

ARTICLES

Data Quality in Online Crowdsourced Surveys: Methodological Challenges and Analytic Insights in Survey Research

Kevin B. Gittner, PhD, MBA^{1,2}, Lauren M. Matheny, PhD, MPH¹, Gregory D. Balkcom, MS, BS¹, Kristine Duncan, MS, BS¹, Andy E. Lewis, BS^{1,2}, James Down, MS, BS¹, An Truong, BS¹

¹ School of Data Science and Analytics, College of Computing and Software Engineering, Kennesaw State University, ² Department of Health Promotion and Physical Education, Wellstar College of Health and Human Services, Kennesaw State University

Keywords: Online survey recruitment, Attention check questions, Crowdsourcing platforms, Data quality, Amazon's Mechanical Turk, Qualtrics, SurveyMonkey, Web surveys, Fraud, Data Integrity, Fraudulent responses

<https://doi.org/10.29115/SP-2026-0017>

Survey Practice

Vol. 20, 2026

Pay-for data collection platforms are increasingly popular in health sciences, due to the ease of use, accessibility, and speed of data collection. However, data quality among platforms may differ. The purpose of this study was to compare data quality from three commonly used, pay-for data collection crowdsourcing platforms, using predefined quality control measures. Data were collected using three major platforms, Qualtrics (N = 705), SurveyMonkey (N = 576), and Amazon Mechanical Turk (N = 1,034), for the methodological purpose of assessing data quality for a scale used by public health practitioners to assess ankle activity level. Data were collected using attempted quota sampling for gender, age, ethnicity and region, for representativeness of the general US population. Twenty quality checks in five categories were utilized to identify poor-quality data: attention verification (e.g. 'please select response B'); demographic verification (e.g. state and zip code match); illogical responses (e.g. respondents separately reporting higher difficulty walking up one flight of stairs than walking up four flights); honesty/reliability verification (e.g. male reported to be pregnant); and assessment of open-ended questions (rejected based on similarity of content or nonsensical responses). Of the 2,315 completed responses received from the three different platforms, 947 (40.91%) passed our quality checks and were considered good-quality responses. SurveyMonkey yielded the highest proportion of good-quality data (70.5%), followed by Qualtrics (61.8%) and Amazon Mechanical Turk (10.2%); however, open-ended responses could not be included in SurveyMonkey due to their 80-item limit, which may have inflated its proportion of good-quality responses. This study demonstrates that data quality issues may differ across platforms and a variety of quality control measures should be implemented for researchers utilizing pay-for data collection services from crowdsourcing platforms.

Introduction

Over the past two decades, digital data collection has become one of the most popular approaches in the health sciences, with widespread adoption throughout public health and clinical care (Forkus et al. 2023; Naumova 2025; Wazny 2018). Online surveys allow researchers to assess individual health outcomes and generate population-level insights through structured data collection. The use of pay-for data collection or crowdsourcing

platforms, has surged in recent years, with the COVID-19 pandemic acting as a catalyst (Hensen et al. 2021; Hlatshwako et al. 2021; Torrejón-Guirado et al. 2024). This rapid expansion, reflected in a 2024 market valuation exceeding \$3.9 billion (The Business Research Company 2026), highlights the growing demand and distinct advantages these platforms offer including efficient, large-scale data collection and improved access to specific and hard-to-reach sub-populations (Ball 2019; Dupuis et al. 2022; Walker et al. 2023).

A valuable application of crowdsourcing platforms in health science research is the collection of patient-reported outcomes (Khan et al. 2022), standardized measures that reflect individuals' experiences across the continuum of care (Davidson et al. 2020; Martin et al. 2005). These platforms offer an efficient way to gather patient-reported outcomes data from diverse, global populations, especially when other sampling methods are impractical. However, concerns remain about the quality of data collected through these services and the integrity of the conclusions drawn from the data (Douglas et al. 2023; Newman et al. 2021; Peer et al. 2022; Torrejón-Guirado et al. 2024).

While crowdsourcing platforms have made data collection more accessible, new challenges necessitate stronger data quality control and assessment standards. The growing presence of malicious programs, such as survey “bots” and automated form-fillers, poses a significant threat to data integrity (Guy et al. 2024; Marshall et al. 2023; Storozuk et al. 2020). The prevalence of fraudulent and inattentive respondents, whose low-quality input often evades standard screening, raises concern that, in some cases, the drawbacks of crowdsourcing platforms may outweigh the benefits (Dupuis et al. 2022; Lebrun et al. 2024). Recent federal indictments exposed large-scale fraud schemes in which individuals were paid to impersonate legitimate survey takers, using virtual private networks and coordinated deception to fabricate survey data for market research firms (U.S. Attorney's Office, District of New Hampshire 2025). Although these threats are increasingly acknowledged, clarity is lacking on how data quality varies across platforms, which differ in user interface, recruitment strategies, and built-in quality control mechanisms. Systematic evaluations are needed to compare platforms and guide selection of tools best suited for high-quality survey data in health research, where reliable data are critical for evaluating treatment outcomes and informing clinical decisions. The purpose of this study was to systematically assess data quality across the three leading pay-for data collection platforms using a variety of rigorous data quality checks embedded within a health survey focused on physical function and activity level.

Methods

Data Collection

A custom online survey was developed and administered using three pay-for-data collection crowdsourcing platforms, including SurveyMonkey, Qualtrics, and Amazon Mechanical Turk (MTurk), to collect responses from individuals in the United States (U.S.) general population. The applied purpose of this work was to document psychometric properties and establish normative values for a recently developed orthopaedic outcomes instrument, the Foot and Ankle Activity Level Scale (FAALS; Matheny et al. 2025). Although the survey content focused on physical function and activity level for that applied component, the study reported here is the embedded methodological portion. In this methodological study, we used a consistent survey across platforms to evaluate and compare data quality through systematic quality control checks. Data were collected using demographic quotas, including age, race, sex, educational level, income, and geographic location to improve external validity by improving representation of the U.S. population, according to the U.S. Census Bureau (2020). This study was approved by an institutional review board, and all respondents provided informed consent to participate in the study and complete the survey.

Recruitment

Recruitment and data collection procedures varied by platform. SurveyMonkey and Qualtrics managed participant recruitment, survey distribution, and compensation through their proprietary systems. In contrast, MTurk provided access to participant panels but required researchers to independently deploy the survey, monitor response progress, and issue compensation. A full-length questionnaire was designed separately within both Qualtrics and SurveyMonkey platforms and distributed through each platform's recruitment system. For MTurk participants, the same Qualtrics questionnaire was shared with the MTurk sampling panel, and respondents were required to record a randomized code upon survey completion to ensure accurate linkage and delivery of incentive payments. The goal was to collect 1,500 survey responses (500 per crowdsourcing platform) deemed "good-quality".

Survey Instrument

The survey included 80 questions for SurveyMonkey and 95 questions for Qualtrics and MTurk. For SurveyMonkey, an 80-question limit required combining race and ethnicity items and removing several open-ended and demographic questions. The 80-item version included a consent question; four demographic items; seven questions on anthropometric characteristics and prior knee or ankle injuries or surgeries; four randomly placed attention verification questions (AVQs), the 22-item Foot and Ankle Activity Level Scale (FAALS) (Matheny 2021; Matheny et al. 2025), and four widely used

patient-reported outcome measures: the 21-item Foot and Ankle Ability Measure (FAAM) Activities of Daily Living (ADL) and 8-item FAAM-Sport subscales (Martin et al. 2005), the single-item Tegner Activity Level Scale (Tegner and Lysholm 1985; Tegner et al. 1986; 1988), and the 12-item Short Form (SF-12) Health Survey (Ware et al. 1996). The use of AVQs as a quality control measure is well documented and has been previously shown to identify poor-quality responses (Agans et al. 2023; Douglas et al. 2023; Liu and Wronski 2018; Newman et al. 2021). All questions, except for the final open-ended question, which asked participants for any further comments, were required for survey progression. The 95-item version (Qualtrics and MTurk) included three additional demographic questions, five supplemental questions on anthropometric and injury history, and seven questions on honesty and demographic verification. A “think aloud” protocol was implemented during survey development to ensure item clarity and interpretability among members of the target population (Fonteyn et al. 1993). As suggested by Eysenbach (2004), a table documenting all aspects of our online data collection was included ([Appendix 2](#)).

Data Quality Assessment

To evaluate accuracy and honesty in responses, data from each platform were analyzed using 20 quality assessments (i.e. checks) ([**Appendix 1**](#)). Not all checks were applied on every platform: 19 were used in SurveyMonkey, 20 in Qualtrics, and 20 in MTurk. The checks fell into five categories: (1) *Attention Verification Question Checks* requiring a specific response (e.g., “please select Moderate Difficulty”); (2) *Demographic Checks* to ensure consistency within or against platform-collected internal data (e.g., state and zip code match); (3) *Logic Checks* to flag illogical answers across items (e.g., rating a half marathon as easier than walking on uneven ground); (4) *Honesty-Reliability Checks* to detect fraudulent or implausible data, including unrealistic body mass index (BMI) values under 8.5 (Suszko et al. 2022) or over 205 (Public Health Nigeria 2022) and survey speeders (completion time less than half the platform-specific median, per Qualtrics guidance); and (5) *Open-Ended Question Checks* to screen open-ended responses for irrelevant or duplicate content (only available in Qualtrics and MTurk). Responses failing any quality checks were rejected and classified as “poor-quality data.”

Survey Pilot Testing

During the pilot survey, issues were identified with duplicate responses, unrealistic height and weight values, and participation by non-U.S. residents, which contributed to differences between the 80- and 95-question versions since SurveyMonkey had an 80-item limit. To address these problems, we added open-ended items to detect duplicates, a BMI check for plausible values, a country-of-residence item, and two demographic verification questions, while also enabling built-in fraud and bot detection in Qualtrics and MTurk. Because MTurk offers minimal concrete guidance on

appropriate incentive structures (Amazon Mechanical Turk, n.d.), compensation decisions were informed by limited guidance in the literature and aimed to meet or exceed U.S. minimum wage (\$7.25/hour), with a target of approximately \$10/hour based on median completion time estimates (Saravanos et al. 2021). Pilot studies were used to estimate completion times and refine incentives. Initial testing started with a \$0.10 payment for screening completion and a \$1.20 payment for completion of the full survey. Based on pilot results, incentives were adjusted to a \$0.05 payment for screening completion and a \$1.45 payment for attentive completion of the full survey, for a total possible payment of \$1.50. This structure was intended to provide reasonable compensation while discouraging low-effort responding.

Statistical Methods

Descriptive statistics, including means, medians, ranges, and percentages were calculated. Interrater reliability was calculated to assess researcher agreement for manual open-ended question rejections, which were also compared to the automated rejection method using a minimum 80% similarity threshold between open-ended question responses (McHugh 2012). A contingency table was created of good- versus poor-quality responses from the three different platforms, and a chi-square test was used to compare the proportion of good-quality data across service platforms. Pairwise post hoc tests were adjusted for multiple comparisons. Statistical analyses were performed in R version 4.2.2 (R Core Team 2022) and in SAS version 9.4 (SAS Institute Inc. 2023).

Results

Across all three platforms, a total of 5,756 responses were recorded. Of these, 2,977 respondents met the inclusion criteria and initiated the survey. Of the 2,977, 662 cases were removed for failing attention verification question checks: 391 from Qualtrics (35.7% of 1,096), 145 from SurveyMonkey (20.1% of 721), and 126 from MTurk (10.9% of 1,160), ensuring consistency in AVQ-related exclusion across platforms. This left 2,315 individuals who completed the survey and were included in the final analysis ([Figure 1](#)). To ensure methodological consistency across platforms, responses that failed AVQ checks were removed prior to analysis. Qualtrics automatically terminated survey sessions when an attention check was failed, so equivalent handling was applied across all datasets for comparability.

Data Quality Assessment

Data quality was assessed by data collection platform ([Figure 2](#)). Of the 2,315 total completed responses, 947 (40.9%) passed all quality checks and were classified as good-quality data. A significant difference in data quality was observed across platforms ($\chi^2 = 740.8$, $p < 0.001$), with post hoc comparisons confirming differences between all three: SurveyMonkey vs.

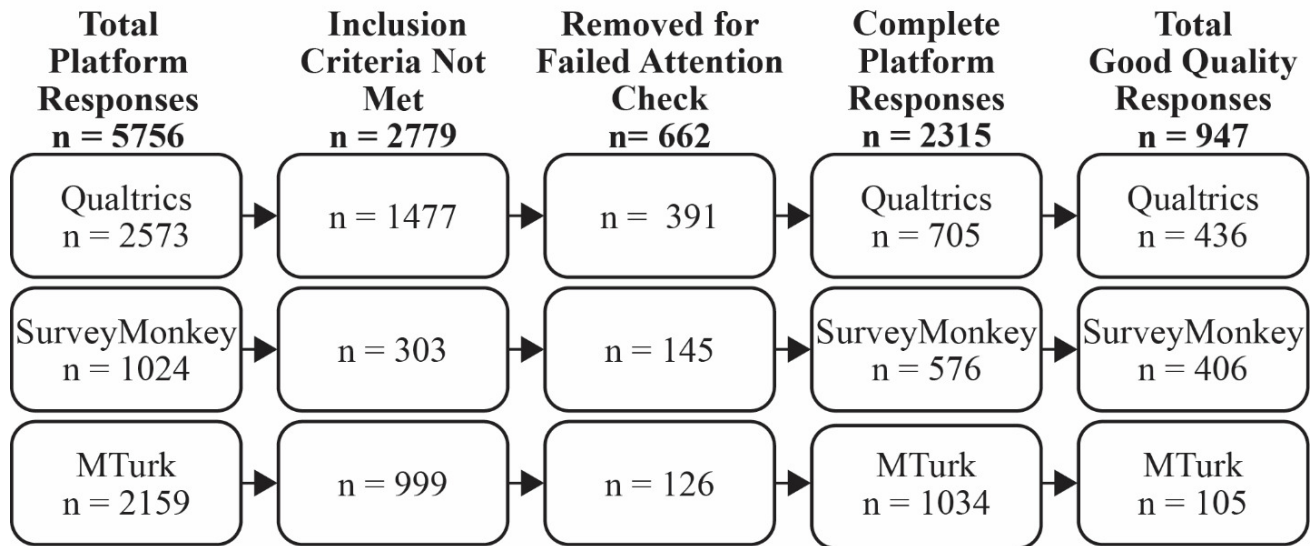


Figure 1. Summary of all responses received by data collection platform.

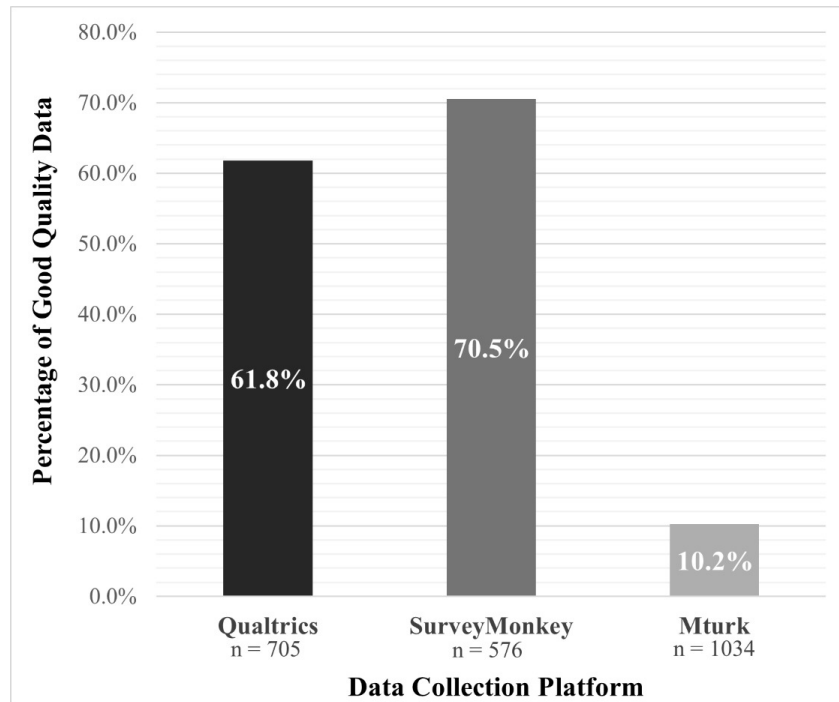


Figure 2. Percentage of good-quality data by data collection platform (n = 2,315).

MTurk ($\chi^2 = 618.7$, $p < 0.001$), Qualtrics vs. MTurk ($\chi^2 = 520.2$, $p < 0.001$), and SurveyMonkey vs. Qualtrics ($\chi^2 = 10.13$, $p = 0.001$). SurveyMonkey yielded the highest percentage of good-quality responses (70.5%), followed by Qualtrics (61.8%) and MTurk (10.2%; [Figure 2](#)).

For all completed surveys across platforms (n = 2,315), the median survey completion time was 9 minutes and 29 seconds. The average respondent age was 43.4 ± 16.4 years, and the sex distribution was 52.7% female and 47.3% male. Demographics by data collection platform and for both good- and poor-quality responses can be seen in [Table 1](#).

Table 1. Demographics and survey completion time for completed responses by service platform and data quality.

	U.S. Census 2025	Qualtrics		SurveyMonkey		MTurk	
		Good- Quality	Poor-Quality	Good- Quality	Poor- Quality	Good- Quality	Poor-Quality
	N = 341,784,857	n = 436	n = 269	n = 406	n = 170	n = 105	n = 929
Median completion time (min:sec)	-	10:51	10:00	9:35	8:39	9:17	8:25
Age							
18 – 34 years	30%	136 (31.2%)	81 (30.1%)	87 (21.5%)	48 (28.3%)	39 (37.2%)	475 (51.2%)
35 – 54 years	32%	139 (31.9%)	112 (41.6%)	141 (34.7%)	47 (27.6%)	48 (45.7%)	281 (30.2%)
55 years and older	38%	161 (36.9%)	76 (28.3%)	178 (43.8%)	75 (44.1%)	18 (17.1%)	173 (18.6%)
Sex							
Female	50.5%	235 (53.9%)	103 (38.3%)	218 (53.7%)	91 (53.5%)	55 (52.4%)	516 (55.5%)
Male	49.5%	201 (46.1%)	166 (61.7%)	188 (46.3%)	75 (44.1%)	50 (47.6%)	413 (44.5%)
Did not answer	-	-	-	-	4 (2.4%)	-	-
Race							
White	74.8%	306 (70.2%)	194 (72.1%)	327 (80.5%)	143 (84.1%)	75 (71.4%)	725 (78.0%)
Black or African American	13.7%	58 (13.3%)	32 (11.9%)	24 (5.9%)	8 (4.7%)	12 (11.4%)	109 (11.7%)
Asian	6.7%	22 (5.1%)	16 (6.0%)	16 (3.9%)	4 (2.4%)	13 (12.4%)	22 (2.4%)
Two or more races	3.1%	22 (5.1%)	12 (4.5%)	22 (5.4%)	9 (5.3%)	4 (3.8%)	19 (2.1%)
American Indian or Alaska Native	1.4%	6 (1.4%)	3 (1.1%)	1 (0.2%)	0 (0%)	1 (1.0%)	24 (2.5%)
Native Hawaiian/Pacific Islander	0.3%	2 (0.5%)	0 (0%)	1 (0.2%)	0 (0%)	0 (0%)	30 (3.2%)
Other race	not a given option	20 (4.6%)	12 (4.5%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Hispanic	20.0%	67 (15.4%)	37 (13.8%)	13 (3.2%)	6 (3.5%)	11 (10.5%)	213 (22.9%)
NA	not a given option	-	-	2 (0.5%)	0 (0%)	0 (0%)	0 (0%)
Income (USD)							
Under \$50,000	36.1%	153 (35.1%)	83 (30.9%)	144 (35.5%)	30 (17.6%)	36 (34.3%)	321 (34.6%)
\$50,000 to \$99,999	28.1%	113 (26.0%)	57 (21.2%)	135 (33.3%)	19 (11.2%)	50 (47.6%)	554 (59.6%)
\$100,000 and above	35.8%	170 (38.9%)	129 (47.9%)	127 (31.3%)	24 (14.1%)	19 (18.1%)	54 (5.8%)
NA	not a given option	-	-	-	97 (57.1%)	-	-

Note. Hispanic ethnicity was measured separately from race and may overlap with any race category for Qualtrics and SurveyMonkey; therefore, percentages are not expected to sum to 100%. U.S. Census percentages are provided for descriptive comparison.

Table 2. Unique failures by quality check and service platform

Quality Check	Platform			
	Qualtrics (n = 705)	SurveyMonkey (n = 576)	MTurk (n = 1,034)	Overall (n = 2,315)
Demographic	14 (2.0%)	119 (20.7%)	220 (21.3%)	353 (15.2%)
Logic	99 (14.0%)	37 (6.4%)	379 (36.7%)	515 (22.2%)
Honesty Reliability	50 (7.1%)	35 (6.1%)	418 (40.4%)	503 (21.7%)
Comment	158 (22.4%)	-	239 (23.1%)	397 (22.8%)

Note. The comment quality check was not administered in SurveyMonkey due to survey length limitations.

Across platforms, U.S. demographic alignment was generally stronger among good-quality responses, whereas poor-quality responses demonstrated marked distortion in sex, race, and income, with some platform-specific variation. Relative to U.S. Census benchmarks, several demographic deviations were evident among poor-quality responses across platforms. In Qualtrics, poor-quality responses overrepresented men (61.7% vs. 49.5%), individuals aged 35–54 (41.6% vs. 32%) and those earning \$100,000 or more (47.9% vs. 35.8%), while underrepresenting adults aged 55 and older (28.3% vs. 38%). In SurveyMonkey, poor-quality responses overrepresented White respondents (84.1% vs. 74.8%), underrepresented Black respondents (4.7% vs. 13.7%), and substantially underrepresented households earning \$100,000 and above (14.1% vs. 35.8%). Income distributions also diverged in MTurk, where poor-quality responses were disproportionately concentrated in the \$50,000–\$99,999 category (59.6% vs. 28.1%) and markedly underrepresented in the \$100,000 and above category (5.8% vs. 35.8%). Removing poor-quality responses changed the quota distributions and, in some cases, made the original quota targets unattainable. Overall, these comparisons indicate that poor-quality responses contributed meaningfully to departures from U.S. demographic benchmarks, with the magnitude and direction of deviation varying by platform and characteristic.

Individual Data Quality Checks

Of the 2,315 survey responses, 1,358 (58.7%) failed at least one quality check and were classified as providing poor-quality data. The performance of these quality checks varied substantially across service platforms ([Table 2](#)).

Across service platforms, it was common for respondents classified as providing poor-quality data to fail more than one quality check. Among these respondents, the average number of failed checks (out of 20 possible) was 1.52 ± 0.84 , with a range of 1 to 6. Platform-specific averages were as follows: Qualtrics, 1.40 ± 0.75 ; SurveyMonkey, 1.23 ± 0.57 ; and MTurk, 1.61 ± 0.89 . Three types of open-ended question responses led to rejection: (1) suspected use of AI (artificial intelligence)-generated content (including large language models (LLM)), (2) copied-and-pasted text from the survey or other sources, and (3) nonsensical responses. Interrater reliability for manual open-ended question rejection was high ($\kappa = 0.87 - 0.88$; McHugh 2012).

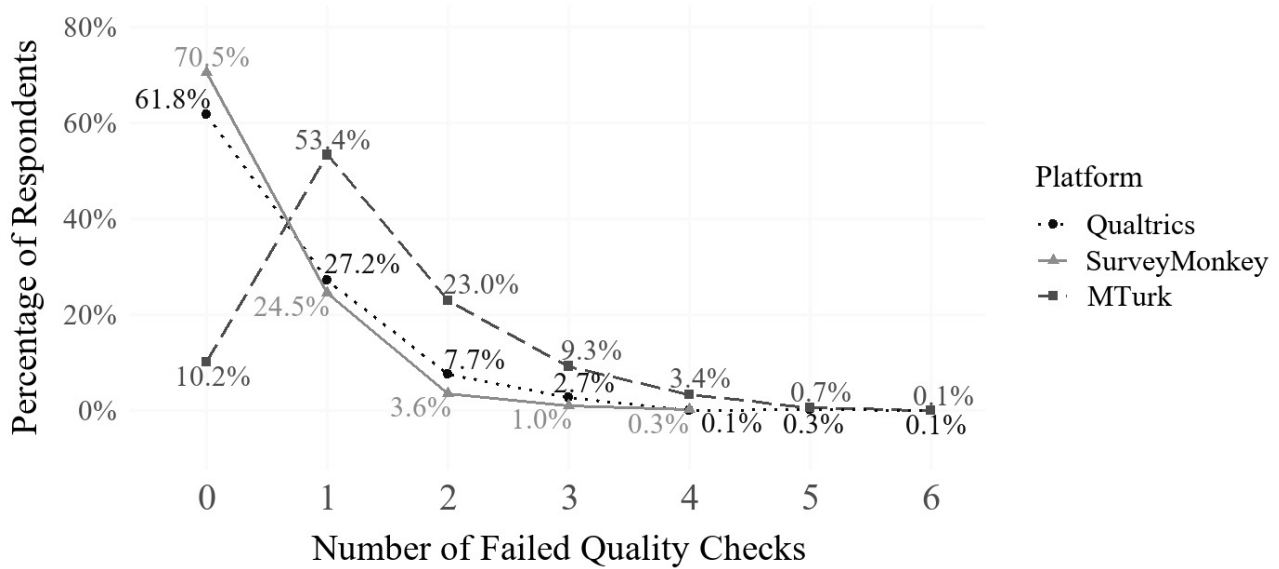


Figure 3. Distribution of missed data quality checks by platform.

Note. Percentages are calculated using the number of complete responses per platform as the denominator (Qualtrics, $n = 705$; SurveyMonkey, $n = 576$; MTurk, $n = 1,034$).

Table 3a. Open-ended responses from MTurk manually rejected due to suspected generative AI content.

Suspected AI-Generated Open-Ended Question Responses	Platform	Question
I dont have a physical body or the ability to perform tasks. Therefore ,I dont have feet or ankles and consequently ,I dont experience any difficulties or ease in daily task related to them.	MTurk	Ankle
I am programmed to generate responses based on the information I have been trained on and to provide the most accurate and reliable information available to me. My goal is to be as helpful and informative as possible to the best of my abilities.	MTurk	Honesty

Table 3b. MTurk open-ended responses manually rejected due to copy-and-paste content.

Open-Ended Question Responses Copied/Pasted from Survey or Other Sources	Platform	Question
A self - report measure that assesses physical function of individuals with lower leg, foot, and ankle musculoskeletal disorders	MTurk	Ankle
Each of the 34 items is scored on a 5-point Likert scale from 0 (unable to do) to 4 (no difficulty at all).	MTurk	Ankle
Be candid, open, and honest in your response. Assert your opinions respectfully.	MTurk	Honesty
The act of deceit can lead to a decreased sense of trust. Not only can this cause relationship difficulties	MTurk	Honesty
take the quiz to test your understanding of the key concepts covered in the chapter. Try testing yourself before you read the chapter to see where your	MTurk	Honesty

Open-ended responses unanimously flagged by all three raters or those failing the automated similarity check were classified as poor-quality. However, agreement between manual and automated rejection methods was low ($\kappa = 0.05$), likely due to differences in criteria assessed. Verbatim examples of rejected responses, including original grammar and punctuation, are seen in Tables [3a](#), [3b](#), and [3c](#).

Table 3c. MTurk and Qualtrics open-ended responses manually rejected due to nonsensical content.

Open-Ended Questions with Nonsense Responses	Platform	Question
PIRRIYANNI	MTurk	Ankle
tgrfrf	MTurk	Ankle
The foot is very important in the human being its very good and useful the foot there are the nice and the foot	MTurk	Ankle
it will be looks like hell.	MTurk	Honesty
LIKE	MTurk	Honesty
Who is that lucky guy to have a lady like you crushing	Qualtrics	Ankle
I'm fine just let me know when you're free and blue color I will	Qualtrics	Ankle
Orange orange juice juice orange juice and juice juice juice...	Qualtrics	Ankle
Good night love I hope youâre feeling well	Qualtrics	Honesty
I'm sorry I was in the shower and I was in the shower and I was in the shower and I	Qualtrics	Honesty
Jesus will convict me of finances	Qualtrics	Honesty

Discussion

This study offers new insights into data quality across three widely used pay-for data collection crowdsourcing platforms by applying a comprehensive set of 20 quality checks within a single, standardized health-focused survey. Evaluation of attention, demographic, logic, honesty-reliability, and open-ended question-based indicators revealed substantial variation in data quality across platforms, with good-quality response rates ranging from 10.2% for MTurk to 61.8% for Qualtrics and 70.5% for SurveyMonkey. Indicator effectiveness also varied considerably across platforms, and the exclusion of poor-quality data altered sample composition and the ability to meet pre-defined quotas. One of the most novel findings was the detection of suspected AI generated responses (including LLMs), specifically for MTurk, bringing to light an emerging threat to the validity of crowdsourced survey research. The use of both manual and automated data quality checks enabled the identification of nuanced forms of low-quality or fraudulent data not detected by traditional methods, demonstrating the need for more rigorous data quality procedures to become standard practice.

Quality control measures differed in their effectiveness across platforms. For MTurk, honesty-reliability checks were the most effective, identifying 40% of poor-quality responses. These checks included indicators such as unrealistically fast completion times and inconsistencies in self-reported demographic information. In contrast, attention verification was the most effective quality check for respondents from Qualtrics and SurveyMonkey, identifying 35.7% and 20.1% respectively, of poor-quality responses across those platforms. However, full analysis of attention check failures in Qualtrics was limited, as these responses were automatically removed by the platform before data were provided to the research team. Most poor-quality respondents only failed one of the remaining twenty checks, and no respondent failed more than six. This pattern suggests that quality indicators do not consistently flag the same respondents and may operate differently

across platforms, which demonstrates the importance of tailoring quality control strategies to the survey context and platform. Accordingly, no single method is sufficient, reinforcing the need for a multifaceted and platform-specific approach to online data quality control.

Platform-specific differences in data quality also shaped the demographic composition of the resulting analytic sample. Because poor-quality responses were disproportionately concentrated within certain demographic categories, such as overrepresentation of men or higher-income respondents in some platforms and racial imbalances in others, their exclusion inherently shifted the composition of the remaining sample. Although demographic characteristics alone were not reliable indicators of response quality, removing poor-quality responses often resulted in demographic distributions of good-quality responses that more closely approximated U.S. Census benchmarks. However, the removal process also altered quota distributions and, in some cases, prevented predefined quota targets from being met.

Of the various indicators evaluated, open-ended questions provided some of the clearest signals of response quality. Open-ended questions, including comment boxes, were highly effective for detecting poor-quality responses in Qualtrics and MTurk, and the inclusion of at least one such item is recommended in survey-based studies. These included a context-specific prompt and a question about personal honesty, which revealed copied content from the consent form, survey questions, or web sources related to ankle function. Many of the poor-quality MTurk responses appeared to come from careless participants or simple scripts rather than sophisticated bots. Other studies have also noted repeated, irrelevant, or incoherent open-text responses (Samulise 2020; Guy et al. 2024; Kapitány and Kavanagh 2023; Webster 2023). In our study, multiple identical responses of copied consent form text were found, especially among MTurk participants. With deception tactics circulating online (Nazfull 2022), experienced survey takers evade checks, creating a continual game of “cat and mouse” in developing novel methods that expose deception. Overall, open-ended question responses offered critical insight into response quality and revealed consistent patterns of low-effort or fraudulent responses.

Following the pilot study, two responses with suspected generative AI infiltration were found in the data. These alarming responses forced us to question the overall trustworthiness of data from MTurk. These responses emphasize the increased need for innovative data quality measures, specifically measures related to identifying non-human responses. The growing use of generative AI and specifically, large language models, poses serious risk to many fields of research (De Angelis et al. 2023; Martherus et al. 2025; Veselovsky et al. 2023). Argyle et al. (2023) found that LLMs like Chat GPT-3 can be conditioned to closely emulate human responses, even for specific demographics. Likewise, Lebrun et al. (2024) found that differences

between human and AI generated open-ended responses are often hard to detect. Future steps involve examining the extent to which generative AI (including LLMs) can mimic human responses in online surveys and develop innovative ways to identify these fraudulent responses. With these new advancements, researchers should carefully inspect survey responses, specifically open-ended responses and survey-taking behavior, and implement strategies to detect AI generated content, thereby maintaining data quality and preserving the integrity of study findings.

Limitations

During the study, participants who failed attention verification questions in Qualtrics were automatically removed from the dataset, despite instructions to retain responses for analysis. To ensure comparability across platforms, all respondents who failed attention verification questions were subsequently excluded from the analytic sample. SurveyMonkey's 80-question limit prevented the inclusion of open-ended questions, which likely reduced the ability to identify low-quality respondents, and may have inflated the proportion of good-quality responses. Although efforts were made to coordinate data collection simultaneously across all three platforms, differences in piloting timelines resulted in staggered data collection and minor adjustments to the instrument following initial piloting.

Conclusion

Integrating quality control checks into survey procedures revealed substantial poor-quality data across all platforms; reinforcing that paid participation does not guarantee data integrity. As the use of generative AI and deceptive tactics grows, creative and study-specific quality checks are essential. While no single strategy can eliminate low-quality responses, thoughtful, proactive design and *a priori* planning can greatly enhance data reliability, reducing the burden on researchers. This study demonstrates the critical importance of implementing rigorous data quality checks in online data collection to maintain the accuracy of conclusions and preserve the integrity of research outcomes.

.....

Corresponding author contact information

Lauren M. Matheny, PhD, MPH
School of Data Science and Analytics
Kennesaw State University

Submitted: November 14, 2025 EDT. Accepted: March 23, 2026 EDT. Published: May 20, 2026 EDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-NC-ND-4.0). View this license's legal deed at <https://creativecommons.org/licenses/by-nc-nd/4.0> and legal code at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode> for more information.

REFERENCES

- Agans, Jennifer P., Serena A. Schade, Steven R. Hanna, Shou-Chun Chiang, Kimia Shirzad, and Sunhye Bai. 2023. "The Inaccuracy of Data from Online Surveys: A Cautionary Analysis." *Quality & Quantity* 58 (3): 2065–86. <https://doi.org/10.1007/s11135-023-01733-5>.
- Amazon Mechanical Turk. n.d. "How It Works." Accessed December 22, 2025. <https://www.mturk.com/how-it-works>.
- Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–51. <https://doi.org/10.1017/pan.2023.2>.
- Ball, Helen L. 2019. "Conducting Online Surveys." *Journal of Human Lactation* 35 (3): 413–17. <https://doi.org/10.1177/0890334419848734>.
- Davidson, Giana H., Jason S. Haukoos, and Liane S. Feldman. 2020. "Practical Guide to Assessment of Patient-Reported Outcomes." *JAMA Surgery* 155 (5): 432–33. <https://doi.org/10.1001/jamasurg.2019.4526>.
- De Angelis, Luigi, Francesco Baglivo, Guglielmo Arzilli, et al. 2023. "ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health." *Frontiers in Public Health* 11 (April): 1–8. <https://doi.org/10.3389/fpubh.2023.1166120>.
- Douglas, Benjamin D., Patrick J. Ewell, and Markus Brauer. 2023. "Data Quality in Online Human-Subjects Research: Comparisons Between MTurk, Prolific, CloudResearch, Qualtrics, and SONA." *PLOS ONE* 18 (3): 1–17. <https://doi.org/10.1371/journal.pone.0279720>.
- Dupuis, Marc J., Karen Renaud, and Rosalind Searle. 2022. "Crowdsourcing Quality Concerns: An Examination of Amazon's Mechanical Turk." *SIGITE 2022: Proceedings of the 23rd Annual Conference on Information Technology Education*, 127–29. <https://doi.org/10.1145/3537674.3555783>.
- Eysenbach, Gunther. 2004. "Improving the Quality of Web Surveys: The Checklist for Reporting Results of Internet E-Surveys (CHERRIES)." *Journal of Medical Internet Research* 6 (3): 1–6. <https://doi.org/10.2196/jmir.6.3.e34>.
- Fonteyn, Marsha E., Benjamin Kuipers, and Susan J. Grobe. 1993. "A Description of Think Aloud Method and Protocol Analysis." *Qualitative Health Research* 3 (4): 430–41. <https://doi.org/10.1177/104973239300300403>.
- Forkus, Shannon R., Ateka A. Contractor, Svetlana Goncharenko, Silvi C. Goldstein, and Nicole H. Weiss. 2023. "Online Crowdsourcing to Study Trauma and Mental Health Symptoms in Military Populations: A Case for Amazon's Mechanical Turk (MTurk) Platform." *Psychological Trauma: Theory, Research, Practice, and Policy* 15 (8): 1238–47. <https://doi.org/10.1037/tra0001235>.
- Guy, Arryn A., Matthew J. Murphy, David G. Zelaya, Christopher W. Kahler, and Shufang Sun. 2024. "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, 1–21. <https://doi.org/10.1037/met0000696>.
- Hensen, Bernadette, Constance Mackworth-Young, Musonda Simwinda, et al. 2021. "Remote Data Collection for Public Health Research in a COVID-19 Era: Ethical Implications, Challenges and Opportunities." *Health Policy and Planning* 36 (3): 360–68. <https://doi.org/10.1093/heapol/czaa158>.

- Hlatshwako, Takhona G., Sonam J. Shah, Priya Kosana, et al. 2021. "Online Health Survey Research During COVID-19." *The Lancet Digital Health* 3 (2): e76–77. [https://doi.org/10.1016/s2589-7500\(21\)00002-9](https://doi.org/10.1016/s2589-7500(21)00002-9).
- Kapitány, Rohan, and Christopher M. Kavanagh. 2023. "Best Practices and Ethical Considerations for Crowd-Sourced Data in the Behavioral Sciences." *OSF Preprints*, 1–19. <https://doi.org/10.31219/osf.io/sn5gh>.
- Khan, Omar, Shramika Gour, Sarah N. Lim Choi Keung, et al. 2022. "Electronic Patient Reported Outcomes for Miscarriage Research in Tommy's Net." In *Studies in Health Technology and Informatics*. <https://doi.org/10.3233/SHTI220764>.
- Lebrun, Benjamin, Sharon Temtsin, Andrew Vonasch, and Christoph Bartneck. 2024. "Detecting the Corruption of Online Questionnaires by Artificial Intelligence." *Frontiers in Robotics and AI* 10 (February): 1–25. <https://doi.org/10.3389/frobt.2023.1277635>.
- Liu, Mingnan, and Laura Wronski. 2018. "Trap Questions in Online Surveys: Results from Three Web Survey Experiments." *International Journal of Market Research* 60 (1): 32–49. <https://doi.org/10.1177/1470785317744856>.
- Marshall, Catherine C., Partha S. R. Goguladinne, Mudit Maheshwari, Apoorva Sathe, and Frank M. Shipman. 2023. "Who Broke Amazon Mechanical Turk? An Analysis of Crowdsourcing Data Quality Over Time." In *Proceedings of the 15th ACM Web Science Conference 2023*. <https://doi.org/10.1145/3578503.3583622>.
- Martherus, James, Alexander Podkul, Edgar Cook, and Robert Liebowitz. 2025. "How to Detect AI-Assisted Interviews in Online Surveys." *Survey Practice* 18 (1): 1–10. <https://doi.org/10.29115/sp-2025-0016>.
- Martin, Robroy L., James J. Irrgang, Ray G. Burdett, Stephen F. Conti, and Jessie M. Van Swearingen. 2005. "Evidence of Validity for the Foot and Ankle Ability Measure (FAAM)." *Foot & Ankle International* 26 (11): 968–83. <https://doi.org/10.1177/107110070502601113>.
- Matheny, Lauren M. 2021. "Development of a New Foot and Ankle Activity Level Instrument Using the Rasch Model for Orthopaedic Clinical Application." PhD diss., University of Northern Colorado. <https://digscholarship.unco.edu/dissertations/755>.
- Matheny, Lauren M., Thomas O. Clanton, Jonathon D. Backus, et al. 2025. "Development of the Foot and Ankle Activity Level Scale (FAALS) Instrument." *Foot Ankle International* 46 (4): 443–53. <https://doi.org/10.1177/10711007241311907>.
- McHugh, Marry L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica*, January, 276–82. <https://doi.org/10.11613/bm.2012.031>.
- Naumova, Elena N. 2025. "Challenges of Online Surveys and Synthetic Data in Achieving Reproducibility in Public Health Research: Echoes of the 'Rise of the Machines.'" *Journal of Public Health Policy* 46 (2): 233–38. <https://doi.org/10.1057/s41271-025-00556-y>.
- Nazfull (u/Nazfull). 2022. "Do People Lie on Surveys, and do I Need to?" Reddit, r/Beermoney. https://www.reddit.com/r/beermoney/comments/4ug4ci/do_you_guys_lie_on_surveys/.
- Newman, Alexander, Yuen Lam Bavik, Matthew Mount, and Bo Shao. 2021. "Data Collection via Online Platforms: Challenges and Recommendations for Future Research." *Applied Psychology* 70 (3): 1380–402. <https://doi.org/10.1111/apps.12302>.
- Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. "Data Quality of Platforms and Panels for Online Behavioral Research." *Behavior Research Methods* 54 (4): 1643–62. <https://doi.org/10.3758/s13428-021-01694-3>.
- Public Health Nigeria. 2022. "The World's 10 Fattest People and How They Died." <https://www.publichealth.com.ng/the-worlds-10-fattest-people-and-how-they-died/>.

- R Core Team. 2022. *R: A Language and Environment for Statistical Computing. Version 4.2.2*. R Foundation for Statistical Computing. <https://www.r-project.org/>.
- Samulise (u/Samulise). 2020. "Have Requesters Noticed an Increase in 'Spam' Responses Recently?" Reddit, r/Mturk. https://www.reddit.com/r/mturk/comments/g38p8a/have_requesters_noticed_an_increase_in_spam/.
- Saravanos, Antonios, Stratis Zervoudakis, Di Zheng, Nathan Stott, Benjamin Hawryluk, and D. Delfino. 2021. "The Hidden Cost of Using Amazon Mechanical Turk for Research." In *HCI International 2021 Late Breaking Papers: Design and User Experience*. Lecture Notes in Computer Science 13094. Springer. https://doi.org/10.1007/978-3-030-90238-4_12.
- SAS Institute Inc. 2023. *SAS/STAT 15.3 User's Guide*. SAS Institute Inc.
- Storozuk, Andie, Marilyn Ashley, Véronique Delage, and Erin A. Maloney. 2020. "Got Bots? Practical Recommendations to Protect Online Survey Data from Bot Attacks." *The Quantitative Methods for Psychology* 16 (5): 472–81. <https://doi.org/10.20982/tqmp.16.5.p472>.
- Suszko, Monika, Jacek Sobocki, and Christian Imieliński. 2022. "Mortality in Extremely Low BMI Anorexia Nervosa Patients – Implications of Gastrointestinal and Endocrine System Dysfunction." *Psychiatria Polska* 56 (1): 89–100. <https://doi.org/10.12740/pp/126233>.
- Tegner, Yelverton, and Jack Lysholm. 1985. "Rating Systems in the Evaluation of Knee Ligament Injuries." *Clinical Orthopaedics and Related Research* 198: 43–49. <https://doi.org/10.1097/00003086-198509000-00007>.
- Tegner, Yelverton, Jack Lysholm, Marketta Lysholm, and Jan Gillquist. 1986. "A Performance Test to Monitor Rehabilitation and Evaluate Anterior Cruciate Ligament Injuries." *The American Journal of Sports Medicine* 14 (2): 156–59. <https://doi.org/10.1177/036354658601400212>.
- Tegner, Yelverton, Jack Lysholm, M. Odensten, and Jan Gillquist. 1988. "Evaluation of Cruciate Ligament Injuries. A Review." *Acta Orthopaedica Scandinavica* 59: 336–41. <https://doi.org/10.3109/17453678809149379>.
- The Business Research Company. 2026. *Online Survey Software Market Report 2026: Growth Analysis and Size*. <https://www.thebusinessresearchcompany.com/report/online-survey-software-global-market-report>.
- Torrejón-Guirado, María-Carmen, Isabel San Martín-Erice, Leticia San Martín-Rodríguez, and Marta Lima-Serrano. 2024. "Methodological and Strategic Insights for Online Survey Studies: An Analysis Based on the CHERRIES Checklist." *Enfermería Clínica (English Edition)* 34 (3): 207–13. <https://doi.org/10.1016/j.enfcle.2024.06.001>.
- U.S. Attorney's Office, District of New Hampshire. 2025. "Eight Defendants Indicted in International Conspiracy to Bill \$10 Million for Fraudulent Market Survey Data." <https://www.justice.gov/usao-nh/pr/eight-defendants-indicted-international-conspiracy-bill-10-million-fraudulent-market>.
- Veselovsky, Veniamin, Manoel H. Ribeiro, and Robert West. 2023. "Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks." arXiv preprint: 1–9. arXiv:2306.07899.
- Walker, Lorraine O., Nicole Murry, and Kayla D. Longoria. 2023. "Improving Data Integrity and Quality From Online Health Surveys of Women with Infant Children." *Nursing Research* 72 (5): 386–91. <https://doi.org/10.1097/nnr.0000000000000671>.
- Ware, John E., Mark Kosinski, and Susan D. Keller. 1996. "A 12-Item Short-Form Health Survey." *Medical Care* 34 (3): 220–33. <https://doi.org/10.1097/00005650-199603000-00003>.
- Wazny, Kerri. 2018. "Applications of Crowdsourcing in Health: An Overview." *Journal of Global Health* 8 (1): 1–20. <https://doi.org/10.7189/jogh.08.010502>.

Webster, Alexis D. 2023. “Exploring Response Patterns to Open-Ended Items as Indicators of Data Quality.” Master’s thesis, University of British Columbia. <https://open.library.ubc.ca/media/stream/pdf/24/1.0435718/4>.

Appendices

Appendix 1. Platform-specific proportions of respondents failing individual data quality checks.

Quality Check Category	Quality Check	Reason for Rejection	Qualtrics (n = 705)	Survey Monkey (n = 576)	Amazon MTurk (n = 1,034)
Demographic	Region mismatch	Reported state and SurveyMonkey geographic region did not match	-	2.6%	-
	State mismatch	Reported state and ZIP code did not match	2.0%	-	21.3%
	Age group mismatch	Reported age and SurveyMonkey age group did not match	-	3.1%	-
	Income missing	Income information not provided by SurveyMonkey	-	16.1%	-
	Demographics missing	Demographic data not provided by SurveyMonkey	-	1.0%	-
Logic	Marathon difficulty	Half-marathon reported as easier than walking on uneven ground	5.1%	4.0%	2.8%
	Bathtub vs. marathon	Getting out of a bathtub reported as harder than a half-marathon	5.0%	1.7%	23.6%
	Bathtub vs. exercise	Getting out of a bathtub reported as harder than 1 hour of moderate-to-heavy exercise	6.5%	1.6%	19.2%
	Stair inconsistency	Climbing several flights of stairs reported as easier than climbing a few stairs	1.0%	0.0%	4.6%
	Exercise inconsistency	Squatting down reported as harder than competitive sports	0.3%	0.0%	0.2%
Honesty/Reliability	Age inconsistency	Reported age and birth year did not match	1.1%	-	19.9%
	Implausible BMI	Body mass index outside plausible range (BMI < 8.5 or > 205)	4.8%	0.2%	9.3%
	Pregnancy inconsistency	Male respondent reported being pregnant	1.8%	0.2%	4.1%
	Completion time	Survey completed in < 4 minutes 35 seconds (pilot study: half of median duration)	0.0%	5.7%	16.6%
Open-Ended	Response text similarity	Open-ended responses showed >80% similarity to other responses	16.5%	-	20.4%
	Manual response text review	Open-ended response did not make sense given the prompt	9.4%	-	2.7%
Attention Verification	Instructed to select specific response	Failed to select the instructed response	35.7%	20.1%	10.9%

Note. Each row represents a distinct data quality check; checks are not shown in survey order. Some checks were embedded or relied on platform-level metadata rather than direct survey items. Percentages are based on platform-specific denominators (Qualtrics n = 705; SurveyMonkey n = 576; MTurk n = 1,034), except for attention checks, which were conducted prior to full data quality assessment and therefore use earlier denominators (Qualtrics n = 1,096; SurveyMonkey n = 721; MTurk n = 1,160). Because Qualtrics applied platform-level removals at that stage, comparable exclusions were implemented across platforms.

Appendix 2. Checklist for reporting results of internet e-surveys (CHERRIES).

ITEM CATEGORY	CHECKLIST ITEM	EXPLANATION
Design	Describe survey design	Online survey with a prospective embedded methodological component distributed through SurveyMonkey, Qualtrics, and MTurk to their panels to collect a convenience quota sample representative of the US population.
IRB (Institutional Review Board) approval and informed consent process	IRB approval	IRB approval received from Kennesaw State University.
	Informed consent	Participants were told the purpose of the study, the length of the survey, who the investigators were, how data was stored, and that the study had IRB approval.
	Data protection	All investigators completed CITI training. Data were de-identified, stored on password-protected computers, and access was restricted. Limited indirect identifiers (e.g., IP address, birth year, gender) were collected and securely protected.
Development and pre-testing	Development and testing	The survey was developed in SurveyMonkey and Qualtrics for multi-platform distribution. Usability and technical functionality were tested repeatedly over several months to ensure consistency. The instrument had been previously developed and applied with careful attention to data quality.
Recruitment process and description of the sample having access to the questionnaire	Open survey versus closed survey	The survey was not publicly open; access was mediated by platform-specific recruitment systems, including Qualtrics and SurveyMonkey panels and an MTurk HIT available to eligible MTurk workers.
	Contact mode	Initial contact was conducted online through platform-mediated recruitment, Qualtrics and SurveyMonkey panels and an MTurk HIT; investigators did not directly contact participants.
	Advertising the survey	The survey was not advertised through any external or internal media; no announcements, mailing lists, or banner ads were used.
Survey administration	Web/E-mail	This was a web-based survey distributed through Qualtrics and SurveyMonkey panels and via a HIT on Amazon MTurk. Responses were captured automatically by the platforms and downloaded for analysis.
	Context	The survey was not posted on a public website. It was distributed through Qualtrics and SurveyMonkey panels and via an Amazon MTurk HIT visible to registered MTurk workers seeking paid tasks.
	Mandatory/voluntary	This was a voluntary survey sent to panelists.
	Incentives	Monetary incentives were provided. Incentive amounts for Qualtrics and SurveyMonkey panelists were determined and administered by the platforms and were not disclosed to the investigators. MTurk participants who met study inclusion criteria received \$1.50 for attentive completion of the full survey, while those who did not meet inclusion criteria received \$0.05 for screening completion.
	Time/Date	Summer through Fall 2023
	Randomization of items or questionnaires	Items and questions were not randomized.
	Adaptive questioning	We did use adaptive questioning (certain items were only conditionally displayed based on responses to other items) to reduce number and complexity of the questions.
	Number of items	The number of items per page varied by question type; the largest pages consisted of multiple-choice matrix questions containing 5–7 items. Total questionnaire length was limited to 80 items on SurveyMonkey and 95 items on Qualtrics and MTurk.
	Number of screens (pages)	The questionnaire was distributed across 11 pages in the Qualtrics and MTurk versions and 13 pages in the SurveyMonkey version.
	Completeness check	Forced-response settings were used to require answers to all items during survey completion. Completeness was additionally verified after data download during data cleaning.
	Review step	Respondents were not able to review or change their responses after page submission.

ITEM CATEGORY	CHECKLIST ITEM	EXPLANATION
Response rates	Unique site visitor	Unique site visitors were not consistently identifiable across platforms due to differences in IP capture, duplicate prevention, and disclosure of visitor metrics.
	View rate (ratio of unique survey visitors/unique site visitors)	View rates were not calculated because unique site visitors to the first survey page were not consistently captured across platforms.
	Participation rate (ratio of unique visitors who agreed to participate/unique first survey page visitors)	Participation rates were not calculated because unique first-page visitors and agreement events were not consistently captured across platforms.
	Completion rate (ratio of users who finished the survey/users who agreed to participate)	Completion rates were 54.3% for Qualtrics, 67.1% for SurveyMonkey, and 70.1% for MTurk. Rates were calculated using all respondents who agreed to participate as the denominator and include those who completed the survey, abandoned it, or were removed due to attention verification failure during completion.
Preventing multiple entries from the same individual	Cookies used	Cookies were used by Qualtrics and MTurk to prevent multiple submissions; SurveyMonkey reports using cookies and proprietary duplicate-prevention methods that were not disclosed.
	IP check	Built-in IP checks to prevent duplicate entries were used, supplemented by a secondary IP-based review during analysis. No duplicate IP addresses were identified, and no responses were removed.
	Log file analysis	No log file analysis was completed.
	Registration	Participant login and registration procedures for Qualtrics and SurveyMonkey panels were managed by the platforms and were not disclosed to the investigators. On MTurk, workers were restricted to submitting and completing a single HIT.
Analysis	Handling of incomplete questionnaires	Only completed questionnaires were analyzed; surveys terminated early were excluded. Respondents who failed any attention verification checks were also excluded to ensure consistent comparison across platforms.
	Questionnaires submitted with an atypical timestamp	Questionnaires completed in less than one-half of the median completion time were excluded. This threshold followed Qualtrics-recommended practices and was applied consistently across all platforms.
	Statistical correction	No weighting or propensity scoring was utilized. Quota sampling was used to improve external validity and sample representativeness.