Connections
Volume 46 Issue 1 · pp. 56–65 · DOI: 10.2478/connections-2025-0004

Connections

My closest relationship is with “Yur Mama”: Data quality in the CHIP50 ego network module

Research Article

Zachary P. Neal*, Jennifer Watling Neal

Department of Psychology, Michigan State University, East Lansing, MI 48824, USA

Received 22 August 2025; Accepted 21 September 2025

Abstract: Data quality issues including problematic responses from non-human bots, malicious respondents, and uncareful respondents can threaten the validity of online survey-based network data. In this study, we describe how we identified and handled potentially problematic responses in the ego network module of the Civic Health and Institutions Project, a 50 States Survey (CHIP50-NET). Specifically, we identified and excluded 4,282 respondents who named non-agent alters or invalid alters or who provided inconsistent responses (17% of the sample). Excluding these potentially problematic respondents from the CHIP50-NET sample had only modest effects on sample demographics. However, these exclusions did yield a prevalence estimate of an uncommon demographic group (adults who do not want children) that was closer to a recent, external benchmark estimate, providing evidence of the validity of our data cleaning efforts. We conclude with implications and recommendations for using the CHIP50-NET dataset specifically, and for cleaning data when conducting large-scale online network surveys more generally.

Keywords: *Big data • Data cleaning • Data quality • Ego network • Name generator*

1. Introduction

Researchers have recognized the importance of examining ego networks using nationally representative surveys for decades. For example, Burt (1984) described how incorporating ego network items into the US general social survey could increase the precision of measurement of individuals’ social contexts and could be used to examine phenomena like segregation and social isolation (e.g., Marsden et al., 2020; McPherson et al., 2006). More recently, the emergence of online survey methods and “big data” have not only yielded new opportunities to collect large-scale survey-based ego network data but have also presented new challenges (Perry et al., 2018).

Among these new challenges, data quality issues can threaten the validity of large-scale, survey-based ego network data. Online surveys are prone to infiltration by nonhuman bots and malicious respondents engaged in survey fraud (Lawlor et al., 2021). In addition, online survey respondents may be more likely to ignore, misread, or misunderstand questions. Such careless responding may be particularly prevalent in online ego network surveys, which typically involve complex questions to measure network composition and structure (Perry et al., 2018).

In this study, we describe how we applied the REAL framework (Lawlor et al., 2021) to identify and handle potentially problematic responses in the ego network module of the Civic Health and Institutions Project, a 50 States Survey (CHIP50-NET). In addition, we validate our application of the REAL Framework and share the resulting cleaned data with other researchers who might be interested in using the CHIP50-NET dataset. We caution network researchers who use large-scale online surveys that automated cleaning methods may not be sufficient and provide recommendations for improving the quality of online large-scale ego network data in future studies.

2. Background

The “Civic Health and Institutions Project, a 50 States Survey” (CHIP50) is a multi-university collaborative effort to collect data from representative samples in every US

* *E-mail: zpneal@msu.edu*

state on a range of topics of interest to, and selected based on proposals submitted by, social scientists and health researchers (Baum et al., n.d.). The project is an extension of the earlier "COVID States Project," which began collecting data in April 2020. Analysis of data generated by these projects has supported research in multiple domains, including political science (e.g., Baum et al., 2024), public health (e.g., Santillana et al., 2024), and mental health (e.g., Perlis et al., 2021).

In May 2024, CHIP50 solicited proposals from researchers for additional questions that would accompany an ego network module designed to capture respondents' ego networks. We submitted a proposal to collect data that would enable research on the ego networks of adults who do not have and do not want to have children, and thus are *childfree*. Our proposal was successful and our questions – three to determine whether the respondent was childfree, and one to determine whether each nominated alter was a parent – were included in the module.

The CHIP50 survey that included the ego network module (CHIP50-NET) was conducted between 30 August, 2024 and 8 October, 2024 by PureSpectrum, a survey data collection firm. The survey included several common features designed to ensure high-quality data, including attention check questions, a Completely Automated Public Turing Test to tell Computers and Humans Apart (CAPTCHA), large language model (LLM) traps, and bot detection services supplied by Google and RelevantID. Following data collection, the CHIP50 team cleaned the data using these tools, as well as checking for both inconsistent and "straightlining" (i.e., answering the same way to multiple consecutive questions) responses. The use of these tools and practices is widely recommended (Mustanski, 2001; Nosek et al., 2002; Pequegnat et al., 2007; Teitcher et al., 2015), and follows the REAL (reflect, expect, analyze, label) framework (Lawlor et al., 2021).

In January 2025, the cleaned data from 25,183 respondents were released to researchers whose proposals had been selected, and thus whose questions were included in the data. Despite the CHIP50 team's extensive efforts to ensure data quality and clean the data, we identified a number of potentially problematic responses that required closer inspection before analysis could proceed. In this work, we use our experience with the CHIP50-NET data as an opportunity to illustrate the data quality issues that can persist even after extensive validation and cleaning efforts. We also use it as an opportunity to describe strategies for identifying and addressing problematic responses in large, anonymous, online ego network surveys.

3. Methods

In the past couple of years, many frameworks have been proposed to confront the challenge of identifying fraudulent online survey participation (Comachio et al., 2025; Johnson et al., 2024; Mistry et al., 2024; Ng et al., 2025). Our review of the CHIP50-NET data was guided by the REAL framework, which is among the earliest and widely-cited, and involves four steps (Lawlor et al., 2021). The first step, *reflect*, asks researchers to consider how problematic responses could occur before beginning data collection, and to design the data collection instrument and practices in ways that minimize these risks. Because the CHIP50-NET data had already been collected, this step was not relevant for us. Therefore, in this section we describe our use of the remaining three steps: expect, analyze, and label.

3.1 Expect

The *expect* step of the REAL framework asks researchers to articulate what they expect to observe in the data. The CHIP50-NET survey was designed to elicit respondents' ego networks using the following name generator question to identify alters: "Think about the five people in your life with whom you have the strongest, closest relationship. Please write their initials below to remind yourself who each person is as you answer the following questions." Therefore, we expected to see responses to this question that consisted of short strings of letters (i.e., initials), roles (e.g., "Mom"), and actual names (e.g., "John").

3.2 Analyze

The *analyze* step of the REAL framework asks researchers to examine responses that were unexpected, with the goal of determining whether they are problematic. Importantly, "what constitutes a suspicious response pattern depends on the context," and therefore, the analyze step may require examining not only the unexpected response itself, but examining the unexpected response in the context of responses to other questions (p. 6). Inspecting responses to the name generator question, sometimes in combination with responses to other questions, revealed three broad categories of potentially problematic responses and respondents.

3.2.1 Non-agent alters

In some cases, respondents answered the name generator by indicating that they have a strong, close relationship with a non-agent, that is, a thing with which one cannot have a strong, close relationship.[1] We identified five unique types of non-agent alters. Determining whether a reported alter is a non-agent can be ambiguous. For each type, we resolved ambiguities by examining the response in the context of other alters reported by the same respondent. In principle, such non-agent alters could be omitted from analysis, while still retaining the respondent and any other alters they reported as valid. However, providing nonsensical responses to the name generator question raised suspicion about a respondent's responses to other questions, and about whether they were a human respondent.

3.2.1.1 Numbers

Some respondents provided solely a number for one or more alters. A number could be used as a shorthand for a specific person. However, closer inspection of respondents' response patterns suggested that this was not typical because the numbers frequently formed a pattern. For example, eight respondents all reported the same ascending integer sequence as their five alters ("1," "2," "3," "4," "5"), and an additional six respondents all reported the same repeating integer sequence as their five alters ("1," "1," "1," "1," "1"). One respondent reported one potentially valid alter who was identified by initials, then four additional alters identified by a repeating sequence ("ct," "11," "22," "33," "44").

3.2.1.2 Companies

There were several instances where respondents listed the names of companies such as "Amazon" or "Hulu" for one or more alters. It is possible that a respondent listed a company name as a nickname or mnemonic device for a specific person. However, response patterns in the data suggest that this is unlikely because multiple respondents listed the same companies in the same order. For example, eight respondents reported that their first alter was "Amazon," their second alter was "Walmart," and their third alter was "Kroger." An additional six respondents reported that their first alter was "Amazon," their second alter was "Walmart," and their third alter was "Hulu."

3.2.1.3 Words

Sometimes, respondents listed words like "strongest," "thirsty," "unfortunately," or "tofu" for one or more alters. When a word was used, we checked the pattern of responses to determine whether it might have been used as a nickname or mnemonic device for a specific person. However, in most cases, the pattern of responses was suspicious. For example, two respondents who used the word "strongest" repeated it for all five of their listed alters. Additionally, when respondents listed a word for one named alter, they tended to list other words for all other named alters.

3.2.1.4 Random text

Some respondents provided phrases or strings of random text for one or more alters. In some cases, rude phrases were used. For example, one respondent listed several questionable alter names including "yur mama," "dead poop," and "smell pee." In other cases, long strings of nonsensical text were listed as alters such as "i'm not sure if you can find a good place" or "we are all going to be a long time." The use of random text suggested respondents were not taking the survey seriously or raised questions about whether respondents were bots.

3.2.1.5 Random keystrokes

In addition to random text, some respondents provided long random keystrokes such as "bzbzbdb" or "nzvhfkf" for one or more alters. In most cases, when a random keystroke was provided for one alter, it was provided for other listed alters as well. Like random text, random keystrokes suggested unserious responding or raised questions about whether respondents were bots.

3.2.2 Invalid alters

In some cases, respondents answered the name generator by indicating that they have a strong, close relationship with an agent whom they may sincerely believe they have a strong, close relationship, but who nonetheless represents an invalid response. We identified seven unique types of invalid alters. Determining whether a reported alter is invalid can be ambiguous. For each type, we resolved ambiguities by examining the response in the context of other alters reported by the same respondent. In principle, such invalid alters could be omitted from analysis, while still

1 Actor network theory adopts a broad view of "agents" and might allow that a human individual could have a strong, close relationship with an inanimate object. However, the ego networks collected by CHIP50-NET were designed in the social network analysis tradition, where this is generally not possible.

retaining the respondent and any other alters they reported as valid. However, in some cases, the responses raised concerns about whether the respondent was taking the survey seriously. In other cases, because subsequent questions in the survey used these alters' names (e.g., does <alter1> know <alter2>?), the questions may have appeared nonsensical to the respondent and thus, may have introduced additional reporting errors.

3.2.2.1 Duplicates

Some respondents entered the same text to identify more than one alter (e.g., alter1 = "son," alter2 = "son"). Although it is plausible that a respondent may have had a strong, close relationship with two different sons, the use of duplicate alter names meant that subsequent questions about the alters were ambiguous and may have led to reporting error. For example, when collecting information about the alters' own relationships to capture the ego network's structure, this respondent would have been asked questions like "Does son know <alter3>," but would have been unable to determine which son is referenced in the question. Similarly, name interpreter questions designed to collect additional information about the alters (e.g., how old is son) could not have been answered accurately.

3.2.2.2 No one

Several respondents entered text intended to indicate that they did not have any (more) close alters (e.g., "no one," "nobody"). It would have been possible to retain these respondents and simply exclude entries intended to state that there were no (more) alters from their ego network data. However, during the survey administration, entries like "no one" and "nobody" were piped into subsequent questions designed to capture the ego network structure. For example, a respondent who listed one named alter before listing "no one" for a second alter would have received questions like "Does <alter1> know no one"? which could have caused confusion and inaccurate responding.

3.2.2.3 Self

In some cases, respondents provided themselves as an alter (e.g., "myself"). In most cases, respondents listed themselves alongside other alter names (e.g., "mom," "brother," "myself"). However, some respondents exclusively listed themselves as alters. Although self-loops can sometimes be of interest in social networks, it is not clear that that they make sense in the context of the CHIP50-NET name generator, which asks about strongest, closest relationships. Additionally, although it may have been possible for a respondent to accurately answer a question like "Does <alter1> know myself," including such ties would have significantly distort the ego network's density and structure.

3.2.2.4 Collectives

Some respondents listed collective entities such as "my children," "my family," "siblings," or "coworkers" for one or more alters. Because collective entities represent more than one individual as a single alter in the network, they are problematic for multiple reasons. First, listing collective entities as alters would have made it impossible to accurately calculate the size and composition of respondents' ego networks. Second, listing collective entities would have made it impossible to accurately delineate the structure of relationships between individuals in the respondents' ego networks.

3.2.2.5 Deities

There were several instances were respondents listed deities as one or more of their alters. For example, 123 respondents named "god" as an alter, usually alongside other alter names. In addition, other respondents listed "christ" or "the holy spirit" as alters. These responses may have reflected respondents' sincere beliefs that they had a strong, close relationship with a deity. However, including these responses would have led to problems in interpreting respondents' network size, composition, and structure.

3.2.2.6 Celebrities

Some respondents named celebrities as one or more of their alters. These celebrities included politicians (e.g., "Joe Biden," "Kamala Harris," "Donald Trump"), performers (e.g., "Katy Perry," "Britney Spears," "Nicki Minaj"), and athletes (e.g., "Lionel Messi"). It is difficult to tell whether these responses represented respondents' sincere beliefs about their relationship with these celebrities or unserious responding. However, it is implausible that respondents actually had strong, close relationships with these individuals.

3.2.2.7 Pets

In some cases, respondents listed a pet as one or more of their alters. For example, 20 respondents listed "dog" and

another 15 respondents listed "cat" as an alter. Although respondents can have strong, close relationships with their pets, most ego network research assumes that alters are human. Therefore, similar to the issues described above for deities, including pets could also lead to problems in interpreting respondents' network size, composition, and structure.

3.2.3 Inconsistent responses

A final category of potentially problematic responses can be identified when a response to one question was inconsistent with a response to another question. There were three opportunities to identify such inconsistent responses. In principle, the variables on which a respondent's responses are inconsistent could be omitted from analysis, while still retaining the respondent and their other responses as valid. However, the inconsistencies suggested that the respondent was not reading and answering questions carefully, and raised concerns about the accuracy of the respondent's answers to other questions.

3.2.3.1 Parent alters

One name interpreter question (included based on our proposal) asked respondents to report whether each named alter "has ever had biological or adopted children." In some cases, respondents named alters who are known to have children. For example, many respondents reported having a strong, close relationship with "mom" or "dad." In these cases, it was possible to verify that the respondent provided a consistent response to this name interpreter question. For example, it was inconsistent to identify "mom" as an alter, and later report that she has never had children.

3.2.3.2 Child alters

One question (included based on our proposal) asked respondents to report whether they "have, or have ever had, any biological or adopted children." In some cases, respondents named alters who are known to be the respondent's own children. For example, many respondents reported having a strong, close relationship with "son" or "daughter." In these cases, it was possible to verify that the respondent provided a consistent response to this name interpreter question. For example, it was inconsistent to identify "daughter" as an alter, and later report that the respondent has never had biological or adopted children.

3.2.3.3 Politics

Respondents were asked to report the strength of their political party affiliation on a seven-point Likert scale ranging from "strong Republican" (1) to "strong Democrat" (7). Additionally, respondents were asked to report the strength of their political ideology on a seven-point Likert scale ranging from "extremely liberal" (1) to "extremely conservative" (7). Notably, the scale of these two questions were reversed – Republican occurred at the low end of the party scale, but conservative appeared at the high end of the ideology scale – which created an opportunity to check for consistency. Although political party and political ideology are distinct, they are typically closely associated. For example, Republicans are typically conservative, while Democrats are typically liberal. Therefore, we treated a response pair as inconsistent if they were five or more scale units separated from the expected pattern. For example, we treated a respondent who reported being a "strong Republican" or "Republican" *and* being "extremely liberal" as inconsistent.

3.3 Label

Finally, the *label* step of the REAL framework asks researchers to decide which responses and respondents should be labeled as problematic. We used both automated scans and manual inspection of the data to identify each instance of each type of potentially problematic response. Respondents with one or more potentially problematic responses of any type were labeled as a potentially problematic respondent. This set of potentially problematic respondents could include non-human bots, malicious human respondents, and otherwise benign human respondents who ignored, misread, or misunderstood questions. However, we did not attempt to distinguish between them. Finally, we asked the CHIP50 team to compute new sampling weights for the reduced data that exclude these potentially problematic respondents.

Excluding all respondents with one of more of any type of potentially problematic response was a particularly conservative approach, or what the REAL framework would describe as a low threshold of suspicion. Some of the responses identified as problematic may be accurate, and some of the respondents flagged as problematic may be valid. Additionally, problematic responses to certain questions may introduce error only for analyses involving those questions, but not for analyses involving responses to different questions by the same respondent. For example, including respondents who provided problematic responses to the name generator question may

introduce error into an analysis of ego networks, but including the same respondents may not introduce error into an analysis of demographic characteristics. Nonetheless, we adopted this conservative approach for three reasons. First, for a given type of potentially problematic response, we were unable to consistently judge which are actually problematic, and therefore, erred on the side of caution. Second, we wanted to generate a single, high-quality CHIP50-NET subsample that could be used by all researchers regardless of the specific items they planned to analyze. Finally, we share the full data and our code, which allows other researchers to adopt more liberal exclusion rules when appropriate for specific analyses.

3.4 Analysis plan

To describe our labeling decisions, we report the number of respondents labeled as having one or more of each type of problematic response. To examine the effect of these labeling decisions, we computed the demographic characteristics of the original raw sample and the updated cleaned sample, testing whether the difference is statistically significant. Finally, to determine whether excluding the flagged respondents yields higher-quality data, we compute estimates of the prevalence of childfree adults (our original research focus in CHIP50-NET) in both the full and cleaned samples, and compare these estimates to previously published estimates. The data and code to reproduce the results reported below, and to use the full or cleaned sample for future analysis, are available at https://osf.io/w5qmp.

4. Results

4.1 Frequency of types of errors

A total of 4,282 respondents, representing 17% of the sample, were labeled as problematic because they provided one or more problematic responses. Table 1 reports the number of respondents who provided one or more of each type of problematic response. Because a respondent may have provided more than one type of problematic response, these counts are not mutually exclusive and sum to more than 4,282.

Non-agent alter responses to the name generator question were relatively rare. For example, only 38 respondents named a company as one or more of their strongest, closest relationships, while 8 entered a series of random keystrokes. Although these types of problematic responses were uncommon, they also raised the most suspicion that the respondent was a bot or malicious human respondent.

Type of response	Number of respondents
Non-agent alters	
Numbers	61
Companies	38
Words	53
Random text	42
Random keystrokes	8
Invalid alters	
Duplicates	2,631
No one	148
Self	57
Collective	310
Deity	129
Celebrity	32
Pet	9
Inconsistent responses	
Parent alters	514
Child alters	130
Politics	770

Table 1. Frequency of problematic responses by type.
Source: Author's contribution.

Invalid alter responses to the name generator question were more common. The most common type of invalid alter was a duplicate, where 2,631 respondents named the same person as their strongest, closest relationship multiple times. Respondents reporting a strong, close relationship with a collective (e.g., "parents," N = 310) or a deity (e.g., "God," N = 144) were also common. It was less common for respondents to report having a relationship with themselves (N = 57), a celebrity (N = 32), or a pet (N = 9).

Finally, many respondents provided responses to multiple questions that were inconsistent with each other. Focusing on responses to the name generator question, 514 respondents named a parent but reported that this person did not have children, while 130 respondents named their own child as an alter but reported not having any children. Focusing on questions about the respondent's political ideology and party affiliation, 770 reported having highly discrepant ideology-party pair (e.g., an extremely liberal Republican).

4.2 Effect on demographic and network characteristics

Table 2 reports the demographic and network characteristics of the unweighted raw (N = 25,183) and cleaned (N = 20,901) samples. The exclusion of certain respondents in the cleaned data had limited impact on the sample's demographic composition, which was not statistically significantly different from the raw sample in terms of income, education, sex, partnership status, or state of residence. The cleaned sample was statistically significantly older (49.25 vs 48.52, t = −4.73, p < 0.01) and more rural (2.97 vs 2.93, t = −2.52, p = 0.01), but the differences were small. Additionally, the cleaned sample had statistically significantly more White respondents (71% vs 68%) and fewer African American respondents (14% vs 15%; χ^2 = 35.88, p < 0.01), but again the differences were small.

The exclusion of certain respondents also had limited impact on the sample's network characteristics, which was not statistically significantly different from the raw sample in terms of size or proportion kin. Respondents in cleaned sample communicated statistically significantly less frequently with their alters (3.44 vs 3.49 on a five-point scale; t = 4.93, p < 0.01), but the difference was very small.

4.3 Validation

The results in Tables 1 and 2 indicate that our application of the REAL framework to CHIP50-NET yielded a smaller but demographically similar sample. But, did this cleaned sample provide higher quality data? To answer this question, we returned to our original focus in the CHIP50-NET data: childfree adults. Childfree adults are adults who do

Characteristic	Raw sample	Cleaned sample	Test
N	25,183	20,901	—
Age	48.52 (16.45)	49.25 (16.59)	t = −4.73, p < 0.01
Income	62356.32 (74385.01)	62202.57 (72801.71)	t = 0.22, p = 0.82
Urbanicity	2.93 (1.58)	2.97 (1.58)	t = −2.52, p = 0.01
Education	3.22 (1.07)	3.23 (1.06)	t = −1.41, p = 0.16
Male	0.46	0.46	t = 1.65, p = 0.1
Single	0.32	0.31	t = 1.36, p = 0.17
Race			
White	0.68	0.71	χ^2 = 35.88, p < 0.01
African American	0.15	0.14	
Hispanic	0.07	0.07	
Asian American	0.05	0.05	
Other	0.02	0.02	
Pacific Islander	0.02	0.01	
Native American	< 0.01	< 0.01	
State (selected)			
CA	0.05	0.05	χ^2 = 20.39, p = 1
FL	0.04	0.04	
NY	0.04	0.03	
TX	0.04	0.04	
GA	0.03	0.02	
Network characteristics			
Size	4.57 (1.03)	4.55 (1.05)	t = 1.38, p = 0.17
Frequency	3.49 (0.93)	3.44 (0.91)	t = 4.93, p < 0.01
Kin	0.85 (0.19)	0.85 (0.19)	t = −0.03, p = 0.98

Table 2. Demographic and ego network characteristics of the unweighted raw and cleaned samples.

Source: Author's contribution.

not have and do not want children. Based on data from the National Survey of Family Growth collected in 2022–2023, 29.4% of non-parents age 18–44 in the United States were recently estimated to be childfree (Neal & Neal, 2025). This estimate offers a benchmark against which estimates from the CHIP50-NET can be validated.

The corresponding estimate from the full CHIP50-NET data was 32.1% (SE = 0.008), which is statistically significantly larger than the benchmark estimate ($z = -2.865$, $p = 0.004$). In contrast, the corresponding estimate from the cleaned CHIP50-NET data was 30.4% (SE = 0.009), which is not statistically significantly different from the benchmark estimate ($z = -1.005$, $p = 0.315$). The fact that the cleaned sample yielded the same estimate as a recent prior study provided evidence of the cleaned sample's validity, and illustrated that using the full sample would have yielded an overestimate of childfree prevalence in the United States.

5. Discussion

The CHIP50-NET survey collected ego network data from a state-by-state representative sample of adults in the United States. When we started examining the data for some planned comparisons of parents and childfree adults, despite the data having already been cleaned by the CHIP team (Baum et al., n.d.), we noticed some unusual responses and response patterns. Applying the REAL framework to further clean the data (Lawlor et al., 2021), we excluded 4,282 (17%) respondents who provided one or more problematic responses. These exclusions did not meaningfully change the demographic composition of the sample, but did yield a more plausible estimate of childfree prevalence when compared to existing benchmark estimates.

Our experience of working with the CHIP50-NET data highlighted two important lessons. First, traditional automated data quality and cleaning methods, such as CAPTCHA or LLM traps, can be inadequate for ensuring quality data from large, anonymous, online samples. Second, when network data are collected and responses to a name generator question are available, these responses provide a rich opportunity to more rigorously clean the data. We identified three broad categories of problematic responses to the name generator question. Non-agent alter responses occurred when respondents reported having strong, close relationships with repeating numerical patterns (e.g., "1," "1," "1," "1," "1"; $N = 61$), companies (e.g., "Amazon," $N = 38$), words (e.g., "thirsty," $N = 53$), random text (e.g., "yur mama," $N = 42$), or random keystrokes (e.g., "bzbzbdb," $N = 8$). Invalid alter responses occurred when respondents reported having strong, close relationships with collectives (e.g., "parents," $N = 310$), deities (e.g., "God," $N = 129$), or celebrities (e.g., "Katy Perry," $N = 32$). Finally, inconsistent responses occurred when respondents reported having a strong relationship with a parent and reported that their parent had never had children ($N = 514$), or reported having a strong relationship with their own child and reported that they had never had children themselves ($N = 130$).

Our application of the REAL framework to clean the CHIP50-NET data is subject to some limitations. First, because it is not possible to determine which potentially problematic responses were actually erroneous, we adopted a conservative approach by excluding all respondents who provided any type of problematic response. Second, given the other data available in the CHIP50-NET survey, we could only evaluate the consistency of alter responses for the subset of respondents who reported having a strong, close relationship with a parent or child. Third, although we were able to identify several categories of problematic responses, it is not an exhaustive list of potentially or ambiguously problematic responses. One noteworthy omission is that no respondents reported having a close relationship with an AI chatbot; this may become increasingly frequent, but its validity as a response to an ego network name generator is ambiguous. Finally, although this approach enables the identification and exclusion of problematic respondents, we were unable to determine whether any given problematic respondent was a bot, a malicious human respondent, or an uncareful human respondent.

Despite these limitations, based on our results, we conclude by two overarching recommendations for network data collection. First, to limit the number of problematic responses that must be identified and removed, and consistent with the "reflect" step in the REAL framework (Lawlor et al., 2021), *we recommend taking steps early to collect high-quality data*. Highly-trained human interviewers following detailed survey scripts can identify invalid responses, and solicit corrections with carefully worded followup probes. In cases where live interviewers are infeasible, automated response validation may be useful to block certain types of invalid responses (e.g., a numeric response to a non-numeric question). However, both live interviewers and automated validators require anticipating potential invalid responses and designing appropriate real-time responses, which may still be defeated by increasingly sophisticated AI-powered survey bots.

Second, researchers analyzing data from large, anonymous, online samples should not rely solely on automated data quality and cleaning tools, and should not assume that data collected by others have been adequately cleaned.

Instead, *we recommend that researchers should also manually review and clean their data, then verify the cleaned data's validity by comparison with available (often demographic) benchmarks*. The richness of network data elicited using name generator questions offers a particularly powerful opportunity to perform such manual checks. Specifically, the nominated alters can be evaluated for validity by asking whether it a thing with which the respondent could have the specified type of relationship. When additional information about the alters is obtained through name interpreter questions, it is also possible to evaluate response consistency for a subset of respondents. While this recommendation can be applied broadly to any network data collected from large, anonymous, online samples, for researchers interested in studying ego networks using the CHIP50-NET data, we recommend using the cleaned and re-weighted dataset described in this study and available at https://osf.io/w5qmp.

Funding information

Authors state no funding involved.

Author contributions

The authors contributed equally to this work.

Conflict of interest statement

Authors state no conflict of interest.

Data availability statement

The data used in this work come from the Civic Health and Institutions Project (chip50.org). The funders and the principal investigators of the CHIP50 Project bear no responsibility for the analyses reported here or the content of the work. The data and code necessary to reproduce the analyses reported in this manuscript are available at https://osf.io/w5qmp.

References

Baum, M., Druckman, J., Lazer, D., & Ognyanova, K. (n.d.). *The Civic Health and Institutions Project, a 50 States Survey (CHIP50) [NSF SES-2241884, SES-2241885, and SES-2241886]*. https://www.chip50.org/.

Baum, M. A., Druckman, J. N., Simonson, M. D., Lin, J., & Perlis, R. H. (2024). The political consequences of depression: How conspiracy beliefs, participatory inclinations, and depression affect support for political violence. *American Journal of Political Science*, *68*(2), 575–594. doi: 10.1111/ajps.12827.

Burt, R. S. (1984). Network items and the general social survey. *Social Networks*, *6*(4), 293–339. doi: 10.1016/0378-8733(84)90007-8.

Comachio, J., Poulsen, A., Bamgboje-Ayodele, A., Tan, A., Ayre, J., Raeside, R., Roy, R., & O'Hagan, E. (2025). Identifying and counteracting fraudulent responses in online recruitment for health research: A scoping review. *BMJ Evidence-Based Medicine*, *30*(3), 173–182. doi: 10.1136/bmjebm-2024-113170.

Johnson, M. S., Adams, V. M., & Byrne, J. (2024). Addressing fraudulent responses in online surveys: Insights from a web-based participatory mapping study. *People and Nature*, *6*(1), 147–164. doi: 10.1002/pan3.10557.

Lawlor, J., Thomas, C., Guhin, A. T., Kenyon, K., Lerner, M. D., UCAS Consortium, & Drahota, A. (2021). Suspicious and fraudulent online survey participation: Introducing the real framework. *Methodological Innovations*, *14*(3), 20597991211050467. doi: 10.1177/20597991211050467.

Marsden, P. V., Smith, T. W., & Hout, M. (2020). Tracking US social change over a half-century: The General Social Survey at fifty. *Annual Review of Sociology*, *46*(1), 109–134. doi: 10.1146/annurev-soc-121919-054838.

McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2006). Social isolation in America: Changes in core discussion networks over two decades. *American Sociological Review*, *71*(3), 353–375. doi: 10.1177/000312240607100301.

Mistry, K., Merrick, S., Cabecinha, M., Daniels, S., Ragan, J., Epstein, M., Lever, L., Venables, Z.C., & Levell, N. J. (2024). Fraudulent participation in online qualitative studies: Practical recommendations on an emerging phenomenon. *Qualitative Health Research*, 10497323241288181. doi: 10.1177/10497323241288181.

Mustanski, B. S. (2001). Getting wired: Exploiting the internet for the collection of valid sexuality data. *Journal of Sex Research*, *38*(4), 292–301. doi: 10.1080/00224490109552100.

Neal, J. W., & Neal, Z. P. (2025). Tracking types of nonparents in the United States. *Journal of Marriage and Family*, *87*(4), 1747–1762. doi: 10.1111/jomf.13097.

Ng, W. Z., Erdembileg, S., Liu, J. C., Tucker, J. D., & Tan, R. K. J. (2025). Increasing rigor in online health surveys

through the reduction of fraudulent data. *Journal of Medical Internet Research*, *27*, e68092. doi: 10.2196/68092.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Eresearch: Ethics, security, design, and control in psychological research on the internet. *Journal of Social Issues*, *58*(1), 161–176. doi: 10.1111/1540-4560.00254.

Pequegnat, W., Rosser, B. S., Bowen, A. M., Bull, S. S., DiClemente, R. J., Bockting, W. O., Elford, J., Fishbein, M., Gurak, L., Horvath, K., & Konstan, J. (2007). Conducting internet-based HIV/STD prevention survey research: Considerations in design and evaluation. *AIDS and Behavior*, *11*, 505–521. doi: 10.1007/s10461-006-9172-9.

Perlis, R. H., Ognyanova, K., Quintana, A., Green, J., Santillana, M., Lin, J., Druckman, J., Lazer, D., Simonson, M. D., Baum, M. A., & Chwe, H. (2021). Gender-specificity of resilience in major depressive disorder. *Depression and Anxiety*, *38*(10), 1026–1033. doi: 10.1002/da.23203.

Perry, B. L., Pescosolido, B. A., & Borgatti, S. P. (2018). *Egocentric network analysis: Foundations, methods, and models*. Cambridge University Press.

Santillana, M., Uslu, A. A., Urmi, T., Quintana-Mathe, A., Druckman, J. N., Ognyanova, K., Baum, M., Perlis, R. H., & Lazer, D. (2024). Tracking COVID-19 infections using survey data on rapid at-home tests. *JAMA Network Open*, *7*(9), e2435442. doi: 10.1001/jamanetworkopen.2024.35442.

Teitcher, J. E., Bockting, W. O., Bauermeister, J. A., Hoefer, C. J., Miner, M. H., & Klitzman, R. L. (2015). Detecting, preventing, and responding to “fraudsters” in internet research: Ethics and tradeoffs. *Journal of Law, Medicine & Ethics*, *43*(1), 116–133. doi: 10.1111/jlme.12200.