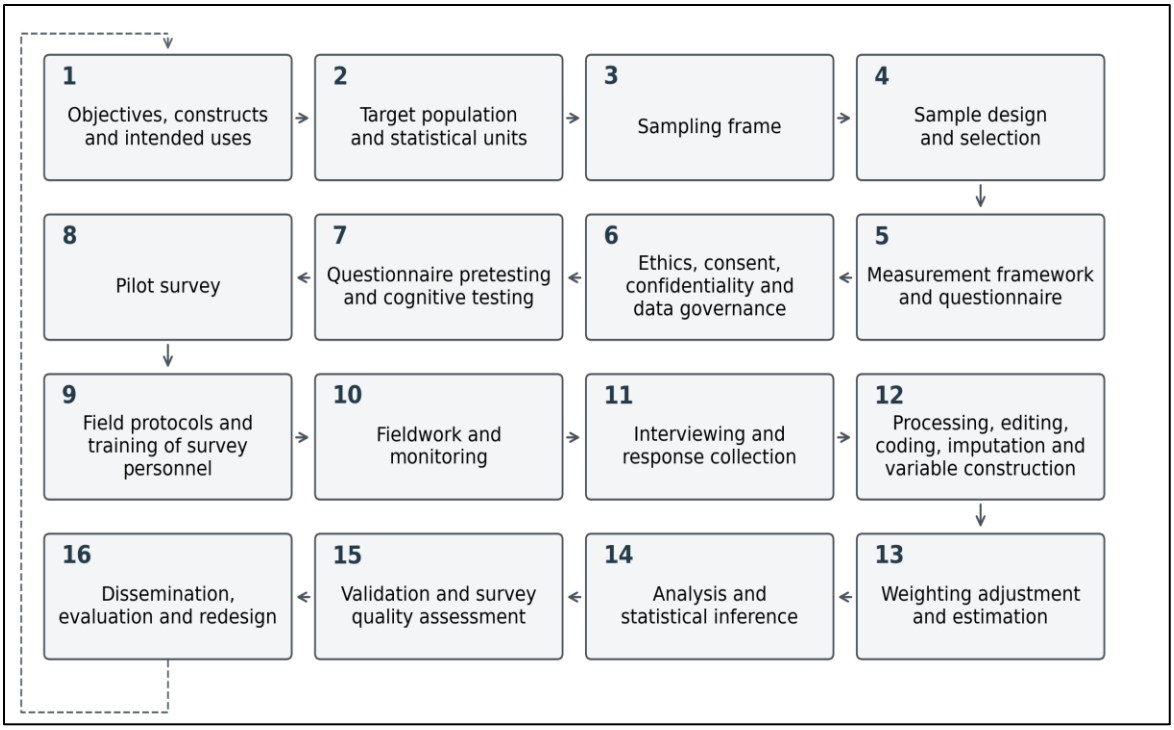


Figure 1. Sixteen-stage survey lifecycle used to organize the review

Source: Author's synthesis of the stages of a survey, drawing on Groves et al. (2009).

Survey conduct connects all stages, from the specification of objectives and constructs to post-survey evaluation. AI and related computational methods are now used across many of these stages through machine learning, natural-language processing, computer vision, speech technologies, optimization, generative models, and privacy-preserving computation. This literature is dispersed across social sciences, computer science and survey methodology. The following review organizes it around sixteen stages of the survey cycle, examining what these tools do, how they have been used in published research and statistical practice, and their implications for stage-specific errors and TSE.

4.1 Survey objectives, constructs, and intended uses

This stage defines the objectives, constructs, indicators, reference period, precision requirements and intended analyses. It aligns the survey with its research, policy, marketing or other stated legitimate use. It establishes the target concepts and estimands. Poor specification can yield precise answers to the wrong question (Groves et al., 2009).

Current evidence does not suggest that AI can make these decisions independently or reliably. Its more defensible role begins after the construct has been defined, when it can help researchers develop possible survey items. Hommel et al. (2022) developed a transformer-based method for generating items for predefined psychological constructs. Götz et al. (2024) also designed and evaluated a neural-language-model-based system that produced candidate items for established and newly proposed constructs. Terry et al. (2025) used ChatGPT-3.5 to generate items measuring grit and tested them in two independent student samples. Comparison with the Short Grit Scale supported its factor structure, reliability and validity in these samples; this does not establish broader superiority. In market research, Keane and McNaughton (2026) developed a human–AI workflow for an entrepreneurial-mindset scale. Their proof of concept shows how generative AI

can help expand an item pool, remove repetition, and suggest alternative wording, while researchers remain responsible for the important conceptual and methodological choices.

The evidence therefore supports a limited but useful role for AI in turning predefined constructs into candidate survey items. It does not support handing over decisions about survey objectives, construct definitions, estimands, or intended uses. Researchers must judge whether proposed items cover the construct adequately, avoid contamination and repetition, work appropriately across cultural and demographic groups, and meet accepted standards of reliability and validity. AI may make item generation faster and more varied, but whether it reduces specification error across survey settings is not yet established.

4.2 Target population and statistical units

The target population identifies the people, households, firms, establishments, transactions or areas about which evidence is required. It fixes the scope of inference and the relevant statistical unit. Clear eligibility rules connect observations to estimands. Any ambiguity can produce misclassification and conclusions about an ill-defined population (United Nations, 2005).

Stevens et al. (2015) used random forests with census, remotely sensed, and ancillary data to produce detailed population maps. Wardrop et al. (2018) extended this type of geospatial modelling to countries without recent census information, generating high-resolution estimates of where people were likely to live. Such models allow survey organizations to move beyond broad administrative totals and obtain a more detailed picture of population concentrations, settlements, and geographical domains.

AI-based matching and classification methods can also help determine which records represent the same statistical unit. Record linkage and entity resolution compare names, addresses, identification numbers, and other attributes across datasets to establish whether separate records refer to the same person, household, enterprise, establishment, or location (Christen, 2012). In business statistics, Kühnemann et al. (2020) used text from company websites, word embeddings, naïve Bayes, and support-vector machines to predict enterprises' main economic activities across 111 NACE classes. This provided a practical way of supporting the classification and updating of units held in statistical business registers used as sampling frames in some surveys.

These studies support population mapping and unit classification as inputs to survey design. Their outputs concern predicted locations or categories, rather than observed eligibility for a specified survey. Spatial detail or classification therefore cannot by itself establish that the target population has been defined more accurately.

4.3 Sampling frame

The sampling frame is the operational list or procedure used to identify and reach eligible units. It provides the basis for sample selection. Its coverage determines which units can be represented. Omissions, duplication, ineligible entries and outdated records can create coverage error and distorted estimates (United Nations, 2005).

AI is being used in three well-established areas of sampling-frame construction and maintenance: linking records, classifying units, and mapping populations. Probabilistic and machin

e-learning record-linkage methods compare information such as names, addresses, dates, and relationships to establish whether records held in different administrative systems refer to the same person, household, or business. This makes it easier to combine sources and remove duplicate records when developing household and business sampling frames (Christen, 2012).

Geospatial machine learning offers support for area-based frames. Stevens et al. (2015) combined random forests with remotely sensed data to produce fine-resolution population estimates from census counts. Wardrop et al. (2018) used geospatial data and statistical modelling to estimate population distributions where up-to-date census information was unavailable. The population surfaces produced by these methods can provide useful inputs when constructing or updating area frames in locations where conventional lists are incomplete.

These applications support frame development through linkage, classification and spatial prediction. However, a population surface is not a verified list of eligible units, and deduplication is not a coverage assessment. These studies should inform frame construction without being read as direct evidence of reduced survey under coverage.

4.4 Sample design and selection

This stage specifies the sampling method, strata, clusters, sample size, selection probabilities and within-unit selection. It produces a sample capable of supporting population estimates. The design governs precision, weighting and variance estimation. Selection bias or miscalculated uncertainty can weaken conclusions (Lohr, 2022).

An important use of AI-related methods in sample design is the optimization of stratification and sample allocation. Ballin and Barcaroli (2013) developed a genetic algorithm that compares alternative ways of combining elementary groups into sampling strata. It then searches for the design requiring the smallest sample while meeting predetermined precision requirements for several variables and population domains. This approach was later made available through the *SamplingStrata* R package, which allows survey statisticians to incorporate sampling-frame variables, anticipated variances, domain requirements and precision targets into the design process (Barcaroli, 2014).

O'Luing et al. (2019) refined this method by developing a grouping genetic algorithm tailored to the joint problem of forming strata and allocating the sample among them. Tests on publicly available datasets showed that it found better solutions than a conventional genetic algorithm using a similar amount of computational effort. O'Luing et al. (2022) later applied simulated annealing with delta evaluation to the same design problem. Their comparisons showed that the method could produce solutions of comparable quality in considerably less computation time.

Together, these studies show how computational search methods can examine far more possible stratifications and allocations than would normally be practical by hand. Their main contribution is to identify sample designs that satisfy stated precision requirements with relatively small total samples. AI-related optimization therefore helps translate information held on the sampling frame into a workable and efficient design. Survey statisticians remain responsible for setting the variables, domains, precision thresholds and sample-size requirements against which each proposed design is assessed.

In their application to the Italian Farm Structure Survey, Ballin and Barcaroli (2013) identified designs meeting prespecified multivariate precision constraints with smaller samples than the existing design. This demonstrates reduced sample requirements at a specified level of precision, rather than a direct reduction in sampling error.

O'Luing et al. (2019) improved design solutions relative to a conventional genetic algorithm under accuracy constraints. The gain concerns required sample size, not observed reductions in survey bias.

4.5 Measurement framework and questionnaire

The measurement framework translates constructs into questions, response options, recall periods, routing and derived variables. It elicits observable reports that correspond to the underlying concepts. Wording and response design shape validity and comparability. Misinterpretation, recall difficulty and response effects can generate measurement error (Tourangeau et al., 2000).

Adhikari et al. (2025) evaluated LLM-assisted questionnaire generation and adaptation with 238 US and 118 South African participants. Comparisons concerned generated versus LLM-pretested political-trust questions and original versus adapted climate-change questions. Ratings of clarity, specificity, relevance and perceived bias assess question wording in these settings; they do not establish reduced measurement error in completed population surveys.

AI can also make open-ended questionnaires more responsive. Xiao et al. (2020) compared an AI-powered chatbot with a conventional online survey in a field study involving approximately 600 participants. Both versions asked the same questions, but the chatbot could interpret answers, acknowledge what respondents had said and ask for further detail where appropriate. An examination of more than 5,200 responses found that participants using the chatbot were more engaged and provided answers that were more informative, relevant, specific and clearly expressed (Xiao et al., 2020).

The evidence supports candidate generation, contextual adaptation and fuller open-ended responses as distinct contributions. Wording ratings and response elaboration address particular features of measurement; neither establishes that answers more accurately capture the intended construct. Researchers must therefore evaluate proposed questions and probing rules against the survey's measurement requirements before adoption.

4.6 Ethics, consent, confidentiality and data governance

This stage establishes informed consent, confidentiality, secure data handling, access and retention arrangements. It protects both participants and researchers and defines legitimate data use. Clear protections support responsible collection and disclosure control. Breaches can expose respondents, violate professional duties and undermine participation and data integrity (ICC & ESOMAR, 2016).

The strongest published evidence at this stage concerns the use of computational methods to protect respondents when survey or census results are released. Differential privacy offers a formal way of limiting how much any one person's record can influence published statistics. It does this by introducing carefully controlled randomization and accounting for the protection provided through a privacy budget (Dwork & Roth, 2014). The US Census Bureau adopted this approach in developing the disclosure-avoidance system for the 2020 Census, showing that it could be incorporated into a major official-statistics programme (Abowd, 2018).

Synthetic data offer another means of protecting confidential information. Rather than releasing respondents' original records, researchers use statistical models to generate artificial records that retain selected relationships and distributions found in the source data. Drechsler (2011) explains how these datasets can give researchers access to useful information while reducing the direct disclosure of data supplied by identifiable respondents. Their usefulness and confidentiality protection can then be assessed against standards set by the statistical organization before they are released.

These methods make privacy protection part of the survey-data production process rather than a final administrative check. Differential privacy measures and records the privacy loss

associated with statistical releases, while synthetic-data methods provide an alternative to distributing original microdata (Abowd, 2018; Drechsler, 2011). The responsible institution decides the acceptable privacy budget, priority outputs, permitted uses and release conditions; computational systems help implement and document those decisions.

Reiter (2005) evaluated CART-generated partially synthetic microdata through simulations based on US Current Population Survey data, examining analytical validity and disclosure risk. The relevant comparison is between inferences from synthesized and original data, alongside confidentiality protection. This supports assessing a release procedure against specified analyses and disclosure scenarios, with applicability assessed for each intended release.

4.7 Questionnaire pretesting and cognitive testing

Pretesting examines how respondents understand questions, retrieve information, form judgements and select answers. It identifies ambiguous wording, excessive burden and unsuitable response categories before fieldwork. Cognitive evidence links question design to the response process. Undetected comprehension and retrieval problems can result in systematic measurement error (Willis, 2005).

Computational language tools can support questionnaire pretesting by identifying possible comprehension problems and suggesting revisions before fieldwork begins. An early example is the Question Understanding Aid (QUAID), developed by Graesser et al. (2006). This natural-language-processing system examines draft questions for problems such as unfamiliar terminology, vague or ambiguous expressions, complicated syntax and excessive demands on working memory. Its performance was investigated through comparisons with expert judgements, question-revision exercises and an eye-tracking study of how respondents processed survey questions. These evaluations assess diagnosis and question revision against expert and response-process evidence. They support preliminary wording assessment, rather than directly demonstrating reduced error in population estimates.

Adhikari et al. (2025) compared LLM-generated and LLM-pretested political-trust questions, and original and culturally adapted climate-change questions. Evaluations involved 238 US and 118 South African participants, with expert assessment of the first comparison. The outcomes were judgements of clarity, relevance, specificity and perceived bias. Their relevance is to preliminary instrument assessment within the studied topics and recruited groups. These ratings do not independently verify respondents' comprehension or the accuracy of their eventual answers; human cognitive testing remains a separate source of evidence.

4.8 Pilot survey

A pilot runs the survey system on a limited scale under realistic conditions. It tests recruitment, questionnaire flow, interview duration, field logistics, supervision, transmission and preliminary estimation. It assesses operational feasibility across connected procedures. If unresolved design and implementation faults enter the main survey, they become costly to correct (United Nations, 2005).

The published evidence for AI at the pilot stage pertains to trials of conversational survey interfaces. These trials allow survey teams to compare AI-assisted administration with a conventional questionnaire before deciding whether and how the technology should be used more widely. Kim et al. (2019) experimentally compared chatbot and standard web-survey formats. Their analysis showed how the choice of interface and conversational style could affect engagement, satisficing and response quality, illustrating the value of testing these features during a pilot.

Xiao et al. (2020) compared a probing chatbot with a conventional Qualtrics survey among approximately 600 participants. Analysis of more than 5,200 responses favored the chatbot on engagement and rated informativeness, relevance, specificity and clarity. This provides evidence for choosing and refining a conversational interface. However, it evaluates response characteristics within a particular collection task, rather than an integrated pilot of recruitment, sampling, logistics and estimation. Interface testing should therefore inform a pilot without being equated with validation of the complete survey system.

4.9 Field protocols and training of survey personnel

Protocols and training standardize respondent selection, consent, questioning, probing, coding, callbacks and quality checks. They prepare staff to apply the design consistently. Standardization limits interviewer-related variation across cases and locations. Divergent practices can introduce interviewer effects, recording errors and inconsistent treatment of respondents (Fowler & Mangione, 1990).

AI could support training by matching staff to assignments, personalizing instruction, rehearsing interviews, assessing competence, and providing just-in-time assistance. LLMs could transform manuals and specifications into examples, quizzes, multilingual explanations, and role-play scenarios, while retrieval-augmented generation could ground answers in the approved protocol. Birnbaum et al. (2012) developed an automated quality-control system for surveys conducted on mobile devices. By analysing data accumulated by individual interviewers, the system could identify recurring errors and alert supervisors to interviews requiring closer examination. This gave survey managers a practical way to recognize interviewer-specific problems while fieldwork was still under way.

Cohen and Warner (2021) examined quality control in the 2016–2017 AmericasBarometer. Their classification task compared screening based on a small set of automated and human-coded indicators with the final interview sample retained after a fuller quality-control process. The reduced set recovered much of that screening, supporting a more economical allocation of review effort. The benchmark was the survey's existing screening outcome, rather than independently known truth for every interview. The findings therefore support supervisory prioritization in this setting; they do not directly demonstrate the effectiveness of AI-generated training or a reduction in interviewer error following corrective instruction.

4.10 Fieldwork and monitoring

Fieldwork implements contact, recruitment, callbacks, supervision and case verification while monitoring progress and performance. It secures observations from the selected units according to protocol. Timely indicators reveal nonresponse, interviewer effects, fabrication and protocol deviations. Coverage, nonresponse and measurement errors can accumulate unnoticed (Biemer & Lyberg, 2003).

Kern et al. (2023) evaluated future-wave nonresponse prediction in the German probability-based GESIS Panel using participation histories and temporal cross-validation. Aggregating information across waves produced competitive, robust prediction over the test waves. This evaluates predictive performance under longitudinal validation, rather than the effects of an implemented retention intervention. Better identification of likely nonrespondents does not itself demonstrate increased participation or reduced nonresponse bias.

Machine learning has also been used to monitor adherence to fieldwork procedures. Cohen and Warner (2021) analyzed the 2016–2017 AmericasBarometer using information from audio

recordings, geolocation, interview duration and questionnaire paradata. They found that a relatively small set of automated and human-coded indicators reproduced much of the interview screening achieved by a more extensive quality-control process.

Schwanhäuser et al. (2022) compared detection methods using a German refugee survey with verified falsification. Their assessment included false-positive and false-negative rates, accuracy and chance-adjusted agreement. Known cases provide a benchmark for identifying suspicious interviewers in this dataset. The results concern detection, rather than the effect of subsequent supervisory action on population estimates.

Birnbaum et al. (2013) examined behavioral traces recorded by mobile devices as indicators of interviewer fabrication. The application concerns identifying suspicious interviews from recorded behaviour. Detection performance should be distinguished from prevention or correction: identifying cases for investigation does not itself establish how much fabrication remains in released data or how resulting survey estimates change.

4.11 Interviewing and response collection

Interviewing and response collection administer questions and record answers through the selected mode(s). They obtain complete, accurate and consistently elicited reports from sampled units. Mode, interviewer behaviour and respondent interaction affect the measurement process. Leading probes, recording mistakes and incomplete responses can reduce comparability and increase measurement error (Groves et al., 2009).

Research into AI-assisted interviewing has focused mainly on conversational agents that administer survey questions and collect respondents' answers. In an experimental comparison, Kim et al. (2019) found that chatbot surveys produced more differentiated responses and reduced some signs of satisficing compared with conventional web questionnaires. Celino and Re Calegari (2020) also tested a conversational survey interface, reporting comparable scale reliability, stronger response-quality indicators and greater user acceptance than for a conventional online questionnaire.

Xiao et al. (2020) compared a probing chatbot with a conventional Qualtrics survey among approximately 600 participants. More than 5,200 responses were assessed for informativeness, relevance, specificity and clarity, with greater engagement and better ratings for the chatbot. These are task-specific response-quality outcomes; richer answers alone do not establish more accurate reporting of the intended construct or improved population inference.

The published evidence therefore shows that conversational AI can administer predefined questionnaires, manage bounded conversational exchanges and encourage more detailed open-ended responses. These applications do not give AI control over the substance of the interview. Instead, they extend interviewer-like support to self-administered surveys while researchers continue to determine the questions, approve the available probes and assess the validity of the resulting data.

4.12 Processing, editing, coding, imputation and variable construction

Processing converts collected responses into documented analytical variables through editing, coding, classification, linkage, imputation and derivation. It produces a coherent dataset for estimation and analysis. Reproducible rules preserve meaning across records and waves. Coding, imputation or construction errors can alter distributions, relationships and substantive findings (Biemer & Lyberg, 2003).

Evidence for using AI in survey-data processing relates to the coding of open-ended answers and the imputation of missing values. Schonlau and Couper (2016) developed a semi-automated method that assigns codes when the model is sufficiently confident and sends uncertain responses to human coders. This makes it possible to convert large volumes of narrative material into analysable variables while reserving expert attention for the complex cases. Gweon et al. (2017) applied statistical-learning methods to respondents' job descriptions and showed how they could be used to assign detailed occupational codes.

He and Schonlau (2020) compared single- and double-coded training data under different coding budgets. Their findings concern how training-label quality and expert resolution of disagreements affect classification accuracy under the examined budget conditions.

Dagdoug et al. (2023a) compared nonparametric and machine-learning imputation procedures in varied settings, including high-dimensional data. Outcomes were the bias and efficiency of estimates derived after imputation, rather than prediction accuracy alone. Several procedures performed well, supporting evaluation of imputation through its consequences for the study parameters.

Targeting coding error is one use of AI-augmented surveys. Gweon et al. (2017) evaluated three statistical-learning procedures for occupational coding using German General Social Survey data. Their hierarchical and hybrid methods improved coding accuracy over the underlying statistical-learning algorithm and the benchmark approaches examined. The improvement directly reduces misclassification when narrative occupational descriptions are converted into statistical categories.

Gweon and Schonlau (2024) compared fine-tuned BERT with support-vector machines and boosting on two open-ended survey datasets. Once 200–400 manually coded training observations were available, BERT's classification advantage increased, yielding higher coding accuracy than the non-pre-trained alternatives. Fine-tuning and labelled training size therefore condition the observed accuracy gain; speed was not the reported basis for superiority.

He and Schonlau (2022) addressed errors already present in manually coded open-ended responses. A model trained on a randomly double-coded subset ranked the remaining single-coded answers by their probability of being wrong. Across three datasets, targeted recoding identified, on average, two to three times as many errors as selecting answers at random. The gain concerns errors identified through additional double coding; correction still requires adjudication rather than automatic acceptance of predicted errors.

These imputation results concern the examined data settings, not a universal ranking of procedures.

4.13 Weighting adjustment and estimation

This stage calculates design weights, applies nonresponse and calibration adjustments, and estimates population quantities and their variances. It converts sample data into population estimates. Weights reflect selection probabilities and justified adjustments, while design-based variance estimation quantifies precision. Improper weighting or unjustified adjustments can produce biased estimates and invalid standard errors (Valliant et al., 2018).

Kern et al. (2021) combined machine-learning propensity estimation with kernel weighting for non-probability samples. They compared these pseudo-weights with kernel weights based on logistic regression, using simulations and a real-data example to assess population-mean estimation. Boosting yielded lower bias in several settings without increasing variance. The

relevant condition is whether the propensity model captures the selection process; flexibility alone cannot establish that adjustment removes selection bias.

Machine learning can also strengthen estimation from probability samples. Breidt and Opsomer (2017) showed how flexible nonparametric methods, including regression trees and local polynomial regression, can be incorporated into generalized-difference estimators. This allows survey researchers to exploit nonlinear patterns in auxiliary data while retaining a design-based basis for inference.

Dagdoug et al. (2023b) developed this idea further by constructing random-forest model-assisted estimators for finite-population totals and means. They derived appropriate variance estimators and introduced a model-calibration procedure covering several survey variables. Their simulations produced favourable results for bias, efficiency and confidence-interval coverage under the conditions studied, and they illustrated the method using radio-audience survey data.

The evidence therefore shows that machine learning can improve the way auxiliary information is used in weighting and estimation. Its role is to provide flexible predictions within established statistical frameworks, while survey statisticians continue to preserve the relevant design information, choose suitable auxiliary variables and evaluate the balance and precision of the estimates.

The gains reported by Kern et al. (2021) included lower bias and root mean squared error than commonly used propensity-score weighting procedures, addressing adjustment error directly.

4.14 Analysis and statistical inference

Analysis evaluates patterns, differences, associations and models; inference states what can be generalized beyond the observed sample. It answers the substantive (economic or marketing) research questions. Valid inference incorporates the sample design and relevant uncertainty. Naive analyses can misstate standard errors, significance and underlying relationships (Lumley, 2010).

Machine learning contributes to survey analysis in two well-established areas: extracting information from open-ended answers and estimating complex statistical (or causal) relationships. Roberts et al. (2014) developed the structural topic model expressly for analysing open-ended survey responses. It identifies recurring themes while allowing their prevalence and content to vary with respondent characteristics or experimental assignments. Using survey and experimental data, the authors showed how researchers could relate the themes expressed in respondents' own words to their personal attributes and exposure to different treatments.

Machine learning also offers useful methods when survey data contain outcomes, treatments and a large number of possible explanatory variables. Chernozhukov et al. (2018) developed double/debiased machine learning, which combines flexible prediction with cross-fitting and orthogonal scores. Under appropriate identification and regularity conditions, this approach can estimate treatment or structural parameters and provide valid measures of uncertainty. Wager and Athey (2018) developed causal forests, enabling researchers to examine whether treatment effects differ systematically across respondents with different observed characteristics.

The published evidence therefore shows that machine learning can enrich survey analysis by turning narrative responses into quantitative measures, capturing nonlinear relationships and identifying variation in treatment effects.

4.15 Validation and survey quality assessment

Validation compares survey results with administrative records, external benchmarks, earlier rounds, paradata and internal consistency checks. It evaluates fitness for use. A total-survey-

error perspective integrates coverage, sampling, nonresponse, measurement, processing and model errors. If important defects remain undetected, reported quality may be overstated (Biemer, 2010).

AI and related computational methods have been applied to two main areas of survey validation: identifying irregularities in fieldwork and assessing the quality of synthetic survey responses. Using a nationally representative survey containing confirmed cases of interviewer falsification, Schwanhäuser et al. (2022) compared several statistical detection methods, including multivariate and cluster-based techniques. Their study shows how information from several quality indicators can be brought together to identify interviews and interviewers that warrant closer investigation.

A separate body of research uses human survey data as a benchmark for evaluating responses generated by language models. Argyle et al. (2023) conditioned a language model on sociodemographic profiles and compared the resulting responses with patterns found in United States public-opinion surveys. Santurkar et al. (2023) developed OpinionQA from established opinion polls and used it to assess how closely language-model responses reflected the views of 60 demographic groups. Bisbee et al. (2024) conducted a broader evaluation that considered response distributions, relationships between variables, subgroup differences and experimental treatment effects.

The validation tasks differ: confirmed falsification provides a benchmark for detection, whereas human-response patterns benchmark synthetic outputs. Agreement with either benchmark is evidence about the assessed task, not validation of the survey lifecycle. Researchers must still investigate flagged cases and judge whether the criteria support the intended use.

He and Schonlau (2022) found two to three times as many coding errors through model-targeted additional double coding as through random selection, across three datasets. This improves review yield within budget; it does not measure the remaining dataset-wide error.

4.16 Dissemination, evaluation and redesign

This stage releases findings and methodological documentation, evaluates performance and feeds lessons into subsequent rounds. It makes results interpretable, reproducible and usable. Metadata, uncertainty measures and revision records support appropriate analysis and comparison. Poor dissemination can lead users to misread estimates, hinder replication and allow recurring weaknesses to persist (United Nations, 2019).

Reiter (2005) used CART to generate partially synthetic public-use microdata, retaining sampled units while replacing selected values. Evaluation concerned analytical validity and disclosure risk in simulations based on US Current Population Survey data. The contribution is therefore a tested release method whose usefulness depends on the analyses retained and the disclosure scenarios examined, rather than evidence that synthetic files preserve every relationship or eliminate disclosure risk.

Sakshaug and Raghunathan (2014) extended this work to restricted data from the National Health Interview Survey. They used a hierarchical Bayesian model to generate synthetic microdata for small geographical areas, taking account of design features such as stratification, clustering and survey weights. Analytical validity was assessed against actual data, through simulation and cross-validation; these comparisons address small-area inference under the specified synthesis model.

These studies make analytical usefulness an object of evaluation against source data, rather than an assumed consequence of synthesis. Their contribution is to disclosure-protected access and the inferences assessed. Statistical agencies must still decide whether a proposed release meets its

intended analytical uses and confidentiality requirements; neither synthesis nor successful model fitting establishes suitability for every analysis.

Taken together, the studies support distinct stage-specific applications, evaluated against different comparators and outcomes. They do not jointly demonstrate that linking these applications reduces lifecycle-wide TSE. Integration remains a proposal requiring evaluation across the transitions between stages. Recently, Bollenbacher et al. (2026) presented an early-stage agentic-AI development programme for survey operations, envisaging evaluation against the TSE framework, including measurement, sampling, coverage, nonresponse, and processing errors, while retaining humans in the loop. In the next section, we explore the risks of insufficient human oversight.

5. Potential Risks of Overreliance on AI Across the Survey Lifecycle

AI may introduce representation, measurement, processing, estimation and inferential errors when outputs replace observations or methodological judgement. Evidential support varies: survey comparisons document selected risks, broader automation research suggests mechanisms, and other concerns remain conceptual. These distinctions matter when assessing consequences for particular stages and lifecycle-wide TSE.

Synthetic respondents illustrate the distinction between numerical precision and population validity. Generated answers reflect learned patterns rather than observations of the target population's current experiences, incentives or beliefs. Increasing their number may estimate a model's response distribution more precisely without improving population validity. Santurkar et al. (2023) compared language-model opinion distributions with US survey responses across demographic groups, finding persistent misalignment despite demographic prompting. Bisbee et al. (2024) prompted ChatGPT 3.5 Turbo with demographic and political personas and compared generated feeling-thermometer responses with American National Election Studies data. Similar aggregate means concealed lower variation and differences in conditional relationships; synthetic and human data could therefore support different substantive inferences. These comparisons demonstrate limitations in the tested models, prompts and opinion measures, rather than a universal ranking of synthetic approaches. They do not establish that every human-survey estimate is correct. Synthetic respondents may inform pretesting, scenario exploration and sensitivity analysis, but agreement on selected benchmarks does not validate replacement of new observations of current human beliefs, expectations or behaviour.

Adaptive AI interviewing raises a possible risk of intervention-induced measurement, but conversational flexibility need not itself reduce accuracy. Schober and Conrad (1997) experimentally compared standardized and flexible human interviewing using fictional situations with known answers. Clarification improved accuracy when respondents' circumstances fitted question concepts poorly, while increasing interview duration. This supports clarification under the tested conditions; it does not demonstrate harmful effects of generative AI probing. In a chatbot–Qualtrics comparison, Xiao et al. (2020) found greater engagement and more informative, relevant, specific and clear open-ended responses. These outcomes establish improvements in response characteristics, rather than independent verification of respondents' underlying beliefs or factual accuracy. Variable AI explanations might nevertheless introduce information or change interpretation, making the interaction part of what is measured. This remains a mechanism requiring direct evaluation against approved elicitation protocols. Fluency, detail, engagement and internal

consistency alone cannot establish measurement validity or comparable interpretation across respondents, population groups and survey waves.

Automation bias has empirical support outside survey research. Skitka et al. (1999) compared performance on a simulated flight task with and without an automated monitoring and decision aid. Aided participants missed events that the system failed to flag and followed erroneous recommendations despite contradictory indicators; unaided participants performed better on monitoring. This concerns conventional automation, not generative AI or survey production, so it does not quantify survey-error consequences. Accepting plausible but unsuitable questionnaires, codes or interpretations is a corresponding survey risk requiring investigation. Bainbridge's (1983) analysis of industrial automation also explains why reduced practice may weaken skills needed for intervention. This is a conceptual rationale for maintaining survey expertise, rather than evidence of measured skill loss among survey researchers.

Representation risks are supported most directly by comparisons of model-generated opinions with human responses. Santurkar et al. (2023) evaluated nine language models against OpinionQA, using US opinion polls covering 60 demographic groups. Alignment differed across groups, and demographic steering did not consistently remove misalignment. The English-language, US-centered benchmark limits conclusions about other populations and languages; it also does not isolate training-data imbalance as the cause. The findings support subgroup validation of synthetic responses, rather than directly demonstrating under coverage in AI-built sampling frames. Reliance on digital traces could nevertheless omit people with limited connectivity or literacy, minority-language users and informal enterprises. These are plausible coverage mechanisms requiring comparison with independently assessed frames. Likewise, differential coding or classification errors require labelled subgroup benchmarks. Aggregate performance alone cannot establish that representation and measurement quality are comparable across the target population.

Reproducibility risks have direct support in synthetic-response research. Bisbee et al. (2024) repeated identical prompts in April, June and July 2023 and found changes in generated responses, including altered relationships after a model update. These comparisons concern the tested ChatGPT versions and prompts, not measured discontinuities in an operational survey series. Nevertheless, undocumented changes could confound population change with changes in measurement. Restricted access to training information, model versions or decision processes can impede diagnosis. Dependence on providers also creates possible exposure to altered availability, access terms, retention policies and prices. Correlated errors across linked AI components remain a further methodological concern, rather than a demonstrated lifecycle-wide effect; shared model assumptions could propagate mistakes from frame construction through processing, estimation and dissemination.

Privacy risks require separating empirical disclosure assessments from formal guarantees. Reiter (2005) evaluated CART-generated partially synthetic microdata using Current Population Survey-based simulations: many inferences were preserved, but some relationships were inadequately captured. This evaluates release utility and disclosure risk, not breaches involving external generative AI services. Differential privacy bounds disclosure from released outputs under its assumptions, with accuracy consequences (Dwork & Roth, 2014); it does not establish secure collection, storage or authorised secondary use. External processing of identifiable responses or linked records therefore requires governance safeguards. These are confidentiality concerns, rather than documented breach rates or measured TSE effects in AI-assisted surveys.

Finally, accountability is a methodological requirement rather than an empirically estimated effect of automation. Survey interpretation depends on judgements about constructs, coverage, missingness, weighting, uncertainty, causal identification and limitations. Automatically generated interpretations could conflate prediction with explanation or association with causation, but the studies discussed here do not quantify how frequently this occurs in survey reporting. Nor do they establish that greater automation necessarily removes accountability. The concern arises when responsibilities for checking assumptions, approving outputs and correcting errors are undefined. Human authority should therefore remain explicit for consequential decisions affecting representation, measurement, inference and publication. These safeguards are recommendations requiring evaluation, not demonstrated reductions in errors across the complete survey lifecycle.

6(a). A Cautious Human–AI Hybrid Approach to Surveys

A defensible human–AI survey system should specify the target population, reference period, construct, estimand and permissible elicitation protocol before selecting an AI tool. Humans should retain authority over construct definition, eligibility, informed consent, sample design, consequential classifications, inferential assumptions and dissemination. AI should undertake bounded and auditable functions, such as retrieving earlier instruments, identifying potential deficiencies in sampling frames, flagging problematic questions, predicting contact difficulty, transcribing interviews, proposing codes, detecting anomalies and drafting summaries. Its outputs should require validation; generated content should not be treated as human observations, and methodological decisions and final conclusions should remain subject to human approval.

Each AI component should be evaluated against a benchmark appropriate to its function. AI-generated questionnaire suggestions should undergo cognitive interviewing and piloting. Synthetic respondents should be compared with newly collected human data and used principally for pretesting, stress testing or sensitivity analysis. Adaptive AI interviewing should be evaluated experimentally against a standardized protocol. Coding and imputation systems should be tested on independent holdout data that permit assessment across relevant population subgroups. AI-generated interpretations should be reconciled with approved statistical tables, uncertainty estimates and methodological documentation (Bisbee et al., 2024; Breidt & Opsomer, 2017; Schober & Conrad, 1997). Human verification should continue after initial validation so that model drift, performance deterioration and unexpected failure modes remain observable.

The survey system should preserve the information required for reproducibility and audit, including raw responses, prompts, model versions, confidence measures, AI interventions and human overrides. Low-confidence, novel, sensitive or consequential cases should be escalated to trained personnel. Where feasible, human-only and AI-assisted subsamples should be retained periodically to estimate the direction and magnitude of AI-induced measurement, mode or processing effects. Privacy and confidentiality should be protected through data minimization, secure local processing where practicable, access controls, retention limits and formally evaluated disclosure-protection procedures.

The allocation of responsibilities can be expressed through a human-supervised complementarity principle proposed here: *delegate routine and well-specified tasks to AI, retain human responsibility for complex and consequential judgements, and permit AI assistance throughout under risk-proportionate human oversight.* This principle neither leaves AI systems unsupervised nor prevents researchers from using AI when exercising human judgement. It recognizes that survey error can be managed but not eliminated. A human–AI system should

preserve the defining purposes of survey research—observing real populations, eliciting their information under documented conditions, quantifying uncertainty and supporting accountable decisions—while using AI to improve scale, timeliness, consistency, accessibility and the allocation of scarce human expertise.

Figure 2 summarizes the human–AI hybrid approach. It connects human specification, bounded AI assistance, task-specific validation, escalation and continuing monitoring. The feedback paths indicate reassessment, while auditability, privacy safeguards and risk-proportionate human oversight apply throughout.

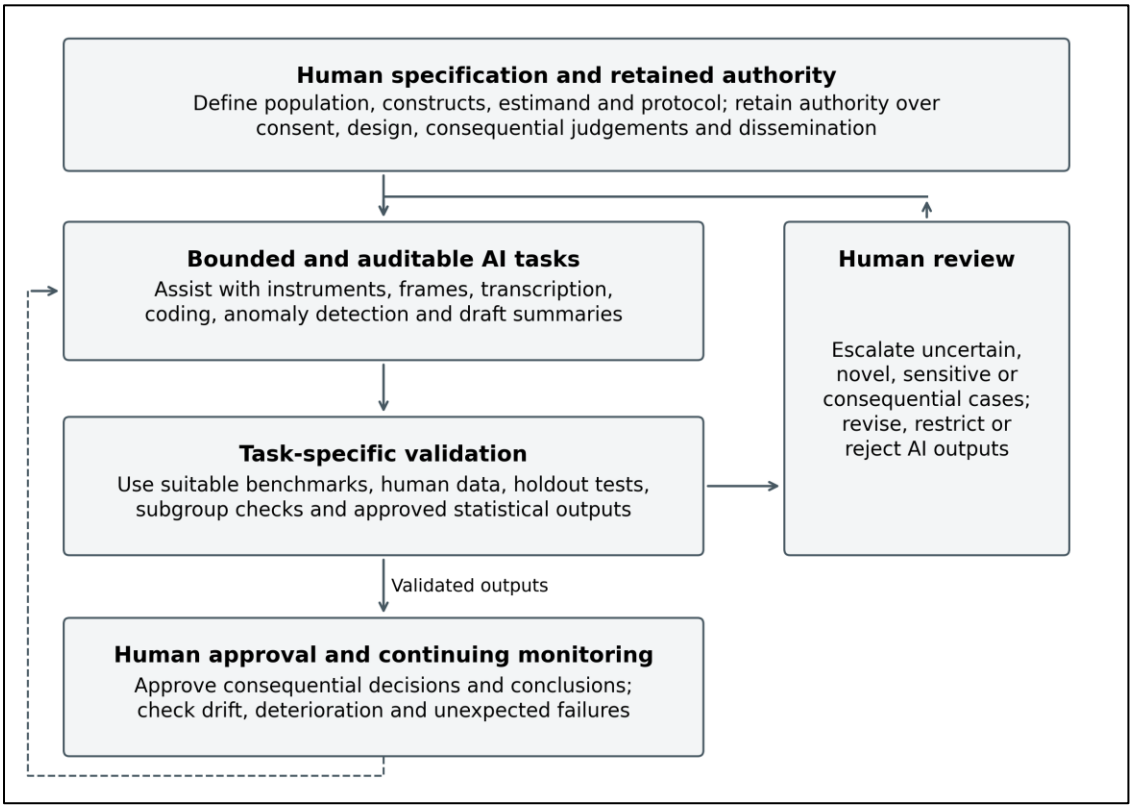


Figure 2. Human-supervised framework for AI-assisted surveys

Note: Author's conceptual synthesis of Human–AI Hybrid Approach to Surveys. The figure summarizes the allocation of responsibilities and safeguards described in this section; these arrangements require empirical evaluation. Monitoring may trigger renewed validation, and escalated cases return to human review before further use.

6(b). Operational Implementation Across Survey Organizations/ Universities/Central Banks

Figure 3 translates the hybrid principle into proposed responsibilities and intervention points. The Principal Researcher/Investigator (PRI) retains responsibility and approves consequential design and release decisions. The Survey Methodology Expert (SME) reviews methodological validity; the Operational Head (OPH) implements protocols and flags exceptions. The Manager Technical /Data Handler (MTD) manages configuration, data, access, versions and maintenance. The Privacy and ethics reviewer (PER) assesses consent, confidentiality and participant protection. Where agents coordinate tasks or use tools, an AI oversight coordinator (AIC) checks permissions, reviews alerts, records interventions and pauses or escalates activity; otherwise OPH and MTD share these duties. Operational or technical clearance cannot substitute for methodological approval.

Before execution, PRI approves constructs, eligibility, sampling, consent, instruments, permitted probing, processing rules and inferential assumptions following SME review and, where relevant, PER review. Routine outputs follow approved protocols; material departures require renewed approval. SME specifies task-appropriate benchmarks and documented acceptance criteria, drawing on Section 6(a), and reviews validation findings. Inadequate performance requires correction and reassessment before consequential use. PRI authorizes publication following methodological and disclosure review.

OPH, MTD or AIC pauses or restricts uncertain, novel, sensitive, anomalous or consequential cases for SME or PER review; PRI resolves consequential departures or authorizes an approved fallback. Records retain models, permissions, protocols, validation findings, material changes, overrides and release decisions. Updates, instrument or population changes, deteriorating performance and subgroup disparities trigger renewed validation. Checks address relevant representation, measurement, processing or inference errors at the affected stages.

A small university team could combine PRI, OPH and technical duties while obtaining additional independent methodological or ethics review for consequential decisions. A commercial organization could separate production from methodological review across recurring projects. A national statistical office or central bank could assign standing methodological, technical and confidentiality responsibilities for repeated or sensitive surveys. Allocation should reflect expertise, resources, frequency, sensitivity and consequences of error, while preserving accountable approval, validation and participant protection in every setting.

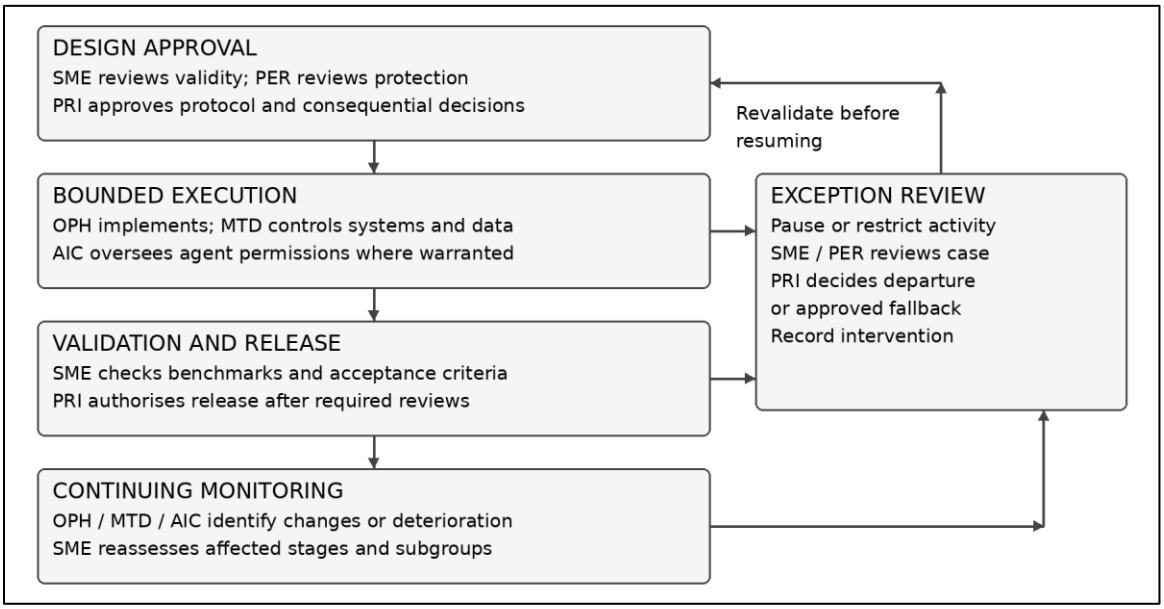


Figure 3. Proposed responsibilities and intervention points for AI-assisted surveys

7. Future Research Agenda

AI is increasingly being applied across multiple stages of the survey lifecycle. However, no published study was identified that develops and evaluates an integrated agentic AI workflow covering the complete survey lifecycle within a total survey error framework. Future research should therefore develop and empirically evaluate a human-supervised, TSE-guided agentic workflow that coordinates survey objectives, population and frame specification, sampling,

questionnaire development, fieldwork, data processing, estimation, statistical inference and dissemination. Evaluation should compare oversight arrangements across organizational settings.

8. Conclusion

This review examined AI applications across the sixteen stages of a survey lifecycle and assessed their implications for stage-specific errors and total survey error (TSE). AI does not enter survey research as a single technology or perform a single function. AI and related computational methods, including machine learning, natural-language processing, computer vision, conversational systems, generative models and privacy-preserving methods, have been applied to frame construction, sample design, questionnaire assessment, interviewing, fieldwork monitoring, coding, imputation, weighting, estimation, analysis, validation and dissemination. AI may operate as a design assistant, measurement instrument, interviewer, respondent aid, processing engine, estimation tool, analytical collaborator, synthetic respondent or object of measurement. Its methodological meaning depends on where it enters the lifecycle and whose information it represents.

The synthesis distinguishes evidence about survey quality from operational and predictive gains. Coding comparisons assess classification accuracy or verified errors located; imputation studies assess bias and efficiency after adjustment; weighting comparisons assess population-mean estimation against conventional alternatives. By contrast, nonresponse prediction evaluates future participation forecasts, conversational studies assess engagement and response characteristics, and sample optimization evaluates sample requirements under precision constraints. Reported gains depend on the studied task, training information, comparator and validation setting. These findings support selected applications, rather than a general claim of superiority. No study identified by the review directly demonstrated reduced TSE across the complete lifecycle. Improvements at one stage cannot establish lower TSE, and richer responses or better predictions do not independently demonstrate measurement validity or reduced nonresponse bias.

This qualification also applies to risks. Comparisons of synthetic and human responses document discrepancies under particular models, prompts and benchmarks; interviewing studies assess clarification or response characteristics under specified protocols. These findings support targeted validation, rather than universal conclusions about AI-induced error. Evidence from conventional automation suggests mechanisms of overreliance, while provider dependence, correlated errors and accountability concerns remain partly conceptual. Privacy guarantees and disclosure assessments also require their own assumptions and evaluation settings. Evidential strength therefore varies across risks and stages. AI may reduce, transfer, conceal or create representation, measurement, processing, estimation and inferential errors, but their combined effect requires lifecycle-wide empirical evaluation rather than extrapolation from isolated comparisons.

The review's principal contribution is to integrate fragmented evidence within a 16-stage lifecycle and an explicit TSE framework. It shows why survey quality is to be evaluated across connected stages rather than inferred from isolated gains. It also proposes a human-supervised, risk-proportionate human–AI hybrid approach to conducting surveys. Surveys should specify the population, reference period, construct, estimand and permissible elicitation protocol before selecting AI tools. AI should undertake bounded and auditable tasks, while humans retain authority over consent, eligibility, sample design, consequential classifications, inferential assumptions and publication. Appropriate benchmarks, continuing human verification, subgroup evaluation,

documentation, escalation of uncertain cases and privacy safeguards are required throughout the lifecycle of a survey. The operational extension specifies responsibilities and approval points adaptable to different organizations.

The most important research priority should now be to develop and empirically evaluate a human-supervised, TSE-guided agentic workflow for the complete probability-survey lifecycle. Such a system should coordinate decisions from objectives, population and frame specification through sampling, measurement, fieldwork, processing, estimation, design-aware inference and dissemination, while permitting errors to be examined at each transition and wave. The enduring methodological standard should not be the extent of automation, but whether a survey maintains a defensible connection between its target population, intended constructs, observed information, quantified uncertainty and accountable conclusions.

References

Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (p. 2867). Association for Computing Machinery. https://doi.org/10.1145/3219819.3226070

Adhikari, D. M., Hartland, A., Weber, I., & Cannanure, V. K. (2025). Exploring LLMs for automated generation and adaptation of questionnaires. In Proceedings of the 7th ACM Conference on Conversational User Interfaces. Association for Computing Machinery. https://doi.org/10.1145/3719160.3736606

Amaya, A. E., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. Journal of Survey Statistics and Methodology, 8(1), 89–119. https://doi.org/10.1093/jssam/smz056

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), 337–351. https://doi.org/10.1017/pan.2023.2

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Ballin, M., & Barcaroli, G. (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. Survey Methodology, 39(2), 369–393.

Barcaroli, G. (2014). SamplingStrata: An R package for the optimization of stratified sampling. Journal of Statistical Software, 61(4), 1–24. https://doi.org/10.18637/jss.v061.i04

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. Public Opinion Quarterly, 71(2), 287–311. https://doi.org/10.1093/poq/nfm006

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. Public Opinion Quarterly, 74(5), 817–848. https://doi.org/10.1093/poq/nfq058

Biemer, P. P., & Lyberg, L. E. (2003). Introduction to survey quality. John Wiley & Sons. https://doi.org/10.1002/0471458740

Birnbaum, B., Borriello, G., Flaxman, A. D., DeRenzi, B., & Karlin, A. R. (2013). Using behavioral data to identify interviewer fabrication in surveys. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2911–2920). Association for Computing Machinery. https://doi.org/10.1145/2470654.2481404

Birnbaum, B., DeRenzi, B., Flaxman, A. D., & Lesh, N. (2012). Automated quality control for mobile data collection. In Proceedings of the 2nd ACM Symposium on Computing for Development (Article 1, pp. 1–10). Association for Computing Machinery. https://doi.org/10.1145/2160601.2160603

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. Political Analysis, 32(4), 401–416. https://doi.org/10.1017/pan.2024.5

Bollenbacher, J., Liao, D., Peytchev, A., & Timbrook, J. (2026, May 22). 4 ways AI is enhancing survey research in 2026. RTI International. https://www.rti.org/insights/ai-enhancing-survey-research-2026

Breidt, F. J., & Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. Statistical Science, 32(2), 190–205. https://doi.org/10.1214/16-STS589

Buskirk, T. D., & Kirchner, A. (2020). Why machines matter for survey and social science researchers: Exploring applications of machine learning methods for design, data collection, and analysis. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), Big data meets survey science: A collection of innovative methods (pp. 9–62). Wiley. https://doi.org/10.1002/9781118976357.ch1

Celino, I., & Re Calegari, G. (2020). Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. International Journal of Human-Computer Studies, 139, Article 102410. https://doi.org/10.1016/j.ijhcs.2020.102410

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer. https://doi.org/10.1007/978-3-642-31164-2

Cohen, M. J., & Warner, Z. (2021). How to get better survey data more efficiently. Political Analysis, 29(2), 121–138. https://doi.org/10.1017/pan.2020.20

Couper, M. P. (2000). Web surveys: A review of issues and approaches. Public Opinion Quarterly, 64(4), 464–494. https://doi.org/10.1086/318641

Dagdoug, M., Goga, C., & Haziza, D. (2023a). Imputation procedures in surveys using nonparametric and machine learning methods: An empirical comparison. Journal of Survey Statistics and Methodology, 11(1), 141–188. https://doi.org/10.1093/jssam/smab004

Dagdoug, M., Goga, C., & Haziza, D. (2023b). Model-assisted estimation through random forests in finite population sampling. Journal of the American Statistical Association, 118(542), 1234–1251. https://doi.org/10.1080/01621459.2021.1987250

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. Journal of Official Statistics, 21(2), 233–255.

Drechsler, J. (2011). Synthetic datasets for statistical disclosure control: Theory and implementation. Springer. https://doi.org/10.1007/978-1-4614-0326-5

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

Fowler, F. J., Jr., & Mangione, T. W. (1990). Standardized survey interviewing: Minimizing interviewer-related error. SAGE Publications. https://doi.org/10.4135/9781412985925

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that tests question comprehensibility. Public Opinion Quarterly, 70(1), 3–22. https://doi.org/10.1093/poq/nfj012

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3), 267–297. https://doi.org/10.1093/pan/mps028

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. Public Opinion Quarterly, 74(5), 849–879. https://doi.org/10.1093/poq/nfq065

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Survey methodology (2nd ed.). John Wiley & Sons.

Gweon, H., & Schonlau, M. (2024). Automated classification for open-ended questions with BERT. Journal of Survey Statistics and Methodology, 12(2), 493–504. https://doi.org/10.1093/jssam/smad015

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. Journal of Official Statistics, 33(1), 101–122. https://doi.org/10.1515/jos-2017-0006

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2024). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. Psychological Methods, 29(3), 494–518. https://doi.org/10.1037/met0000540

He, Z., & Schonlau, M. (2020). Automatic coding of text answers to open-ended questions: Should you double code the training data? Social Science Computer Review, 38(6), 754–765. https://doi.org/10.1177/0894439319846622

He, Z., & Schonlau, M. (2022). A model-assisted approach for finding coding errors in manual coding of open-ended questions. Journal of Survey Statistics and Methodology, 10(2), 365–376. https://doi.org/10.1093/jssam/smab022

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modelling for construct-specific automatic item generation. Psychometrika, 87(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

ICC, & ESOMAR. (2016). ICC/ESOMAR international code on market, opinion and social research and data analytics. https://iccwbo.org/news-publications/policies-reports/iccesomar-international-code-market-opinion-social-research-data-analytics/

Keane, D., & McNaughton, R. B. (2026). Using generative AI to enhance psychometric scale development in market research. International Journal of Market Research, 68(2), 194–218. https://doi.org/10.1177/14707853251384769

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. Survey Research Methods, 13(1), 73–93. https://doi.org/10.18148/srm/2019.v13i1.7395

Kern, C., Li, Y., & Wang, L. (2021). Boosted kernel weighting—Using statistical learning to improve inference from nonprobability samples. Journal of Survey Statistics and Methodology, 9(5), 1088–1113. https://doi.org/10.1093/jssam/smaa028

Kern, C., Weiß, B., & Kolb, J.-P. (2023). Predicting nonresponse in future waves of a probability-based mixed-mode panel with machine learning. Journal of Survey Statistics and Methodology, 11(1), 100–123. https://doi.org/10.1093/jssam/smab009

Kim, S., Lee, J., & Gweon, G. (2019). Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Paper 86, pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300316

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. Public Opinion Quarterly, 72(5), 847–865. https://doi.org/10.1093/poq/nfn063

Kühnemann, H., van Delden, A., & Windmeijer, D. (2020). Exploring a knowledge-based approach to predicting NACE codes of enterprises based on web page texts. Statistical Journal of the IAOS, 36(3), 807–821. https://doi.org/10.3233/SJI-200675

Lind, L. H., Schober, M. F., Conrad, F. G., & Reichert, H. (2013). Why do survey respondents disclose more when computers ask the questions? Public Opinion Quarterly, 77(4), 888–935. https://doi.org/10.1093/poq/nft038

Lohr, S. L. (2022). Sampling: Design and analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429298899

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior, 37, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Lumley, T. (2010). Complex surveys: A guide to analysis using R. John Wiley & Sons. https://doi.org/10.1002/9780470580066

Manski, C. F. (2004). Measuring expectations. Econometrica, 72(5), 1329–1376. https://doi.org/10.1111/j.1468-0262.2004.00537.x

O'Luing, M., Prestwich, S., & Tarim, S. A. (2019). A grouping genetic algorithm for joint stratification and sample allocation designs. Survey Methodology, 45(3), 513–531.

O'Luing, M., Prestwich, S., & Tarim, S. A. (2022). A simulated annealing algorithm for joint stratification and sample allocation. Survey Methodology, 48(1), 225–249.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30*(3), 286–297. https://doi.org/10.1109/3468.844354

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. Journal of Official Statistics, 21(3), 441–462.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. American Journal of Political Science, 58(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rothschild, D., Marlar, J., Amaya, A., Barari, S., Buskirk, T., Cobb, C., Gennai, J., Hillygus, D., Korlakai Vinayak, R., Krupenkin, M., Lee, S., Steiger, D., & Webb, B. (2026). *Responsible AI integration in survey research*. American Association for Public Opinion Research.

Sakshaug, J. W., & Raghunathan, T. E. (2014). Generating synthetic data to produce public-use microdata for small geographic areas based on complex sample survey data with application to the National Health Interview Survey. Journal of Applied Statistics, 41(10), 2103–2122. https://doi.org/10.1080/02664763.2014.909778

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 29971–30004). PMLR. https://proceedings.mlr.press/v202/santurkar23a.html

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. Psychology & Marketing, 41(6), 1254–1270. https://doi.org/10.1002/mar.21982

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? Public Opinion Quarterly, 61(4), 576–602. https://doi.org/10.1086/297818

Schober, M. F., & Conrad, F. G. (2015). Improving social measurement by understanding interaction in survey interviews. Policy Insights from the Behavioral and Brain Sciences, 2(1), 211–219. https://doi.org/10.1177/2372732215601112

Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. Survey Research Methods, 10(2), 143–152. https://doi.org/10.18148/srm/2016.v10i2.6213

Schwanhäuser, S., Sakshaug, J. W., & Kosyakova, Y. (2022). How to catch a falsifier: Comparison of statistical detection methods for interviewer falsification. Public Opinion Quarterly, 86(1), 51–81. https://doi.org/10.1093/poq/nfab066

Sen, I., Ahnert, G., von der Heyde, L., Lasser, J., Weiß, B., & Strohmaier, M. (2026). Total simulated survey error: Designing and diagnosing survey responses from large language models [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2609.10280

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stantcheva, S. (2023). How to run surveys: A guide to creating your own identifying variation and revealing the invisible. Annual Review of Economics, 15, 205–234. https://doi.org/10.1146/annurev-economics-091622-010157

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely sensed and ancillary data. PLOS ONE, 10(2), Article e0107042. https://doi.org/10.1371/journal.pone.0107042

Terry, J., Strait, G., Alsarraf, S., Weinmann, E., & Waychoff, A. (2025). Artificial intelligence in scale development: Evaluating AI-generated survey items against gold standard measures. Current Psychology, 44, 16339–16350. https://doi.org/10.1007/s12144-025-08240-w

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

United Nations. (2005). Designing household survey samples: Practical guidelines. United Nations Statistics Division. https://unstats.un.org/unsd/demographic/sources/surveys/Handbook23June05.pdf

United Nations. (2019). United Nations national quality assurance frameworks manual for official statistics: Including recommendations, the framework and implementation guidance. https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual/

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Valliant, R., Dever, J. A., & Kreuter, F. (2018). Practical tools for designing and weighting survey samples (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-93632-1

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., & Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. Proceedings of the National Academy of Sciences, 115(14), 3529–3537. https://doi.org/10.1073/pnas.1715305115

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. Journal of Advanced Nursing, 52(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. SAGE Publications. https://doi.org/10.4135/9781412983655

Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. ACM Transactions on Computer-Human Interaction, 27(3), Article 15. https://doi.org/10.1145/3381804

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? Computational Linguistics, 50(1), 237–291. https://doi.org/10.1162/coli_a_00502