SoK: Mapping Threats to Defenses in Online Survey Fraud

Shiza Ali[⋆], Wellington Esposito Barbosa[⋆], Matthias Fassl[⋆], Aditi Ganapathi[◇], Jaron Mink[◇], Adam J. Aviv[⋆]

[⋆]*The George Washington University,* [◇]*Arizona State University*

Abstract

Online surveys and recruitment mechanisms, including crowdsourcing platforms such as Prolific and MTurk, as well as social media-based recruitment, have become core infrastructure for human-subjects research. At the same time, their accessibility has made studies increasingly vulnerable to large-scale fraud. Prior work on survey fraud is extensive yet fragmented: different communities use inconsistent definitions, conflate inattentive responding with intentional or automated attacks, and deploy mitigation techniques without explicit threat models. This paper presents a Systematization of Knowledge (SoK) on online survey fraud. Based on a structured review of 124 papers across multiple disciplines that explicitly conceptualize, evaluate, or advance fraud-related mechanisms, we synthesize how fraud is conceptualized, where it arises across the survey lifecycle, and how detection and mitigation strategies are proposed and reported in this literature. Our analysis reveals three recurring gaps: (1) a lack of consistent and explicit definitions that distinguish inattentive responding from adversarial human and automated fraud; (2) systematic misalignment between fraud threats and reported defenses, particularly when recruitment-stage attacks are addressed only post hoc; and (3) inconsistent reporting practices that limit interpretability and reproducibility.

1 Introduction

Online surveys and crowd-sourced data collection have become foundational to empirical research in psychology, public health, social science, human-computer interaction, and usable security. These methods enable rapid, large-scale access to participant populations that are otherwise difficult or costly to reach, and they are now routinely used to study sensitive behaviors, health outcomes, political attitudes, and security-relevant decision-making [9,96]. As online data collection has scaled, researchers across domains have reported growing concerns about fraudulent participation threatening the validity, reliability, and reproducibility of empirical findings [69, 111].

Early work on survey data quality largely framed problematic responses as inattentive or low-effort behavior. Accordingly, much of the literature focused on attention checks, instructional manipulation checks, and response-time thresholds to identify disengaged respondents [1, 83, 86]. While these techniques remain widely used, subsequent studies have documented important limitations, including susceptibility to false positives and the risk of excluding valid participants when applied without clear justification or validation [109,112,121]. Crucially, these approaches were not designed to address adversarial behavior. As online recruitment platforms and paid crowdwork have scaled, the threat landscape has also evolved. More recent work shows that many threats to survey integrity arise from deliberate attempts to exploit recruitment and validation mechanisms, often for financial gain [24, 57]. Documented behaviors include faking eligibility, reusing recruitment links, operating multiple accounts, masking location through VPNs or proxies, and completing surveys at scale using scripts or automated systems [27,41,65]. These behaviors introduce structured bias rather than random noise and can undermine core assumptions of standard survey methodology.

The emergence of large language models (LLMs) further complicates this landscape. Recent studies show that AI-generated survey responses can be fluent, internally consistent, and difficult to distinguish from human-written text using common heuristics [54,66]. Unlike earlier scripted bots, LLM-based systems can adapt responses to survey context, evade attention checks, and mimic demographic or stylistic patterns, weakening many existing detection pipelines [5,123]. As a result, techniques originally developed to filter inattentive responding are increasingly misapplied to adversarial and

USENIX Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23–26, 2026, Hannover, Germany.

automated threats. These challenges impose significant costs on researchers. Health and social science studies report escalating fraud during periods of intensified online recruitment, such as the COVID-19 pandemic, leading to extensive manual verification workflows and repeated data collection [24, 119]. Compensating fraudulent participants results in direct financial loss, while post-hoc filtering and study reruns consume substantial time and resources. Because these attacks target recruitment and validation mechanisms, they cannot be reliably mitigated through response-level checks alone. Despite growing attention to these issues, research on survey fraud remains fragmented. Different communities use inconsistent definitions of "fraud," often conflating inattentive responding, intentional human deception, and automated participation under a single label [65, 114]. Detection and mitigation strategies are frequently proposed in isolation, evaluated on narrow datasets, and deployed without explicit threat models that specify adversary goals, capabilities, or adaptiveness [46, 111]. This fragmentation obscures which defenses address which threats and creates a false sense of security when techniques are applied outside their intended contexts.

In this paper, we present a Systematization of Knowledge (SoK) on online survey fraud. We synthesize prior work across psychology, health sciences, social sciences, economics, and computer science to examine how fraudulent participation is defined, where it arises across the survey lifecycle, and how detection and mitigation strategies are proposed, evaluated, and reported in papers that explicitly engage with survey-fraud-related mechanisms. Our corpus focuses on papers that explicitly define, conceptualize, evaluate, or advance survey-fraud-related mechanisms, rather than all survey-based studies that may use routine data-quality checks. We adopt a security-oriented perspective that distinguishes inattentive responding, intentional human deception, automated participation, and AI-mediated attacks. We also treat recruitment platforms as part of the survey-fraud threat model, because they provide some security-relevant controls while leaving residual risks for researchers to manage.

This SoK is organized around three research questions. For each question, we summarize the main finding.

- **RQ1: How is survey fraud conceptualized and defined across research domains?**
  We find that definitions vary widely. Many papers use the term "fraud" to describe very different behaviors, including inattentive responding, intentional deception, coordinated human fraud, and automated or AI-generated participation. Few studies clearly distinguish between these categories or specify what kind of adversary they assume.
- **RQ2: What detection and mitigation strategies are proposed, and how do they align with specific fraud vectors?**
  We find a clear mismatch between where fraud occurs and where defenses are applied within papers that explicitly discuss fraud-related mechanisms. Many serious threats begin during recruitment, but most defenses are applied later, during survey completion or data cleaning. Across domains, researchers rely heavily on a small set of response-level checks, such as attention or timing measures, and apply them to different types of threats without tailoring defenses to specific risk models. We also find that platform-level controls are important but often under-specified, making it difficult to assess what risks are handled by platforms and what risks remain with researchers.
- **RQ3: How are fraud mitigation practices reported, and what are their implications for interpretability and reproducibility?**
  We find that mitigation decisions are often underdocumented. Papers frequently describe detection methods but do not report decision thresholds, exclusion rates, or how filtering affected the final dataset. This makes it difficult to evaluate the impact of mitigation strategies or to reproduce results.

Accordingly, this SoK makes three primary contributions. First, we consolidate and clarify how survey fraud is defined in literature that explicitly engages with survey-fraud-related mechanisms. Second, we systematize detection and mitigation strategies proposed or evaluated in this literature by mapping them to specific fraud vectors and stages of the survey lifecycle, highlighting recurring mismatches between threats and defenses. Third, we analyze reporting and artifact practices for these fraud-related mechanisms and derive concrete guidance for researchers, reviewers, and platform designers.

By reframing survey fraud as a socio-technical security problem rather than a narrow data-quality concern, this work provides a foundation for more robust, transparent, and sustainable online research infrastructure in the presence of increasingly adversarial participation.

2 Background and Motivation

Online surveys and crowd-sourced data collection serve as core measurement infrastructures across psychology, public health, social science, and usable security. These methods are widely used because they enable rapid recruitment, broad geographic reach, and access to populations that are otherwise difficult or costly to study [9, 96]. As a result, the validity of many empirical claims increasingly depends on the integrity of data collected through online survey platforms.

From Careless Responding to Adversarial Participation. A large body of previous work proposes techniques such as attention checks, instructional manipulation checks, response-time thresholds, and pattern-based indicators to identify disengaged participants [1, 83, 86]. These approaches assume that problematic responses arise from fatigue, misunderstanding, or low motivation, and that such noise can be filtered after data collection without fundamentally altering the dataset. Subsequent research, however, shows that these methods can

generate false positives, disproportionately exclude valid respondents, and inflate effect sizes when applied broadly or without justification [109, 112, 121]. More importantly, not all threats to survey integrity can be explained by inattentiveness alone. Researchers across domains document deliberate behaviors such as misrepresenting eligibility criteria, reusing or reselling recruitment links, operating multiple accounts, and masking location using VPNs or proxy services [27, 65]. These behaviors are often financially motivated and can occur at scale, introducing systematic bias rather than random error. Automation further complicates this landscape. Scripted bots and automated responses have been observed in online surveys [101, 111] and, more recently, LLMs enable AI-mediated participation in which generated responses are fluent, internally consistent, and adaptive to surveys [66, 123]. Empirical evidence suggests that such responses can evade common attention checks and timing-based heuristics, undermining assumptions that non-human participation is easily detectable using surface-level indicators.

The Survey Lifecycle as an Attack Surface. Survey fraud does not occur at a single point in time but unfolds across the survey lifecycle. Researchers commonly distinguish between threats that arise during recruitment (e.g., eligibility misrepresentation, multi-accounting), during survey completion (e.g., inattentive or automated responding), and after data collection (e.g., contamination discovered during cleaning or analysis) [42, 44, 48]. Correspondingly, defenses are deployed at different stages, including recruitment screeners, in-survey checks, and post-hoc filtering [30, 79, 84]. Understanding when fraud occurs is critical because defenses applied late in the lifecycle often cannot fully correct upstream manipulation, particularly when eligibility or identity checks have already been bypassed [47, 68]. Also, different research communities use inconsistent definitions of "fraud," often grouping inattentive responding, intentional deception, and automated participation under a single label [65, 114]. Detection and mitigation strategies are also frequently proposed in isolation [46, 111]. A security-oriented perspective helps clarify these issues by treating survey fraud as a socio-technical phenomenon shaped by platform design, incentive structures, and defensive choices, rather than as individual misconduct alone. From this view, labels such as "careless," "fraudulent," or "bot-generated" responses reflect methodological decisions and embedded assumptions. This perspective motivates the need for a unified, threat-aware systematization of survey fraud, which we develop in the remainder of this paper.

3 Methodology

This section describes how we constructed and analyzed a cross-disciplinary corpus of publications on survey fraud.

Definition of Fraud. For this SoK, we define survey fraud as intentional deception in online survey participation. This includes cases where individuals knowingly provide false information, attempt to bypass study requirements, or use automated tools to generate responses. Examples include eligibility faking, identity or location spoofing, multi-accounting, coordinated participation, scripted bots, and AI-generated responses. We do not treat unintentional low-effort responding as fraud. While careless or inattentive responses can reduce data quality, we focus specifically on behaviors that involve deliberate misrepresentation or attempts to exploit study procedures, often for financial or other incentives. This definition guided both our search strategy and our exclusion criteria.

3.1 Building the Corpus

To construct a scoped corpus of literature on online survey fraud, we followed the PRISMA 2020 protocol [87]. Corpus construction proceeded in three stages: (1) Identification, (2) Screening, and (3) Inclusion. This process was designed to capture work across disciplines that explicitly engages with fraudulent or adversarial participation in online surveys. Figure 1 provides an overview of this process.

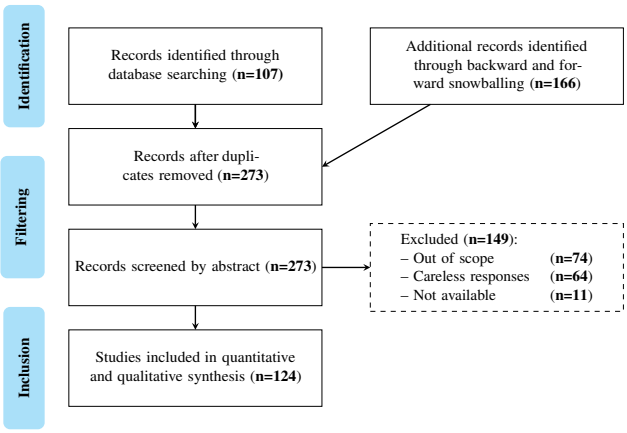

Figure 1: PRISMA flow diagram of the systematic review

Identification. We conducted an initial search across two complementary databases: Google Scholar and DBLP. Google Scholar was selected for its broad coverage across disciplines, while DBLP ensured coverage of peer-reviewed publications in computer science. With this approach, we identified papers from diverse areas, including health, psychology, political science, economics, education, and research methods, as well as subdomains of computer science (human-computer interaction, security, and software engineering).

We refined the search terms iteratively through pilot queries and reference validation, checking whether queries returned highly cited papers on survey fraud and adjusted the terms to improve recall across disciplines. Through this process,

we also included discipline-specific terms that frequently co-occur with discussions of fraudulent or adversarial participation (as defined in **Definition of Fraud**), including: "duplicate," "bot," "AI-mediated," and "crowdwork." While "crowdwork" is not itself a fraud term, it is commonly used in studies examining fraudulent participation on crowdsourcing platforms (e.g., MTurk, Prolific), where eligibility misrepresentation, multi-accounting, and automated responses are often discussed without explicitly using the term "fraud." Our final Boolean query[1] was included in the search protocol.

The initial database search yielded 107 unique publications. To broaden coverage, we conducted backward and forward snowballing on all 107 papers identified through the initial database search. For each paper, we manually examined its reference list (backward snowballing) and used Google Scholar's "cited by" feature to identify subsequent publications that cited it (forward snowballing). This process identified an additional 166 candidate publications. These papers were then subjected to the screening and filtering criteria described below. In total, this strategy identified 273 unique papers for screening.

Filtering. Filtering proceeded in two phases aligned with the research questions introduced in Section 1.

During the filtering phase, we retained papers published between 2010 and 2025 to capture contemporary reports of online survey fraud while reflecting the growth of online recruitment platforms and, more recently, the widespread adoption of large language models (LLMs). Additionally, we retained papers that satisfied at least one of the following criteria:

1. **Conceptualization of Fraud:** The paper provides an explicit or implicit definition of survey fraud, fraudulent participation, or adversarial behavior in online surveys.
2. **Detection or Mitigation:** The paper evaluates or proposes techniques used to detect or mitigate fraudulent participation at any stage of the survey lifecycle.
3. **Guidance or Implications:** The paper offers methodological guidance, design recommendations, or reflections on the implications of survey fraud for research practice, platforms, or policy.

Following the filtering, we excluded 74 from the original 273 papers because they were out of scope, see Figure 1.

Exclusion Criteria for Careless Responses. Careless Responses (CR) can be considered a form of fraud, and the term may be used by researchers interchangeably with survey fraud. However, in this paper, we aimed to differentiate studies that conceptualize CR as a data quality issue from work that frames CR as a fraudulent behavior that requires strategic mitigation to be deterred (see detailed definition of survey fraud and careless responses in section 4.1). We retained careless-responding papers when they were linked to adversarial, deceptive, automated, or financially motivated participation, because these papers help explain how the literature blurs the boundary between data quality problems and fraud. Following that criterion, we excluded papers that focus exclusively on inattentive, careless, or low-effort responding without engaging with adversarial or automated behavior (n=64).

Inclusion. Of the 273 papers included in the abstract screening phase, we excluded studies considered out of scope (n=74), those that primarily framed careless responding as a data quality issue (n=64), and papers for which the full text was unavailable (n=11). The final corpus comprised 124 papers that explicitly engage with survey fraud or fraud-related mechanisms across multiple research domains.

3.2 Analyzing the Papers in the Corpus

We conducted an iterative qualitative analysis to develop and refine a structured codebook. Three coders read the papers in full, with coders taking analytic notes on definitions, assumptions, and methodological choices. They refined codes through regular discussion meetings to resolve differences. To assess the consistency of our screening and coding process, we measured inter-rater reliability (IRR) using Cohen's κ [21]. Three researchers independently coded an initial set of 10 papers resulting in a κ score of 0.38, indicating fair agreement and highlighting ambiguities in code definitions. We then refined code definitions and inclusion criteria through discussion and clarification. Applying the revised codebook to a new set of 10 papers increased IRR to 0.63, reflecting a substantial improvement in shared interpretation. After a further round of refinement, we coded an additional 10 papers and achieved a κ score of 0.72, indicating substantial agreement. This iterative process demonstrates how the codebook stabilized over time and supports the reliability of the coding used in our analysis. Table 1 reports the final IRR values by coding category. After reaching this threshold, coders independently coded the remaining papers in the corpus.

The final codebook (see Appendix A) reflects dimensions that recur across disciplines and that are relevant to understanding survey fraud as a socio-technical security problem. The codes support both quantitative and qualitative synthesis of how different fields explicitly conceptualize fraudulent participation and propose, evaluate, or report fraud-related mechanisms.

3.3 Positionality

Our research team works at the intersection of usable security, trustworthy AI, and human-computer interaction, and

[1] ("online survey") AND ("survey fraud" OR "fraudulent responses" OR "fraudulent participants" OR "fake participants" OR "duplicate responses" OR "bot responses" OR "AI mediated responses" OR "crowdwork responses" OR "scam responses")

Table 1: Inter-rater reliability (Cohen's κ) by code category across three iterative coding rounds.

Code Category	**Round 1** (n=10)	**Round 2** (n=10)	**Round 3** (n=10)
Fraud Definition	0.32	0.58	0.66
Fraud Vector (Behavior)	0.41	0.66	0.75
Detection Stage (Lifecycle)	0.45	0.69	0.78
Detection Technique	0.36	0.62	0.71
Reporting / Outputs	0.34	0.61	0.70
Overall	**0.38**	**0.63**	**0.72**

this interdisciplinary perspective shapes how we study online survey fraud. Collectively, we have experience conducting large-scale quantitative and qualitative studies with crowdworkers on platforms such as Amazon Mechanical Turk and Prolific, as well as running controlled laboratory experiments and participatory design studies on sensitive topics in online safety, usable security, privacy, and human–AI interaction. Since 2019, our team has regularly encountered and managed fraudulent participation in our own online studies. We have developed and refined fraud checks over time in response to recurring issues such as eligibility misrepresentation, duplicate participation, and more recently, AI-generated responses. This direct experience shapes how we frame survey fraud as an evolving adversarial problem and informs our emphasis on stronger recruitment-stage and lifecycle-aware defenses.

3.4 Limitations

Our systematization is shaped by the availability of peer-reviewed literature and by the terminology adopted across different research communities. Although our search strategy was designed to capture cross-disciplinary work that explicitly engages with survey fraud or fraud-related mechanisms, it likely underrepresents informal practitioner knowledge, unpublished platform reports, and emerging fraud behaviors that have not yet appeared in academic venues. In addition, because reporting practices vary substantially across fields, some mitigation techniques, design assumptions, or evaluation details may be under-documented or inconsistently described in the literature. Nonetheless, we believe our analysis encompasses a large body of papers spanning a significant period and multiple domains. This allows us to systematize how survey fraud is explicitly conceptualized in the literature, what fraud vectors and mitigation strategies are proposed or evaluated, and what recommendations are offered across domains.

3.5 Ethics Statement

While this SoK synthesizes findings from studies involving human subjects, we do not directly conduct human-subjects research and therefore believe no additional ethical constraints exist. Nonetheless, we discuss specific ethical considerations related to the survey fraud scenario in the Discussion and Conclusion section.

4 Results

In this section, we examine how prior work in our corpus defines and conceptualizes survey fraud, and show that inconsistent and often implicit threat models obscure important distinctions between inattentive responding, intentional human deception, and automated participation (RQ1). Second, we analyze how fraud vectors are mapped to detection and mitigation strategies across the survey lifecycle in the papers we reviewed, revealing a systematic misalignment between where fraud occurs and where defenses are deployed (RQ2). Third, we study reporting and artifact practices, and find that incomplete documentation of mitigation decisions and outcomes limits interpretability, reproducibility, and comparative evaluation (RQ3).

Study Design Characteristics: Table **A4** (Supplementary Material, see Appendix A) summarizes the composition of the corpus and the study characteristics we coded. Health sciences (n=40) and psychology (n=28) make up the largest portions of the corpus, followed by social sciences (n=19) and computer science (n=19). Within computer science, most papers appear in HCI venues (n=14), while security-focused venues account for only a small fraction (n=4), despite the adversarial and infrastructure-level nature of many documented threats. In terms of contribution type (Table A2), most papers are empirical user studies. Fewer papers are literature reviews (n=9), scoping reviews (n=4), or qualitative studies (n=8). This pattern suggests that prior work in the corpus primarily documents and evaluates fraud in specific applied settings, rather than synthesizing findings across domains. Regarding study motivation (Table A2), the most common goal is to improve survey data quality. Other motivations include examining how fraud affects data quality, responding to high levels of fraud in a specific study, characterizing fraudulent behavior, proposing new detection or mitigation strategies, and, more recently, studying AI-mediated or automated fraud. Taken together, these patterns show that survey fraud is often framed in the corpus as a data integrity problem across the fields represented in our review. However, it is less often framed explicitly as a security or adversarial systems challenge. This gap motivates our cross-disciplinary synthesis.

Threat Model. Our synthesis separates actor classes from the capabilities they may exercise. Table 2 maps actor classes to recurring capabilities, while Table 3 maps those capabilities to intervention stages, defenses, and residual risks. This structure reflects that multiple actor classes can share the same

capability, and that platforms function both as recruitment infrastructure and as a partial layer of defense.

4.1 RQ1: Conceptualization of Fraud

We begin by describing how fraud is defined across the literature. We then step back and identify patterns that emerge across disciplines.

Intentional Data Fabrication and Deception. Several papers define fraud as deliberate falsification of data. These definitions focus on intentional acts of deception that directly compromise data integrity [10, 36, 72]. For example, Birnbaum et al. [10] use the terms "interviewer data fabrication" and "curbstoning" to describe cases where interviewers fabricate responses, often because they cannot reach households or are incentivized to complete surveys quickly. In this framing, fraud is tied to intentional misrepresentation motivated by convenience or compensation. Similarly, work on mischievous responding defines fraud as deliberately providing false or exaggerated answers that distort results [20]. Cimpian et al. [20] describe participants who intentionally report implausible behaviors or identities in ways that create spurious statistical relationships. In their definition, mischievous responders:

> *... might find it funny to report that he eats carrots, fruit, potatoes, and salads each "four or more times a day"; is extremely tall; is unsure whether he has asthma; has never been to the dentist; and that he identifies as "gay," even if none of these is true for this individual in reality. Thus, the presence of mischievous responders creates spurious relationships [...]*

In these definitions, fraud is understood as purposeful deception rather than accidental error.

Careless and Insufficient-Effort Responding. A second set of definitions focuses on careless or low-effort responding [37, 39, 52]. Goldammer et al. [39] define careless responding as patterns in which participants do not attend to survey instructions or item content. These definitions emphasize low motivation, misunderstanding, or inattentiveness. For example, Godinho et al. [37] frames careless responding as being:

> *...characterized by participants' effortless or inattentive response behavior. Originally coined random responding, it is the tendency to respond to items without attention to content. It is generally assumed that such responses are truly random (i.e., equally likely to be chosen) and can be treated empirically as such; [...] Careless responding has also been conceptualized as a subset of a much larger concept known as invalid responding.*

Importantly, this category does not always imply malicious intent. Some authors treat insufficient effort as unintentional, while others acknowledge that low-effort responses may be deliberate, particularly in compensated online environments [33, 72]. This ambiguity blurs the boundary between accidental data quality issues and intentional strategic behavior. Along similar lines, Huang et al. [52] propose the "insufficient effort responding (IER)" label, framing this behavior as a cause rather than a simple response pattern:

> *...we propose the label of insufficient effort responding (IER), defined as a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses. Insufficient effort responding underscores the cause of the response behavior without presupposing specific patterns or outcomes.*

Automated and AI-Mediated Participation. More recent work introduces new terms to describe automated or AI-generated responses [29, 33, 66, 99]. For example, Rilla et al. [99] use the term "LLM pollution" to refer to cases where large language models are used to complete tasks intended to measure human responses. Other studies document scripted bots or coordinated automation on crowdsourcing platforms [29, 33]. In these definitions, fraud is framed as non-human or technology-mediated participation that bypasses assumptions about authenticity and identity.

Fraud in Online Qualitative Studies. Although most of the fraudulent activity documented in the literature focuses on online surveys, our findings suggest that this issue also extends to online qualitative research methods, such as interviews, focus groups, and panels [31, 59, 78, 85, 91, 98, 100, 105]. This is particularly notable among health science and psychology researchers who, during the COVID-19 pandemic, accelerated the adoption of online methodologies to bypass in-person meeting restrictions. Schneider et al. [105] present the following definition:

> *"Fake participants" or "fraudulent respondents" refer to individuals who participate in online research studies without a genuine interest in contributing to the research process. Instead, they may sign up solely to receive incentives, compensation, or rewards offered to research participants; for their own personal agenda; or to disrupt the research process.*

Common fraudulent behaviors include participants misrepresenting their demographics, location, and making false claims to meet eligibility requirements. Specifically, some researchers have reported that as much as 80% of participants in online focus groups were ineligible [85]. Other incidents include participants simulating medical conditions to qualify for the study [91], misrepresenting professional credentials, or attempting to participate multiple times [98]. Due to the smaller sample of studies on interview fraud compared to survey fraud, unless otherwise noted, we primarily report on

Table 2: Actor-capability model for online survey fraud synthesized from the corpus. Symbols summarize recurring associations in the corpus and should not be interpreted as prevalence estimates.

Actor class	Motivation	Low effort	Eligibility	Location	Repeat	Sharing	Automation	AI assist
Inattentive participant	Complete the study with minimal effort	●	–	–	–	–	–	–
Financially motivated individual	Obtain compensation, qualify for more studies, or complete studies faster	●	●	●	●	○	○	○
Coordinated group	Scale participation, share access, or exploit recruitment links	○	●	○	●	●	○	○
Automated system	Complete surveys at scale with little or no human effort	●	○	○	●	○	●	○
LLM-assisted actor	Generate fluent responses, reduce effort, or conceal assistance	○	○	–	–	–	○	●
Platform-recruitment infrastructure	Mediate study access, enforce platform rules, and provide requester-facing controls	–	●	●	●	●	○	○

Legend: ● commonly associated capability; ○ possible or indirect capability; – not typically associated. **Capabilities:** Low effort = satisficing, skipped instructions, straightlining, or random responding; Eligibility = false demographic, health, professional, or experiential claims; Location = location spoofing, VPN/proxy use, or inconsistent geolocation; Repeat = duplicate submissions, repeated participation, or multiple accounts; Sharing = shared links, account reuse, or coordinated access; Automation = scripted or bot-based survey completion; AI assist = LLM- or generative-AI-assisted responses.

the latter throughout the study to provide a more systematic examination of the survey fraud scenario.

Fraud Vector Definitions. We define fraud vectors broadly as the methods a fraudulent participant may employ to deceive a study, either by misrepresenting their demographics or identity to bypass eligibility criteria, or by providing inattentive, false, or automated responses to a survey they may or may not be eligible for. We use the fraud-vector categories summarized in the codebook (Table A3) to distinguish eligibility-related threats, such as location spoofing, identity deception, and duplicate submissions, from response-integrity threats, such as careless responding, mischievous responding, and bot- or AI-generated responses.

Patterns Across Definitions. When we zoom out, two patterns become clear. First, many papers use the term "fraud" to describe behaviors that differ substantially in intent, scale, and motivation [2, 47, 84]. Intentional deception, inattentive responding, coordinated participation, and AI-mediated responses are often grouped together under a single label. Few studies clearly specify which type of adversary they assume. Second, economic incentives are inconsistently acknowledged. Some definitions explicitly reference compensation structures or productivity pressures [10, 72], while others frame fraud primarily as a data quality issue without discussing motivation. As a result, the literature lacks a shared taxonomy that distinguishes accidental low effort from intentional or strategically motivated manipulation. This variation in definitions has direct implications for threat modeling and mitigation design, which we examine in RQ2.

4.2 RQ2: Mitigation Strategies Are Systematically Misaligned with Fraud Vectors

RQ2 examines the detection and mitigation strategies explicitly discussed across the literature and how they relate to different fraud vectors. We first describe the primary categories of mitigation strategies identified in our coding. We then analyze where these strategies are placed across the survey lifecycle. Finally, we examine whether and how these defenses align with the types of fraud they are intended to address. Across the corpus, mitigation strategies cluster into four primary categories: (1) Meta-data based methods (IP address, VPN/proxy detection, browser fingerprinting, time and server side logs), (2) response-level heuristics (e.g., attention checks, error counts, randomness filtering, consistency checks etc.), (3) platform level mechanisms and (4) manual review and post-hoc auditing. Table **A5** (Supplementary Material) reports papers in each category.

Lifecycle Asymmetry in Defense Placement. We find that high-impact fraud vectors, such as misreported demographics and false eligibility claims [36, 72], location spoofing via VPNs or proxies [2, 3], and multi-accounting or duplicate submissions [42, 47, 84], typically originate during recruitment. These behaviors occur before participants meaningfully engage with survey content and can enable repeated, coordinated, or deceptive participation. At the same time, defenses are unevenly distributed across the survey lifecycle within the papers in our corpus that explicitly discuss these fraud vectors or defenses. Table 4 reports the number of papers that explicitly discuss each fraud vector and the lifecycle stage at which it is addressed. These counts should not be interpreted as the prevalence of mitigation practice across survey-based research as a whole. Response-integrity threats are overwhelmingly handled after entry in this corpus: 37 of 41 careless responding papers deploy in-survey checks, and all 32 bot-related papers apply in-survey detection. Only 7 of 41 careless responding papers and 15 of 32 bot-related papers report recruitment-stage controls. We describe the specific in-survey and post-hoc methods used for response-integrity threats in the next subsection. While it can be reasonable to detect inattentive or automated responding during or after sur-

Table 3: Mapping from fraud capabilities to defenses and residual risks using papers from our corpus.

Capability	Stage	Fraud vector examples	Defenses to address it	Residual risks
Low-effort response	○●●	• careless responding • insufficient effort • skipped instructions • straightlining	• attention checks • consistency checks • response-time thresholds • randomness filters	May reflect fatigue, confusion, accessibility needs, or atypical but valid behavior
Eligibility deception	●●○	• false demographic claims • false professional claims • fabricated screener answers	• platform qualifications • study-specific screeners • knowledge checks • consistency checks	Eligibility may remain unverifiable; stronger checks may burden legitimate participants
Location masking	●●○	• VPN or proxy use • false location claims • inconsistent location signals	• IP/geolocation checks • VPN/proxy detection • platform location controls • timezone or locale consistency checks	VPN tools are imperfect: shared networks, or privacy tools may trigger false positives
Repeated access	●○●	• duplicate submissions • repeated participation • multiple platform accounts • device or account reuse	• platform account controls • duplicate-account checks • device/browser fingerprinting • IP/device/log review	Shared devices, institutional networks, and rented accounts remain hard to distinguish
Link/account sharing	●○●	• shared recruitment links • unauthorized survey access • coordinated use of eligible accounts • account reuse	• platform-mediated recruitment • account-level enforcement • referral/source tracking • server-side logs • reporting to platforms	Leaked links and account sharing may remain invisible without platform cooperation
Automated completion	○●●	• scripted form completion • bot-generated responses • repeated automated submissions • CAPTCHA evasion	• CAPTCHA/reCAPTCHA • bot-detection tools • response-pattern analysis • metadata checks • server-side logs	Bots can adapt; fast or atypical humans may be falsely flagged
AI-assisted response	○●●	• LLM-generated open-text responses • AI-assisted screener answers • semantic mimicry	• disclosure questions • task-specific prompts • semantic similarity checks • stylometry • LLM detectors	Short, edited, or mixed-authorship responses remain difficult to classify

Legend: ●○○ *Recruitment* ○●○ *In-survey* ○○● *Post-hoc*

vey completion, this distribution contrasts with recruitment-stage threats that shape who enters the dataset in the first place. Eligibility-related threats show a different pattern. All 17 multi-accounting papers and all 7 location-spoofing papers deploy recruitment-stage defenses, yet many of these studies also rely on downstream filtering (11/17 and 12/17 for multi-accounting; 4/7 for location spoofing). Identity spoofing and false claims are more diffusely addressed across stages (23/47 recruitment, 28/47 in-survey, 31/47 post-hoc), indicating no consistent alignment between threat origin and intervention point among papers that explicitly discuss this vector. Taken together, these counts indicate a structural asymmetry in the explicit fraud-focused literature we reviewed: while recruitment-stage attacks determine who enters the dataset, a substantial portion of the literature relies on in-survey screening and post-hoc exclusion to manage contamination. Such downstream controls may remove problematic data from final analyses, but they do not prevent ineligible participants from accessing surveys, consuming incentives, or adapting to visible detection mechanisms.

Defensive Convergence on Response-Level Heuristics. Across disciplines, researchers rely heavily on a small set of response-level heuristics, including attention checks, response-time thresholds, randomness filtering, and stylometric or semantic analysis. As shown in Table 5, these techniques are applied across nearly all fraud vectors reported in the corpus, even those that begin during recruitment. This pattern persists in studies that explicitly examine adversarial or auto-

Table 4: Lifecycle stage at which fraud vectors are addressed in papers from our corpus.

Fraud vector	Recruitment	In-survey	Post-hoc
Misreported demographics (n=2)	1	1	2
Age misrepresentation (n=1)	0	0	1
Location spoofing (n=7)	7	4	1
Identity spoofing / false claims (n=47)	23	28	31
Multi-accounting / duplicate submissions (n=17)	17	11	12
Careless responding (n=41)	7	37	24
Mischievous responders (n=5)	1	3	4
Bot-generated (n=32)	15	32	23

Table 5: Alignment between reported fraud vectors and detection or mitigation methods.

Fraud vector	Metadata					Response								
	IP / Geo	VPN	Device	Time	Logs	Attention	Error	Random	Semantic	LLM	Style	Honesty	Low-inc.	Knowledge
Misreported demographics / age	○	○	○	◐	○	◐	○	○	○	○	◐	◐	◐	◐
Location spoofing (VPN / proxy)	●	●	◐	◐	○	○	○	○	○	○	○	○	○	○
Identity spoofing / false claims	●	◐	◐	◐	◐	◐	○	○	◐	○	◐	◐	◐	●
Multi-accounting / duplicate submissions	●	◐	◐	●	◐	◐	○	◐	◐	○	◐	○	○	○
Careless responding	○	○	○	●	○	●	◐	●	○	○	◐	◐	◐	○
Mischievous responding	○	○	○	◐	○	◐	○	◐	◐	○	●	◐	◐	○
Bot-generated / AI-mediated	◐	◐	◐	●	◐	◐	○	◐	●	●	●	○	○	○

Legend: ● frequently reported; ◐ occasionally reported or supplementary; ○ rarely reported.

mated fraud. For example, work documenting bot or scripted participation [29, 33] and AI-mediated responses [66, 99] often rely on attention checks or timing-based filters as primary safeguards. These methods were originally designed to detect inattentive or low-effort responding, not adaptive or financially motivated adversaries. As a result, the same defensive toolkit is applied across different threat types. Rather than tailoring defenses to specific fraud vectors, many studies use similar response-level signals regardless of whether the threat involves careless responding, identity deception, coordinated human fraud, or automation.

Recruitment Controls as Supplementary Infrastructure. In contrast, metadata-based and platform-level defenses are structurally positioned at the point where recruitment-stage fraud occurs. Recruitment platforms are therefore important security intermediaries in online survey pipelines. They can provide participant pools, qualification filters, approval or reputation signals, duplicate-account controls, and other platform-specific checks that researchers may rely on when screening participants. Yet Table 5 shows that these controls are used less consistently and are often described as supplementary safeguards rather than primary protections. However, the corpus rarely reports what these platforms specifically verify, what guarantees they provide, or where their protections break down. Although recruitment-stage controls appear in the literature, they are often reported briefly as part of the study setup rather than analyzed as primary security mechanisms. In many cases, these controls are combined with downstream filtering strategies, suggesting that response-level checks remain the dominant line of defense. This reporting pattern indicates that recruitment controls are rarely evaluated independently for their preventive impact. Even when implemented, they are usually combined with downstream filtering instead of serving as front-line defenses. This suggests that infrastructure-level protections are often treated as optional layers rather than integrated components of survey design. Thus, our finding is not that platforms are irrelevant or ineffective, but that their role is often under-specified in the literature. This makes it difficult to distinguish risks mitigated by platform infrastructure from residual risks that researchers must address through study-specific defenses.

Reactive and Non-Scalable Manual Mitigation. Manual review practices further illustrate the reactive nature of current mitigation strategies. As summarized in Table **A5** (Supplementary Material), manual inspection of open-text responses, IP logs, or response patterns is common, particularly in health and social science studies [44, 47, 84]. These reviews are typically triggered after researchers observe unexpected response distributions, suspicious patterns, or evidence of large-scale contamination. While manual review can be effective in identifying anomalous or deceptive responses, it is inherently labor-intensive and difficult to scale. More importantly, it is applied after fraudulent participation has already occurred. Recruitment-stage attacks such as eligibility faking or multi-accounting are often discovered only retrospectively, requiring post-hoc exclusion decisions that may alter sample composition. This pattern suggests that many mitigation strategies are implemented in response to observed anomalies rather than through explicit pre-study threat modeling. Instead of preventive, lifecycle-aware controls, many studies rely on retrospective auditing and manual filtering once contamination becomes visible. Manual mitigation typically involves inspecting open-text responses for similarity or nonsensical content, reviewing IP logs for repeated submissions, examining response time distributions, or cross-checking demographic inconsistencies [29, 44, 45, 47, 50, 68, 77, 98].

Taken together, these results show that mitigation strategies are often selected based on availability and ease of deployment rather than on a principled mapping between fraud vectors and defenses in the fraud-focused literature we reviewed.

As a result, current practices tend to reduce low-effort noise while leaving structural vulnerabilities, particularly those arising early in the survey lifecycle, unaddressed.

4.3 RQ3: Reporting Practices Limit Interpretability and Reproducibility

RQ3 examines how fraud detection and mitigation are reported, and how these reporting choices affect interpretability.

Mitigation as an Opaque Intervention Layer. Across domains, mitigation mechanisms are frequently described, but the operational details of how they are applied remain underspecified. As summarized in Table A1, some studies clearly document how suspected fraud is handled, including preventing survey completion, removing responses, denying compensation, or reporting participants to platforms [29, 47, 48]. However, many papers describe the use of detection mechanisms such as attention checks, response time thresholds, or semantic analysis without specifying key parameters, including decision thresholds, exclusion criteria, or the proportion of responses removed [30, 70, 72]. Similarly, work proposing automated or AI-based detection methods often emphasizes detection performance while providing limited information about downstream effects on the final analytic sample [66, 99]. As a result, readers can see that filtering occurred, but cannot fully reconstruct how mitigation decisions reshaped the dataset. Mitigation itself can introduce bias because aggressive filtering may disproportionately exclude legitimate participants, alter demographic composition, or affect measured effect sizes.

Table A1 further highlights limits to reproducibility. Across domains, guidelines and best-practice recommendations are more common than reusable artifacts. Many studies provide narrative guidance on how to detect or mitigate fraud [29, 47, 84], but do not release code or datasets. Even when new detection methods or algorithms are proposed, artifact availability is limited to a small number of papers: only n=5 papers released code and n=7 papers released a dataset (Table A1). This constrains replication and makes it difficult to compare the effectiveness, costs, and biases of different mitigation strategies across contexts. Taken together, these findings show that current reporting practices limit interpretability and reproducibility.

5 Guidelines for Usable Security and Privacy Researchers

Based on the empirical patterns synthesized in Section 4 we articulate a set of research-grounded guidelines for usable security and privacy researchers who design, evaluate, or review online survey studies. Rather than prescribing best practices, these guidelines reflect systematic gaps observed across the literature and highlight where security- and HCI-informed reasoning can improve the robustness, interpretability, and ethical grounding of survey-based research.

G1. Explicitly specify the threat model underlying fraud mitigation choices. Our review shows that many studies conflate inattentive responding, intentional human deception, and automated or AI-mediated participation, while applying similar defensive techniques to all three. Making the assumed adversary explicit, particularly with respect to intent, scale, and adaptiveness, supports more meaningful evaluation and helps avoid drawing conclusions from defenses that are mismatched to the threat being studied. This does not require assuming that all participants are adversaries. Instead, researchers should state which forms of problematic participation they are trying to address, such as careless responding, eligibility deception, coordinated participation, automation, or AI-assisted response generation. In practice, this means specifying the expected actor class, the point of entry in the survey lifecycle, the signals used to detect the behavior, and the action taken when a response is flagged.

G2. Align defenses with fraud vectors across the survey lifecycle. Our analysis reveals a lifecycle asymmetry in which recruitment-stage fraud vectors are frequently addressed only through in-survey or post-hoc defenses. Researchers should reason explicitly about where threats enter the survey pipeline and assess whether mitigation strategies are preventive or merely corrective.

G3. Prioritize recruitment-stage security controls for eligibility and identity verification. Our findings show that many high-impact fraud vectors originate during recruitment, including eligibility misrepresentation, location spoofing, and multi-accounting. Researchers should treat recruitment as a security boundary rather than a neutral intake stage. Strengthening entry controls—such as eligibility verification, platform qualification mechanisms, and identity checks—can reduce downstream contamination more effectively than relying solely on in-survey filtering. For studies using recruitment platforms such as Prolific or MTurk, this means documenting which platform controls were used, what participant metadata or qualifications were relied on, and what independent checks the researchers added to address residual risks. Researchers should verify study-critical eligibility criteria independently when platform fields are insufficient, record whether location, approval history, account uniqueness, or demographic qualifications came from the platform or from the study instrument, and assume residual exposure to rented accounts, VPNs, shared devices, coordinated participation, and AI-assisted responses.

G4. Treat inattentiveness checks as limited signals rather than general-purpose defenses. Across domains, we observe defensive convergence around a narrow set of response-level heuristics, including attention checks and timing thresholds, applied across heterogeneous threat models. These signals are primarily suited to detecting low-effort re-

sponding and should not be treated as general-purpose safeguards against adaptive or economically motivated adversaries. Researchers should deploy them as one signal among many rather than as primary controls.

G5. Evaluate AI-based detection as a high-impact intervention. While recent work increasingly proposes AI-driven or semantic detection methods, reporting on their error modes and downstream effects remains limited. From a usable security perspective, these systems should be evaluated not only for detection performance, but also for transparency, auditability, and their potential to exclude legitimate participants or introduce systematic bias. Researchers should avoid using AI-based detectors as sole decision-makers for exclusion or compensation decisions. When such systems are used, they should be treated as screening signals, validated against human review or other evidence, and reported with their thresholds, error modes, and downstream effects. Evaluation should report false positives, false negatives, threshold sensitivity, disagreement with human review, effects on exclusion and compensation decisions, and subgroup effects when sample size permits. When feasible, authors should release prompts, model/version information, decision thresholds, annotation procedures, and code or synthetic examples that allow others to reproduce the detector evaluation without exposing participant data.

G6. Report mitigation decisions and their downstream effects on data. Our review finds that many studies describe detection mechanisms without documenting decision thresholds, exclusion rates, or how flagged responses were handled. Explicit reporting of mitigation choices and their effects on sample composition is necessary for reproducibility and for assessing trade-offs between fraud reduction and bias introduction.

G7. Incorporate participant privacy into defense design. Metadata-based and behavioral detection techniques often rely on sensitive signals, yet privacy implications are rarely discussed in detail. Usable security research emphasizes proportionality and data minimization; applying these principles to fraud mitigation helps ensure that defenses do not impose unnecessary harm on legitimate participants.

G8. Make platform and infrastructure dependencies visible. Many mitigation strategies depend on platform-specific affordances, opaque scoring systems, or proprietary controls. Making these dependencies explicit clarifies what aspects of a defense are generalizable and where responsibility lies between researchers and platforms. Researchers should also report where platform guarantees may break down, such as when participants use shared devices, VPNs, rented accounts, AI assistance, or coordinated strategies that platform-level filters may not detect.

G9. Treat online surveys as security-critical research infrastructure. Our findings indicate that applying established security and usability principles, such as explicit threat modeling, defense-in-depth, and usability evaluation to survey research, is necessary to support trustworthy and sustainable empirical work.

G10. Consider the ethical concerns of fraud mitigation strategies. Researchers should engage in meaningful ethical discussions with their peers and Institutional Review Boards (IRBs) about these strategies, particularly when working with underrepresented or hard-to-reach populations, or in any situation where they believe such strategies could harm legitimate participants. These discussions should include the privacy risks of collecting behavioral or metadata signals and the risk of exposing participant data to external AI systems.

6 Discussion and Conclusion

Survey Fraud as a Cross-Domain Problem (RQ1). Our review demonstrates that survey fraud is not confined to any single discipline, but instead affects research in psychology, health sciences, social science, economics, HCI, and security. Despite this shared exposure, these communities differ substantially in how they define, detect, and reason about fraud. Our findings suggest that survey fraud should be understood as a shared socio-technical challenge rather than a series of isolated methodological problems. Addressing it requires coordination across disciplines, platforms, and review cultures, as well as greater alignment between how threats are conceptualized and how defenses are evaluated. From a usable security perspective, this highlights the need for frameworks that support consistent reasoning about adversaries, incentives, and system boundaries across diverse research settings.

Ad Hoc and Misaligned Defenses (RQ2). Across the literature, we observe widespread reliance on ad hoc or improvised mitigation strategies, often adopted in response to immediate experiences of data contamination rather than derived from explicit, threat-aware design. Common techniques such as attention checks, response-time thresholds, and simple pattern-based filters are frequently repurposed to address adversarial or automated fraud, despite having been developed primarily to detect inattentive or low-effort responding. The literature also suggests that fraud vectors are dynamic rather than static. Earlier work largely focused on careless responding and basic automation, whereas more recent studies document coordinated human deception and AI-mediated participation that can scale rapidly and adapt to existing defenses. These newer threats introduce disproportionate harms, including financial loss, systematic bias, and the invalidation or cancellation of entire studies. At the same time, AI is increasingly proposed as part of the solution through automated detection, semantic analysis, or LLM-based screening. While such approaches may improve detection capacity, they also introduce new risks, including opaque decision-making, false positives, and the exclusion of legitimate participants. From a usable security perspective, survey fraud should be treated as an evolving

adversarial problem rather than a static data-quality issue. Defenses should be evaluated not only for detection accuracy, but also for how transparent they are and how they affect legitimate participants.

Another important factor is economic incentives. Many online recruitment platforms offer financial compensation without implementing effective mechanisms for identity verification. This creates opportunities for individuals or automated systems to participate deceptively at scale. Platforms provide partial security such as participant pools, qualification filters, reputation signals, and in some cases, location or identity-related checks. However, their effectiveness depends on what information platforms verify, what signals they expose to researchers, and how well they detect adaptive behaviors such as rented accounts, coordinated participation, VPN use, or AI-assisted responding. Survey fraud is therefore not only careless behavior; it can be a rational response to the incentives built into research platforms. Addressing it requires more than filtering suspicious responses.

Implications for Research Infrastructure and Practice (RQ3). Taken together, our findings suggest that survey fraud is best understood as a socio-technical security problem shaped by incentive structures, platform design, and the distribution of responsibility between researchers and infrastructure providers. Many harms arise not from a single point of failure, but from the interaction between recruitment mechanisms, survey instrumentation, and post-hoc data cleaning practices. When defenses are applied late in the pipeline, researchers are left to manage risks that originate upstream, often through labor-intensive or opaque filtering decisions. As mentioned in 4.3, key details about detection thresholds, exclusion criteria, and downstream effects on sample composition are frequently omitted or under-specified. This limits reproducibility, obscures the costs and biases introduced by mitigation, and makes it difficult for reviewers and readers to assess the robustness of reported findings. Addressing survey fraud, therefore, requires not only better detection techniques but also clearer standards for documentation, stronger integration of defenses into research infrastructure, and shared norms that support transparent and accountable research practice across domains. For platform-mediated studies, this also means reporting which protections were provided by the platform, which risks remained outside the platform's control, and which additional researcher-driven checks were used.

Ethical Considerations of Survey Fraud Mitigation Strategies. Researchers must engage in meaningful ethical discussions with their peers and Institutional Review Boards (IRBs) about the measures they will implement to prevent, detect, and mitigate fraud in their studies. Some fraud detection or mitigation strategies proposed in the literature may unintentionally harm genuine participants, particularly those from underrepresented or hard-to-reach populations. These strategies can also discourage participation in future studies. Similarly, the removal of outlier data needs to be approached with caution, especially when working with underrepresented populations. Genuine responses could be flagged as fraudulent based on subjective criteria, particularly if participants from these groups are unfamiliar with research studies. These participants might be confused by attention checks or reverse-worded items, and may provide simpler open-text responses. Discarding such data could inadvertently exclude the very people for whom the study was designed. AI-based fraud detection introduces additional ethical risks. If researchers send participant responses or behavioral traces to external AI systems, they may expose human-subjects data beyond the context participants expected or consented to. Opaque AI outputs can also make it difficult to explain why a participant was flagged or to assess whether errors disproportionately affect particular groups.

Furthermore, decisions regarding participant compensation, such as whether to advertise it with the study, or withholding payment due to fraud suspicion, should be clearly defined, justified in the study protocol, and approved by the IRB to avoid negative consequences for both participants and researchers. Researchers should also exercise caution when requiring participants to confirm their eligibility, such as by phone, email, or providing documentation (e.g., to verify medical conditions), to receive compensation after the participation has already been completed. Legitimate participants might feel burdened by these additional requirements, forgoing their rightful compensation and being wrongly classified as fraudulent. This could potentially reduce their willingness to participate in future studies.

Why This Matters for Usable Security and HCI. Despite its clear relevance to security and human-centered system design, survey fraud has received limited sustained attention from the HCI and security communities. Existing work in these venues often addresses narrow technical mechanisms without engaging with the broader socio-technical context documented in other fields. Our SoK shows that meaningful progress requires integrating insights from across disciplines and applying threat modeling, lifecycle thinking, and usability considerations to survey research itself. This creates an opportunity for usable security and privacy researchers to contribute frameworks, tools, and evaluation practices that improve both data integrity and participant protections. It also creates an opportunity to study the systems around survey research, including recruitment platforms, researcher workflows, IRB expectations, participant communication, and reporting norms. Taken together, our findings suggest a clearer way to understand survey fraud.

References

[1] James D. Abbey and Margaret G. Meloy. Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, 53-56(1), 2017.

[2] Jon Agley, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo. Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior Research Methods*, 54(2):885–897, April 2022.

[3] Douglas J. Ahler, Carolyn E. Roush, and Gaurav Sood. The micro-task market for lemons: data quality on Amazon's Mechanical Turk. *Political Science Research and Methods*, 13(1):1–20, January 2025.

[4] Lesley Andrew, Emily Gizzarelli, Mohamed Estai, and Ruth Wallace. Participant Misrepresentation in Online Focus Groups: Red Flags and Proactive Measures. *Ethics & Human Research*, 46(1):37–42, January 2024.

[5] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[6] Winfred Arthur Jr, Ryan M. Glaze, Anton J. Villado, and Jason E. Taylor. The Magnitude and Extent of Cheating and Response Distortion Effects on Unproctored Internet-Based Tests of Cognitive Ability and Personality. *International Journal of Selection and Assessment*, 18(1):1–16, 2010.

[7] Scott Barge and Hunter Gehlbach. Using the Theory of Satisficing to Evaluate the Quality of Survey Data. *Research in Higher Education*, 53(2):182–200, March 2012.

[8] Alexandra L. Bartell and Jan H. Spyridakis. Managing risk in internet-based survey research. In *2012 IEEE International Professional Communication Conference*, pages 1–6, October 2012. ISSN: 2158-1002.

[9] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science*, 58(3), 2014.

[10] Benjamin Birnbaum, Gaetano Borriello, Abraham D. Flaxman, Brian DeRenzi, and Anna R. Karlin. Using behavioral data to identify interviewer fabrication in surveys. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2911–2920, New York, NY, USA, 2013. Association for Computing Machinery.

[11] Stephen Bonett, Willey Lin, Patrina Sexton Topper, James Wolfe, Jesse Golinkoff, Aayushi Deshpande, Antonia Villarruel, and José Bauermeister. Correction: Assessing and Improving Data Integrity in Web-Based Surveys: Comparison of Fraud Detection Systems in a COVID-19 Study. *JMIR Formative Research*, 9:e76462, 2025.

[12] James Bonnamy, Bethany Carr, Michelle D. Lazarus, and Clifford Connell. Survey sabotage: Insights into reducing the risk of fraudulent responses in online surveys. *Anatomical Sciences Education*, 18(8), 2025.

[13] Nathan A. Bowling, Jason L. Huang, Cheyna K. Brower, and Caleb B. Bragg. The Quick and the Careless: The Construct Validity of Page Time as a Measure of Insufficient Effort Responding to Surveys. *Organizational Research Methods*, 26(2):323–352, April 2023.

[14] Julii Brainard, Anne Killett, Julie Houghton, Diane Bunn, Laura Watts, Suzanne Mumford, Sarah J. O'Brien, and Kathleen Lane. The Wasps are Clever: Keeping Out and Finding Bot Answers in Internet Surveys Used for Health Research. Preprints.org, April 2022.

[15] Cheyna Katherine Brower. *What Are You Looking At? Using Eye-Tracking to Provide Insight into Careless Responding*. Doctoral dissertation, Wright State University, 2020.

[16] Morgan E. Browning, Sidney L. Satterfield, and Elizabeth E. Lloyd-Richardson. Mischievous responders: data quality lessons learned in mental health research. *Ethics & Behavior*, 34(5), July 2024.

[17] Sara Bybee, Kristin Cloyes, Brian Baucom, Katherine Supiano, Kathi Mooney, and Lee Ellington. Bots and nots: safeguarding online survey research with underrepresented and diverse populations. *Psychology & Sexuality*, 13(4), October 2022.

[18] Chadwick K. Campbell, Samuel Ndukwe, Karine Dubé, John A. Sauceda, and Parya Saberi. Overcoming Challenges of Online Research: Measures to Ensure Enrollment of Eligible Participants. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 91(2):232–236, October 2022.

[19] Alessandro Checco, Jo Bates, and Gianluca Demartini. All That Glitters Is Gold — An Attack Scheme on Gold Questions in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 6:2–11, June 2018.

[20] Joseph R. Cimpian, Jennifer D. Timmer, Michelle A. Birkett, Rachel L. Marro, Blair C. Turner, and Gregory L. Phillips. Bias From Potentially Mischievous Responders on Large-Scale Estimates of Lesbian, Gay, Bisexual, or Questioning (LGBQ)–Heterosexual Youth Health Disparities. *American Journal of Public Health*, 108(S4):S258–S265, November 2018.

[21] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[22] Josielli Comachio, Adam Poulsen, Adeola Bamgboje-Ayodele, Aidan Tan, Julie Ayre, Rebecca Raeside, Rajshri Roy, and Edel O'Hagan. Identifying and counteracting fraudulent responses in online recruitment for health research: a scoping review. *BMJ Evidence-Based Medicine*, 30(3), June 2025.

[23] Beverly G. Conrique, Elizabeth McDade-Montez, and Pamela M. Anderson. Detection and Prevention of Data Fraud in a Study of Community College Career Technical Education Students. *Community College Journal of Research and Practice*, 44(9), September 2020.

[24] Leslie S Craig, Christina L Evans, Brittany D Taylor, Jace Patterson, Kaleb Whitfield, Mekhi Hill, Michelle Nwagwu, Mohamed Mubasher, Robert A Bednarczyk, Gail G McCray, Cheryl L R Gaddis, Natasha Taylor, Emily Thompson, Ursula Douglas, Saundra K Latimer, Sedessie G Spivey, Tabia Henry Akintobi, and Rakale Collins Quarells. Challenges and Lessons Learned in Managing Web-Based Survey Fraud for the Garnering Effective Outreach and Research in Georgia for Impact Alliance–Community Engagement Alliance Survey Administrations. *JMIR Public Health and Surveillance*, 10:e51786–e51786, December 2024.

[25] Anastasia Danilova, Alena Naiakshina, Stefan Horstmann, and Matthew Smith. Do you Really Code? Designing and Evaluating Screening Questions for Online Surveys with Programmers. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 537–548, Madrid, ES, May 2021. IEEE.

[26] Molly R. Davies, Dina Monssen, Helen Sharpe, Karina L. Allen, Beki Simms, Kimberley A. Goldsmith, Sarah Byford, Vanessa Lawrence, and Ulrike Schmidt. Management of fraudulent participants in online research: Practical recommendations from a randomized controlled feasibility trial. *International Journal of Eating Disorders*, 57(6):1311–1321, June 2024.

[27] Sean A. Dennis, Brian M. Goodson, and Christopher A. Pearson. Online Worker Fraud and Evolving Threats to the Integrity of MTurk Data: A Discussion of Virtual Private Servers and the Limitations of IP-Based Screening Procedures. *Behavioral Research in Accounting*, 32(1):119–134, March 2020.

[28] James Dewitt, Benjamin Capistrant, Nidhi Kohli, B R Simon Rosser, Darryl Mitteldorf, Enyinnaya Merengwa, and William West. Addressing Participant Validity in a Small Internet Health Survey (The Restore Study): Protocol and Recommendations for Survey Response Validation. *JMIR Research Protocols*, 7(4):e96, April 2018.

[29] Liesje Donkin, Nathan Henry, Amy Kercher, Mangor Pedersen, Holly Wilson, and Amy Hai Yan Chan. Effective Recruitment or Bot Attack? The Challenge of Internet-Based Research Surveys and Recommendations to Reduce Risk and Improve Robustness. *Interactive Journal of Medical Research*, 14:e60548, March 2025.

[30] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. Are your participants gaming the system? screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2399–2402, New York, NY, USA, 2010. Association for Computing Machinery.

[31] Kerryn Drysdale, Nathanael Wells, Anthony K J Smith, Nilakshi Gunatillaka, Elizabeth Ann Sturgiss, and Tim Wark. Beyond the challenge to research integrity: imposter participation in incentivised qualitative research and its impact on community engagement. *Health Sociology Review*, 32(3), September 2023.

[32] Marc Dupuis, Emanuele Meier, and Félix Cuneo. Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods*, 51(5):2228–2237, October 2019.

[33] Marc Dupuis, Emanuele Meier, Mehdi Gholam-Rezaee, Gerhard Gmel, Marie-Pierre F. Strippoli, and Olivier Renaud. Detecting computer-generated random responding in online questionnaires: An extension of Dupuis, Meier & Cuneo (2019) on dichotomous data. *Personality and Individual Differences*, 157:109812, April 2020.

[34] Jeffrey R. Edwards. Response invalidity in empirical research: Causes, detection, and remedies. *Journal of Operations Management*, 65(1):62–76, January 2019.

[35] Gudrun Eisele, Hugo Vachon, Ginette Lafit, Peter Kuppens, Marlies Houben, Inez Myin-Germeys, and Wolfgang Viechtbauer. The Effects of Sampling Frequency

and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*, 29(2):136–151, March 2022.

[36] Holly Fernandez Lynch, Steven Joffe, Harsha Thirumurthy, Dawei Xie, and Emily A. Largent. Association Between Financial Incentives and Participant Deception About Study Eligibility. *JAMA Network Open*, 2(1):e187355, January 2019.

[37] Alexandra Godinho, Vladyslav Kushnir, and John A. Cunningham. Unfaithful findings: identifying careless responding in addictions research: Editorial. *Addiction*, 111(6):955–956, June 2016.

[38] Masaki Gogami, Yuki Matsuda, Yutaka Arakawa, and Keiichi Yasumoto. Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone. *IEEE Access*, 9:53205–53218, 2021.

[39] Philippe Goldammer, Hubert Annen, Peter Lucas Stöckli, and Klaus Jonas. Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, 31(4):101384, August 2020.

[40] Juan Marcos Gonzalez, Kiran Grover, Thomas W. Leblanc, and Bryce B. Reeve. Did a bot eat your homework? An assessment of the potential impact of bad actors in online administration of preference surveys. *PLOS ONE*, 18(10):e0287766, October 2023.

[41] Brittney Goodrich, Marieke Fenton, Jerrod Penn, John Bovay, and Travis Mountain. Battling bots: Experiences and strategies to mitigate fraudulent responses in online surveys. *Applied Economic Perspectives and Policy*, 45(2):762–784, June 2023.

[42] Jeremy A. Grey, Joseph Konstan, Alex Iantaffi, J. Michael Wilkerson, Dylan Galos, and B. R. Simon Rosser. An Updated Protocol to Detect Invalid Entries in an Online Survey of Men Who Have Sex with Men (MSM): How Do Valid and Invalid Submissions Compare? *AIDS and Behavior*, 19(10):1928–1937, October 2015.

[43] Marybec Griffin, Richard J. Martino, Caleb LoSchiavo, Camilla Comer-Carruthers, Kristen D. Krause, Christopher B. Stults, and Perry N. Halkitis. Ensuring survey research data integrity in the era of internet bots. *Quality & Quantity*, 56(4):2841–2852, August 2022.

[44] Jodie L Guest, Elizabeth Adam, Iaah L Lucas, Cristian J Chandler, Rebecca Filipowicz, Nicole Luisi, Laura Gravens, Kingsley Leung, Tanaka Chavanduka, Erin E Bonar, Jose A Bauermeister, Rob Stephenson, and Patrick S Sullivan. Methods for Authenticating Participants in Fully Web-Based Mobile App Trials from the iReach Project: Cross-sectional Study. *JMIR mHealth and uHealth*, 9(8):e28232, August 2021.

[45] Arryn A. Guy, Matthew J. Murphy, David G. Zelaya, Christopher W. Kahler, and Shufang Sun. Data integrity in an online world: Demonstration of multimodal bot screening tools and considerations for preserving data integrity in two online social and behavioral research studies with marginalized populations. *Psychological Methods*, September 2024.

[46] Daniel Habib and Nishant Jha. AIM against survey fraud. *JAMIA Open*, 4(4):ooab099, October 2021.

[47] Robert W. Hammond, Claudia Parvanta, and Rahel Zemen. Caught in the Act: Detecting Respondent Deceit and Disinterest in On-Line Surveys. A Case Study Using Facial Expression Analysis. *Social Marketing Quarterly*, 28(1):57–77, March 2022. Publisher: SAGE Publications Inc.

[48] Jeffrey J Hardesty, Elizabeth Crespi, Joshua K Sinamo, Qinghua Nian, Alison Breland, Thomas Eissenberg, Ryan David Kennedy, and Joanna E Cohen. From Doubt to Confidence—Overcoming Fraudulent Submissions by Bots and Other Takers of a Web-Based Survey. *Journal of Medical Internet Research*, 26:e60184, December 2024.

[49] Rachael M. Hewitt, Catherine Purcell, and Chris Bundy. Safeguarding online research integrity: concerns from recent experience. *British Journal of Dermatology*, 187(6):999–1000, December 2022.

[50] Kris L Hohn, April A Braswell, and James M DeVita. Preventing and Protecting Against Internet Research Fraud in Anonymous Web-Based Research: Protocol for the Development and Implementation of an Anonymous Web-Based Data Integrity Plan. *JMIR Research Protocols*, 11(9):e38550, September 2022.

[51] Maxwell Hong, Jeffrey T. Steedle, and Ying Cheng. Methods of Detecting Insufficient Effort Responding: Comparisons and Practical Recommendations. *Educational and Psychological Measurement*, 80(2):312–345, April 2020. Publisher: SAGE Publications Inc.

[52] Jason L. Huang, Nathan A. Bowling, Mengqiao Liu, and Yuhui Li. Detecting Insufficient Effort Responding with an Infrequency Scale: Evaluating Validity and Participant Reactions. *Journal of Business and Psychology*, 30(2):299–311, June 2015.

[53] Jason L. Huang, Paul G. Curran, Jessica Keeney, Elizabeth M. Poposki, and Richard P. DeShon. Detecting

and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology*, 27(1):99–114, March 2012.

[54] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, Hamburg Germany, April 2023. ACM.

[55] Michael John Ilagan and Carl F. Falk. Supervised Classes, Unsupervised Mixing Proportions: Detection of Bots in a Likert-Type Questionnaire. *Educational and Psychological Measurement*, 83(2):217–239, April 2023.

[56] Kathryn Irish and Jessica Saba. Bots are the new fraud: A post-hoc exploration of statistical methods to identify bot-generated responses in a corrupt data set. *Personality and Individual Differences*, 213:112289, October 2023.

[57] Roseline Jean Louis and Lisa M. Thompson. Bots and fake participants: ensuring valid and reliable data collection using online participant recruitment methods. *International Journal of Social Research Methodology*, 28(4), July 2025.

[58] Malcolm S. Johnson, Vanessa M. Adams, and Jason Byrne. Addressing fraudulent responses in online surveys: Insights from a web-based participatory mapping study. *People and Nature*, 6(1):147–164, February 2024.

[59] Abigail Jones, Line Caes, Tessa Rugg, Melanie Noel, Sharon Bateman, and Abbie Jordan. Challenging issues of integrity and identity of participants in non-synchronous online qualitative methods. *Methods in Psychology*, 5:100072, December 2021.

[60] Andrew Jones, Jessica Earnest, Martyna Adam, Ross Clarke, Jack Yates, and Charlotte R. Pennington. Careless responding in crowdsourced alcohol research: A systematic review and meta-analysis of practices and prevalence. *Experimental and Clinical Psychopharmacology*, 30(4):381–399, August 2022.

[61] Anjana Karumathil and Ritu Tripathi. Combating Survey Bots in Online Research: An Integrative Literature Review of Insights and Strategies. *AIS Transactions on Human-Computer Interaction*, 17(1):80–109, March 2025.

[62] Bennett King-Nyberg, Erica Fae Thomson, Janet Morris-Reade, Roberta Borgen, and Cassie Taylor. The Bot Toolbox: An Accidental Case Study on How to Eliminate Bots from Your Online Survey. *Journal for Social Thought*, 7(1), September 2023.

[63] Jessica Kramer, Amy Rubin, Wendy Coster, Eric Helmuth, John Hermos, David Rosenbloom, Rich Moed, Meghan Dooley, Ying-Chia Kao, Kendra Liljenquist, Deborah Brief, Justin Enggasser, Terence Keane, Monica Roy, and Mark Lachowicz. Strategies to address participant misrepresentation for eligibility in Web-based research. *International Journal of Methods in Psychiatric Research*, 23(1):120–129, March 2014.

[64] Michaela Krawczyk and Katie A. Siek. When Research Becomes All About the Bots: A Case Study on Fraud Prevention and Participant Validation in the Context of Abortion Storytelling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, pages 1–8, New York, NY, USA, 2024. Association for Computing Machinery.

[65] Jennifer Lawlor, Carl Thomas, Andrew T Guhin, Kendra Kenyon, Matthew D Lerner, UCAS Consortium, and Amy Drahota. Suspicious and fraudulent online survey participation: Introducing the REAL framework. *Methodological Innovations*, 14(3):20597991211050467, September 2021.

[66] Benjamin Lebrun, Sharon Temtsin, Andrew Vonasch, and Christoph Bartneck. Detecting the corruption of online questionnaires by artificial intelligence. *Frontiers in Robotics and AI*, 10:1277635, February 2024.

[67] Dominik Johannes Leiner. Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys. *Survey Research Methods*, 13(3):229–248, December 2019.

[68] S Elisha LePine, Catherine Peasley-Miklus, Meghan L Farrington, William J Young, Michelle T Bover Manderski, Mary Hrywna, and Andrea C Villanti. Ongoing Refinement and Adaptation are Required to Address Participant Deception in Online Nicotine and Tobacco Research Studies. *Nicotine & Tobacco Research*, 25(1):170–172, January 2023.

[69] Ronli Levi, Ronit Ridberg, Melissa Akers, and Hilary Seligman. Survey Fraud and the Integrity of Web-Based Survey Research. *American Journal of Health Promotion*, 36(1):18–20, January 2022.

[70] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, 47(2):519–528, June 2015.

[71] Gemma Loebenberg, Melissa Oldham, Jamie Brown, Larisa Dinu, Susan Michie, Matt Field, Felix Greaves, and Claire Garnett. Bot or Not? Detecting and Managing Participant Deception When Conducting Digital Research Remotely: Case Study of a Randomized Controlled Trial. *Journal of Medical Internet Research*, 25:e46523, September 2023.

[72] Cara C. MacInnis, Harrison C.D. Boss, and Joshua S. Bourdage. More evidence of participant misrepresentation on Mturk and investigating who misrepresents. *Personality and Individual Differences*, 152:109603, January 2020.

[73] Kinnon Ross MacKinnon, Naail Khan, Katherine M Newman, Wren Ariel Gould, Gin Marshall, Travis Salway, Annie Pullen Sansfaçon, Hannah Kia, and June Sh Lam. Introducing Novel Methods to Identify Fraudulent Responses (Sampling With Sisyphus): Web-Based LGBTQ2S+ Mixed-Methods Study. *Journal of Medical Internet Research*, 27:e63252, March 2025.

[74] Michael R. Maniaci and Ronald D. Rogge. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48:61–83, February 2014.

[75] Kaitlyn McLachlan, Emma E. Truffyn, Bianka Dunleavy, Delane Linkiewich, Deborah Powell, Anna Taddio, and C. Meghan McMurtry. Fraudulent participation in psychological research using virtual synchronous interviews: ethical challenges and potential solutions. *Ethics & Behavior*, 35(3), April 2025.

[76] A. Dana Ménard, Suzanne McMurphy, Morgan Sterling, Nicholas Armstrong, Oliver Cheek, and Storm Balint. Bots, Scammers, and Fraudulent Responders: A Year of Disrupted Data Collection. *Ethics & Behavior*, 36(3):235–249, 2026.

[77] Michael H. Miner, Walter O. Bockting, Rebecca Swinburne Romine, and Sivakumaran Raman. Conducting Internet Research With the Transgender Population: Reaching Broad Samples and Collecting Valid Data. *Social Science Computer Review*, 30(2):202–211, May 2012.

[78] Khaylen Mistry, Sophie Merrick, Melissa Cabecinha, Susanna Daniels, John Ragan, Miran Epstein, Louisa Lever, Zoe C. Venables, and Nick J. Levell. Fraudulent Participation in Online Qualitative Studies: Practical Recommendations on an Emerging Phenomenon. *Qualitative Health Research*, page 10497323241288181, November 2024.

[79] Jason W Mitchell, Tanaka M D Chavanduka, Stephen Sullivan, and Rob Stephenson. Recommendations From a Descriptive Evaluation to Improve Screening Procedures for Web-Based Studies With Couples: Cross-Sectional Study. *JMIR Public Health and Surveillance*, 6(2):e15079, May 2020.

[80] Annabelle M. Mournet and Evan M. Kleiman. Internet-Based Mental Health Survey Research: Navigating Internet Bots on Reddit. *Cyberpsychology, Behavior, and Social Networking*, 26(2):73–79, February 2023.

[81] Marek Muszyński. Attention checks and how to use them: Review and practical recommendations. *Ask: Research and Methods*, 32(1):3–38, 2023.

[82] Wen Zhi Ng, Sundarimaa Erdembileg, Jean C J Liu, Joseph D Tucker, and Rayner Kay Jin Tan. Increasing Rigor in Online Health Surveys Through the Reduction of Fraudulent Data. *Journal of Medical Internet Research*, 27:e68092–e68092, August 2025.

[83] A. Susan M. Niessen, Rob R. Meijer, and Jorge N. Tendeiro. Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, 63:1–11, August 2016.

[84] Aasli Abdi Nur, Christine Leibbrand, Sara R Curran, Elizabeth Votruba-Drzal, and Christina Gibson-Davis. Managing and minimizing online survey questionnaire fraud: lessons from the Triple C project. *International Journal of Social Research Methodology*, 27(5), September 2024.

[85] Nicola O'Donnell, Rose-Marie Satherley, Emily Davey, and Gemma Bryan. Fraudulent participants in qualitative child health research: identifying and reducing bot activity. *Archives of Disease in Childhood*, 108(5):415–416, May 2023.

[86] Leonard J. Paas, Sara Dolnicar, and Logi Karlsson. Instructional Manipulation Checks: A longitudinal analysis with implications for MTurk. *International Journal of Research in Marketing*, 35(2):258–269, June 2018.

[87] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, page n71, March 2021.

[88] Joseph J. Palamar and Patricia Acosta. On the Efficacy of Online Drug Surveys during the Time of COVID-19. *Substance Abuse*, 41(3):283–285, July 2020.

[89] Aswati Panicker, Novia Nurain, Zaidat Ibrahim, Chun-Han (Ariel) Wang, Seung Wan Ha, Yuxing Wu, Kay Connelly, Katie A. Siek, and Chia-Fang Chung. Understanding fraudulence in online qualitative studies: From the researcher's perspective. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–17, New York, NY, USA, 2024. Association for Computing Machinery.

[90] Weiping Pei, Arthur Mayer, Kaylynn Tu, and Chuan Yue. Attention Please: Your Attention Check Questions in Survey Studies Can Be Automatically Answered. In *Proceedings of The Web Conference 2020*, WWW '20, pages 1182–1193, New York, NY, USA, 2020. Association for Computing Machinery.

[91] Elizabeth Pellicano, Dawn Adams, Laura Crane, Calliope Hollingue, Connie Allen, Katherine Almendinger, Monique Botha, Tori Haar, Steven K Kapp, and Elizabeth Wheeley. Letter to the Editor: A possible threat to data integrity for online qualitative autism research. *Autism*, 28(3):786–792, March 2024.

[92] Natalia Pinzón, Vikram Koundinya, Ryan E. Galt, William O'R Dowling, Marcela Baukloh, Namah C. Taku-Forchu, Tracy Schohr, Leslie M. Roche, Samuel Ikendi, Mark Cooper, Lauren E. Parker, and Tapan B. Pathak. AI-powered fraud and the erosion of online survey integrity: an analysis of 31 fraud detection strategies. *Frontiers in Research Metrics and Analytics*, 9, December 2024.

[93] Artur Pokropek, Tomasz Żółtak, and Marek Muszyński. Mouse Chase: Detecting Careless and Unmotivated Responders Using Cursor Movements in Web-Based Surveys. *European Journal of Psychological Assessment*, 39(4):299–306, 2023.

[94] Mandi Pratt-Chapman, Jenna Moses, and Hannah Arem. Strategies for the Identification and Prevention of Survey Fraud: Data Analysis of a Web-Based Survey. *JMIR Cancer*, 7(3):e30730, July 2021.

[95] Matthew Price, Johanna E. Hidalgo, Julia N. Kim, Alison C. Legrand, Zoe M.F. Brier, Katherine Van Stolk-Cooke, Amy Hughes Lansing, and Ateka A. Contractor. The cyborg method: A method to identify fraudulent responses from crowdsourced data. *Computers in Human Behavior*, 157:108253, August 2024.

[96] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343, San Francisco, CA, USA, May 2019. IEEE.

[97] Denise L. Reyes. Combatting carelessness: Can placement of quality check items help reduce careless responses? *Current Psychology*, 41(10):6858–6866, October 2022.

[98] Damien Ridge, Laurna Bullock, Hilary Causer, Tamsin Fisher, Samantha Hider, Tom Kingstone, Lauren Gray, Ruth Riley, Nina Smyth, Victoria Silverwood, Johanna Spiers, and Jane Southam. 'Imposter participants' in online qualitative research, a new and increasing threat to data integrity? *Health Expectations*, 26(3):941–944, June 2023.

[99] Raluca Rilla, Tobias Werner, Hiromu Yakura, Iyad Rahwan, and Anne-Marie Nussberger. Recognising, Anticipating, and Mitigating LLM Pollution of Online Behavioural Research, November 2025.

[100] Jacqueline Roehl and Darci Harland. Imposter Participants: Overcoming Methodological Challenges Related to Balancing Participant Privacy with Data Quality When Using Online Recruitment and Data Collection. *The Qualitative Report*, November 2022.

[101] Zachary Joseph Roman, Holger Brandt, and Jason Michael Miller. Automated Bot Detection Using Bayesian Latent Class Models in Online Surveys. *Frontiers in Psychology*, 13:789223, April 2022.

[102] Emma Ruby, Serine Ramlawi, Alexa Clare Bowie, Stephanie Boyd, Alysha Dingwall-Harvey, Ruth Rennicks White, Darine El-Chaâr, and Mark Walker. Identifying Fraudulent Responses in a Study Exploring Delivery Options for Pregnancies Impacted by Gestational Diabetes: Lessons Learned From a Web-Based Survey. *Journal of Medical Internet Research*, 27:e58450, January 2025.

[103] Margaret R. Salinas. Are Your Participants Real? Dealing with Fraud in Recruiting Older Adults Online. *Western Journal of Nursing Research*, 45(1):93–99, January 2023.

[104] Alan Santinele Martino, Arielle Perrotta, and Brenna Janet McGillion. Who can you trust these days?: Dealing with imposter participants during online recruitment and data collection. *Qualitative Research*, 24(5):1291–1301, October 2024.

[105] Jekaterina Schneider, Latika Ahuja, Jessica R. Dietch, Anne-Mairead Folan, Jillian Coleman, and Kathleen Bogart. Addressing fraudulent responses in quantitative and qualitative internet research: case studies

from body image and appearance research. *Ethics & Behavior*, 35(7), October 2025.

[106] Ulrich Schroeders, Christoph Schmidt, and Timo Gnambs. Detecting Careless Responding in Survey Data Using Stochastic Gradient Boosting. *Educational and Psychological Measurement*, 82(1):29–56, February 2022.

[107] Saijal Shahania, Myra Spiliopoulou, and David Broneske. Gotta Catch 'Em All... Or Not?: How LLMs Bypass Traditional Checks & Mimic Human Response Behavior in Web Surveys. In *Proceedings of the ACM Collective Intelligence Conference*, pages 113–128, San Diego CA USA, August 2025. ACM.

[108] Thomas J. Shaw, Cory J. Cascalheira, Emily C. Helminen, Cal D. Brisbin, Skyler D. Jackson, Melissa Simone, Tami P. Sullivan, Abigail W. Batchelder, and Jillian R. Scheer. Yes stormtrooper, these are the droids you are looking for: Identifying and preliminarily evaluating bot and fraud detection strategies in online psychological research. *Psychological Methods*, 30(6), 2025.

[109] Henning Silber, Joss Roßmann, and Tobias Gummer. The Issue of Noncompliance in Attention Check Questions: False Positives in Instructed Response Items. *Field Methods*, 34(4):346–360, November 2022.

[110] Amber I Sophus and Jason W Mitchell. Assessment of Fraud Deterrence and Detection Procedures Used in a Web-Based Survey Study With Adult Black Cisgender Women: Description of Lessons Learned and Recommendations. *JMIR Formative Research*, 9:e59955, March 2025.

[111] Andie Storozuk, Marilyn Ashley, Véronic Delage, and Erin A. Maloney. Got Bots? Practical Recommendations to Protect Online Survey Data from Bot Attacks. *The Quantitative Methods for Psychology*, 16(5):472–481, May 2020.

[112] Morgan D. Stosic, Brett A. Murphy, Fred Duong, Amber A. Fultz, Summer E. Harvey, and Frank Bernieri. Careless Responding: Why Many Findings Are Spurious or Spuriously Inflated. *Advances in Methods and Practices in Psychological Science*, 7(1):25152459241231581, January 2024.

[113] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 367–385, Boston, MA, USA, August 2022. USENIX Association.

[114] Jennifer E. F. Teitcher, Walter O. Bockting, José A. Bauermeister, Chris J. Hoefer, Michael H. Miner, and Robert L. Klitzman. Detecting, Preventing, and Responding to "Fraudsters" in Internet Research: Ethics and Tradeoffs. *Journal of Law, Medicine & Ethics*, 43(1):116–133, 2015.

[115] Amber D. Thompson and Rebecca L. Utz. Online surveys: lessons learned in detecting and protecting against insincerity and bots. *Quality & Quantity*, 59(1):23–39, February 2025.

[116] Frederic Traylor. The threat of AI chatbot responses to crowdsourced open-ended survey questions. *Energy Research & Social Science*, 119:103857, January 2025.

[117] Philip Waggoner, Ryan Kennedy, and Scott Clifford. Detecting Fraud in Online Surveys by Tracing, Scoring, and Visualizing IP Addresses. *Journal of Open Source Software*, 4(37):1285, May 2019.

[118] Lorraine O. Walker, Nicole Murry, and Kayla D. Longoria. Improving Data Integrity and Quality From Online Health Surveys of Women With Infant Children. *Nursing Research*, 72(5):386, October 2023.

[119] June Wang, Gabriela Calderon, Erin R. Hager, Lorece V. Edwards, Andrea A. Berry, Yisi Liu, Janny Dinh, August C. Summers, Katherine A. Connor, Megan E. Collins, Laura Prichett, Beth R. Marshall, and Sara B. Johnson. Identifying and preventing fraudulent responses in online public health surveys: Lessons learned during the COVID-19 pandemic. *PLOS Global Public Health*, 3(8):e0001452, August 2023.

[120] Shengqian Wang, Israt Jahan Jui, and Julie Thorpe. Is Crowdsourcing a Puppet Show? Detecting a New Type of Fraud in Online Platforms. In *Proceedings of the New Security Paradigms Workshop*, NSPW '24, pages 84–95, New York, NY, USA, January 2025. Association for Computing Machinery.

[121] M. K. Ward and Adam W. Meade. Dealing with Careless Responding in Survey Data: Prevention, Identification, and Recommended Best Practices. *Annual Review of Psychology*, 74:577–596, January 2023.

[122] Margaret A. Webb and June P. Tangney. Too Good to Be True: Bots and Bad Data From Mechanical Turk. *Perspectives on Psychological Science*, 19(6):887–890, November 2024. Publisher: SAGE Publications Inc.

[123] Sean Westwood. The potential existential threat of large language models to online survey research. *PNAS*, 2025.

[124] Thomas A. Willis, Alexandra Wright-Hughes, Clare Skinner, Amanda J. Farrin, Suzanne Hartley, Rebecca Walwyn, Ana Weller, Mohamed Althaf, Stephanie Wilson, Chris P. Gale, and Robbie Foy. The detection and management of attempted fraud during an online randomised trial. *Trials*, 24(1):494, August 2023.

[125] Christina Yarrish, Laurie Groshon, Juliet Mitchell, Ashlyn Appelbaum, Samantha Klock, Taylor Winternitz, and Dara G. Friedman-Wheeler. Finding the Signal in the Noise: Minimizing Responses From Bots and Inattentive Humans in Online Research. *The Behavior Therapist*, 42(7):235–242, October 2019.

[126] Edanur Yazici and Ying Wang. Attack the bot: Mode effects and the challenges of conducting a mixed-mode household survey during the Covid-19 pandemic. *International Journal of Social Research Methodology*, 27(6):791–796, November 2024.

[127] Ziyi Zhang, Shuofei Zhu, Jaron Mink, Aiping Xiong, Linhai Song, and Gang Wang. Beyond Bot Detection: Combating Fraudulent Online Survey Takers. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 699–709, New York, NY, USA, 2022. Association for Computing Machinery.

A Supplementary Material Availability

We provide the full codebook and additional tables as supplementary material, including representative papers on the distribution of prior work on survey fraud across research domains (Table A4) and fraud detection and mitigation strategies (Table A5): https://anonymous.4open.science/r/SoK-Mapping-Threats-to-Defenses-in-Online-Survey-Fraud-E9D9.

Table A1: Output, replicability, and artifact-availability codes used to evaluate prior work on survey fraud.

Category	Code / Subcategory	Representative Papers
Primary Output Type	Tool or software system	[46, 62, 107]
	New algorithm or detection method	[10, 13, 23, 32, 38, 53, 55, 73, 79, 83, 93, 95, 101, 114, 117, 119, 126]
	Empirical measurement study	[1–3, 6, 7, 9, 11–13, 15–18, 24, 26–28, 31, 33, 35, 36, 40–43, 45, 52, 56, 57, 64–67, 69–72, 75, 76, 78, 80, 82, 86, 88, 90, 94, 96, 97, 103, 104, 108, 109, 111, 112, 116, 121, 125, 127]
	Guidelines or recommendations	[1, 3, 8, 9, 11, 12, 14, 20, 22, 24, 26, 28, 29, 31, 39, 41–43, 47, 48, 50, 52, 63, 64, 67, 69, 71, 74, 81, 82, 84, 92, 94, 97, 103–105, 108, 110, 111, 114–116, 118, 120, 121, 124, 127]
	Suggested/examples of questions	[1, 14, 25, 79, 86, 90]
Artifact Availability	Code provided	[2, 33, 45, 92, 106]
	Dataset provided	[2, 20, 33, 45, 92, 106, 115]

Table A2: The contribution and motivation of studies included in the SoK.

Category	Subcategory	Representative Papers
Contribution Type	User study	[1–4, 6, 7, 9–13, 16–18, 20, 23–30, 32, 33, 35, 36, 38–50, 52, 53, 55–58, 62–77, 79–84, 86, 88–90, 92–97, 101–104, 106–112, 114–122, 124–127]
	Literature review	[8, 22, 34, 60, 61, 63, 81, 94, 121]
	Scoping review	[34, 37, 51, 99]
	Qualitative studies	[31, 59, 78, 85, 91, 98, 100, 105]
Study Motivation	Improve survey data quality	[1, 2, 7–9, 11, 12, 18, 22–24, 26–29, 31–35, 37–39, 41, 43, 44, 46–48, 52, 53, 55, 58–60, 63, 65–67, 69, 70, 73, 75, 77, 81–83, 86, 88, 90, 93, 94, 97, 109, 112, 114, 117–119, 121, 125, 127]
	Impacts of fraud on data quality	[3, 6, 16, 20, 40, 42, 45, 49, 61, 74, 78, 85, 89, 91, 103, 104, 111, 120]
	Excessive fraud in user study	[14, 17, 49, 56, 57, 62, 64, 71, 76, 80, 95, 98, 100–102, 105, 108, 115, 118, 122, 124, 126]
	Characterize fraudulent behavior	[4, 14, 15, 19, 36, 48, 65, 68, 69, 72, 84, 96, 113, 120]
	Propose a detection or mitigation strategy	[10, 13, 15, 23, 25, 29, 30, 34, 38, 39, 42, 45, 50, 51, 55, 61, 73, 79, 82, 92, 106, 108, 110, 114, 117, 119]
	Study the impact of AI-mediated fraud	[17, 40, 41, 43, 54, 56, 57, 62, 64, 71, 76, 80, 95, 99, 101, 107, 116, 123, 126]

Table A3: Fraud conceptualization codes: definitions, automation level, and types of fraudulent behavior.

Category	Code / Subcategory	Representative Papers
Definition Availability	Definition of Survey Fraud	[3, 6, 8, 10–12, 18, 20, 22–31, 33, 36, 41, 45–50, 53, 56–59, 61, 63–66, 69, 71–73, 75, 76, 78, 79, 82, 84, 88, 94–96, 103, 104, 108, 114, 115, 117, 119]
	Definition of Careless Response	[1, 7, 9, 13, 15, 32, 34, 35, 37–39, 51, 52, 67, 70, 74, 81, 83, 86, 90, 93, 97, 106, 108, 109, 112, 121, 125]
Level of Automation	Traditional Bot	[2, 3, 12, 17, 29, 32, 33, 40, 41, 43, 45, 48, 49, 55–57, 59, 61, 62, 64, 68, 71, 76, 80, 95, 101, 108, 111, 115, 125, 126]
	AI-based Bot	[32, 55, 66, 99, 108]
Type of Fraud	Multiple Submissions	[42, 44, 47, 48, 50, 59, 84, 115]
	Inattentive	[1, 9, 30, 32, 35, 38, 67, 70, 81, 83, 86, 90, 93, 97, 109, 112, 121, 125]
Fraud Vector: Eligibility	Misreported demographics	[4, 98]
	Age misrepresentation	[49]
	Location spoofing (e.g., VPN)	[2–4, 108, 110, 118, 127]
	Identity spoofing / false claims	[2, 8, 11, 12, 18, 22–29, 31, 36, 41, 42, 45–48, 50, 53, 56–58, 63–65, 68, 69, 71–73, 75–79, 82, 84, 88, 94–96, 103, 104, 114, 117, 119]
	Multi-accounting / duplicate submissions	[3, 8, 42, 44, 47, 48, 77, 79, 84, 92, 98, 108, 110, 118, 120, 124, 127]
Fraud Vector: Response Integrity	Careless responding	[1–3, 7, 9, 10, 13, 15, 29, 30, 32, 34, 35, 37–39, 47, 52, 67, 70, 74, 77, 81, 83, 86, 90, 93, 97, 109, 110, 112, 115, 118, 120, 121, 124, 125]
	Mischievous responders	[3, 6, 16, 20, 45]
	Bot-generated	[2, 3, 12, 14, 17, 29, 32, 33, 40, 41, 43, 55–57, 62, 64, 66, 71, 76, 80, 92, 95, 101, 108, 110, 111, 115, 116, 125–127]