HHS Public Access

Author manuscript

Comput Human Behav. Author manuscript; available in PMC 2025 August 01.

Published in final edited form as:

Comput Human Behav. 2024 August ; 157: . doi:10.1016/j.chb.2024.108253.

The Cyborg Method: A Method to Identify Fraudulent Responses from Crowdsourced Data

Matthew Price[a,*], **Johanna E. Hidalgo**[a], **Julia N. Kim**[a], **Alison C. Legrand**[a], **Zoe M.F. Brier**[a], **Katherine van Stolk-Cooke**[b], **Amy Hughes Lansing**[a], **Ateka A. Contractor**[c]

[a]Department of Psychological Science, University of Vermont, 2 Colchester Avenue, Burlington, Vermont, 05405, USA

[b]State University of New York Geneseo, Dept of Psychology, USA

[c]University of North Texas, Dept of Psychology, USA

Abstract

Crowdsourcing is an essential data collection method for psychological research. Concerns about the validity and quality of crowdsourced data persist, however. A recent documented increase in the number of invalid responses within crowdsourced data has highlighted the need for quality control measures. Although a number of approaches are recommended, few have been empirically evaluated. The present study evaluated a Cyborg Method that used automated evaluation of participant meta-data and a review of short answer responses. Two samples were recruited – in the first, the Cyborg Method was applied after data collection to gauge the extent to which invalid responses were collected when *a priori* quality controls were absent. In the second, the Cyborg Method was applied during data collection to determine if the method would proactively screen invalid responses. Results suggested that Cyborg Method identified a substantial portion of invalid responses and both automated and human evaluation components was necessary. Furthermore, the Cyborg Method could be applied proactively to screen invalid responses and substantially reduced the per participant cost of data collection. These results suggest that the Cyborg Method is a promising means by which to collect high quality crowdsourced data.

Keywords

Crowdsourced data; quality control; mechanical turk; psychology

[*]Corresponding Author Matthew Price, Department of Psychological Science, University of Vermont, 2 Colchester Avenue, Burlington, Vermont 05405, Matthew.Price@uvm.edu.

Credit Author Statement

Study conception and design were performed by MP, JK, and ACL. Figure creation and Table Creation were performed by MP, JEH, and JK. The first draft was written by MP and all authors provided substantive feedback on several drafts prior to submission. All authors read and approved the final manuscript.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

1. Benefit of Crowdsourced Data

Crowdsourcing platforms such as Amazon's Mechanical Turk and Prolific are widely used to collect data for psychological research (Buhrmester et al., 2018). The number of published studies that used these platforms has increased 10-fold in the past decade (Mellis & Bickel, 2020). There are several reasons for this rapid increase: allow for data collection via self-report surveys and experimental methods, ease of recruitment, access to hard-to-reach populations, and relatively low costs (Arditte et al., 2016; Shapiro et al., 2013; Strickland & Stoops, 2019; Weber, 2021). There are perennial concerns, however, regarding the validity of crowdsourced data (Chandler et al., 2020; Shapiro et al., 2013). These concerns have grown in recent years due to several documented cases in which large portions of responses in crowdsourced datasets were deemed invalid (Chandler et al., 2020; Dennis et al., 2019). Given the potential benefits of crowdsourced data collection, countermeasures are needed.

1.1 Concerns about Crowdsourced Data

Concerns about the quality of crowdsourced data have grown with recent reports of data contaminated with invalid responses (Chandler et al., 2020; Dennis et al., 2019). The invalid responses were detected when the investigators were conducting the primary outcome analyses and obtained markedly different results than what was hypothesized. They carefully inspected the data and discovered several responses in which self-report measures were answered seemingly at random. These invalid responses significantly differed from valid responses across multiple measures of psychopathology. Furthermore, well-established associations between psychopathology and other constructs, such as social support, differed between valid and invalid responses as well. Without proper detection, these invalid responses would have led to incorrect results and conclusions being drawn.

The inclusion of invalid responses also placed a substantial financial burden on these studies. Most studies use *a priori* power analyses to determine a target sample size to be recruited. This sample size dictates the budget required to complete the study. When a portion of the compensated responses are invalid, however, the target sample size may not be reached. Thus, valid conclusions from the collected data cannot be drawn (Chandler et al., 2020). Given the potential risks of collecting and using invalid responses from crowdsourced data, there is a need for quality control methods to maintain the feasibility of such research projects.

Although basic quality control methods to screen for such invalid responses are integrated into many survey platforms (Qualtrics XM, 2021), increasingly sophisticated strategies are used to circumvent these protections. These strategies include the use of Virtual Private Networks (VPNs) or Virtual Private Servers (VPSs) and bots (Dennis et al., 2019). VPNs allow an internet connection to appear as if it has originated from a different location. Many crowdsourced studies restrict access to those within a particular country (Boas et al., 2020). The user's country is determined using the device's Internet Protocol (IP) address. VPNs alter the IP address, so the device appears as if it is in a location allowed by the study, thus giving study access to individuals outside of the target country. Bots are the use of

automated software to complete surveys or experiments in a manner that mimics a real individual. Bot behavior may include automatically completing survey responses at a pre-specified rate, selecting responses based on a specified pattern, and inserting short written responses or narratives that are gleaned from internet search engines. These strategies allow studies to be completed with invalid responses that require advanced methods to detect.

1.2 Current Quality Control Methods

Several recent investigations have recommended quality control methods. One method is IP evaluation (Aguinis et al., 2021; Chandler et al., 2020; Dennis et al., 2019, 2019; MacInnis et al., 2020). IP evaluation uses 3rd party services that compare an IP address against a list of addresses known to be used by VPNs. VPNs will often assign users to a set of addresses affiliated with a given location, making it possible to flag them as suspicious. This evaluation can be done post-hoc if IP addresses were collected or in real time via an application programming interface (API) embedded within the survey platform. A recent comparison between those who were flagged as using a VPN and those who were not demonstrated that these services can accurately identify invalid responses (Dennis et al., 2019). In this study, individuals who used a VPN were more likely to fail a question that requested a specific answer (e.g., "For this question, please answer *none of the above*"), provided lower quality responses to short answer questions, and were more likely to provide nonsensical responses to short answer questions.

Despite the promise of IP evaluation to identify fraudulent respondents, it has several limitations (Dennis et al., 2019). First, IP evaluation is performed by 3rd party companies who use proprietary methods that are not subject to independent evaluation by researchers. Second, IP address identification occurs at the location of the device. It may incorrectly flag multiple users in a similar location, such as multiple individuals who complete the study while in the same location, as fraudulent. Furthermore, the way internet service providers (ISPs) assign IP addresses may lead an IP evaluation service to label an IP as problematic incorrectly. For example, an ISP can use a dynamic IP address that changes routinely such that it can be difficult to determine if a given user is using a VPN. Finally, certain ISPs may use network address translation (NAT) that allows multiple users in a similar location to share an IP address. These IP addresses are more likely to be flagged as VPNs because of the number of users associated with a single address. These limitations suggest that additional measures are needed beyond IP address evaluation.

A second recommendation for identifying invalid responses is embedding items that require a specific response and then determining the number of responses that met these criteria. The most common methods are "attention-check" or "knowledge-check" items. Attention-check items ask for a specific response (e.g., "Choose option 3") or ask about whether fictional events have occurred (e.g., "Have you conducted business in Wakanda?"). A knowledge-check item is one that asks for information that would only be known by a member of a target population (e.g., specific military information for a veteran sample) (Weiss et al., 2021). This approach has been used to identify potential fraudulent responses as well as those who are inattentive (Natesan Batley et al., 2021; Shapiro et al., 2013; van Stolk-Cooke et al., 2018). This method also has limitations. First, these items add to

the length of the survey. Second, items that ask about fictional information may confuse participants (e.g., unsure if Wakanda is a real location). The additional time spent on these items may contribute to fatigue or inattention. Indeed, some have cautioned against using these items because they may compromise the data obtained from valid participants (Vannette, 2017). Third, the respondents who provide invalid data may be aware of these items and correctly answer these questions. In one study, half of the invalid responses answered attention-check questions correctly (Dennis et al., 2019). Thus, the use of these questions in isolation is likely insufficient (Agley et al., 2021).

1.3 The Cyborg Method as a Quality Control Method

A combination approach that integrates an automated evaluation coupled with a researcher's review of responses is recommended (Chandler et al., 2020). This Cyborg Method, in which automated and human evaluation of the data are combined, allows the strengths of each method to offset the limitations of the other. An automated IP evaluation can rapidly identify invalid responses, thus reducing researcher burden. Those that pass this review then have a set of responses reviewed by investigators. Prior work has suggested that short answer responses are ideal for this purpose (Dennis et al., 2019). Short answer questions can be brief, problematic responses are easily identified, and are difficult for a bot to falsify. It is recommended that the short answer questions ask for a brief description of a personal event or belief (Dolan et al., 2020), which is simple for a valid participant to complete, but would pose a significant challenge for a bot. For example, in conducting a study on individuals who have experienced traumatic events, respondents could describe their most salient trauma. This line of questioning is part of several well-established measures such as the Life Events Checklist (Weathers et al., 2013). This combined approach of automated review of IP addresses and review of written responses has the potential to efficiently increase data quality.

To-date, only a single study has evaluated the benefits of this combined approach (Agley et al., 2021). A large cohort of respondents were randomized to four conditions: no control methods, a bot/VPN evaluation that asked participants a fact-based question that was unrelated to the study, an attention check question that asked about fictional events, and a combination arm that included both the bot/VPN check and the attention check. All arms completed measures about psychopathology as well. It should be noted that the VPN evaluation method was more consistent with a knowledge-check method as opposed to the IP evaluation described above. The results showed that the arm without any control methods had significantly higher scores on all measures of psychopathology. This arm had more invalid responses as well. The combination arm had the greatest proportion of valid responses. There were several limitations, however. First, all quality control methods were based on self-report information including the VPN evaluation. Second, participants were compensated with less than the current market rate for completing crowdsourced studies. Finally, recruitment occurred under a target sample size rather than according to a set budget, which makes it difficult to determine the extent that using this approach reduced overall study costs. Given these limitations and the overall lack of empirical work evaluating quality control methods for crowdsourced data, more work is warranted.

1.4 The Current Study

The present study evaluated the utility of a Cyborg Method that involved automated IP evaluation and a human review of short answer responses. The primary aim was to determine if these methods improved the quality of crowdsourced data. This aim was evaluated by comparing responses identified as valid to those as invalid. It was hypothesized that both components of the Cyborg Method would be needed as opposed to just one component. A secondary aim was to determine the extent to which these methods reduced study costs. Finally, the classification of valid and invalid responses between the Cyborg Method and the attention-check method was compared based on the popularity of the attention check method. It was hypothesized that the Cyborg approach would correctly identify a greater proportion of responses as invalid than attention-check questions alone. The study utilized a sample of individuals who reported a history of exposure to highly stress and traumatic events as well as measures of mental health. This target sample was selected due to prior studies showing that crowdsourcing is capable of recruiting valid samples with such histories (Engle et al., 2020; van Stolk-Cooke et al., 2018). Furthermore, psychometrically valid measures to assess relevant constructs that were sensitive to valid and invalid responses in prior work were available (Chandler et al., 2020; Shapiro et al., 2013). These methods allowed for invalid and responses to be detected and evaluated.

2. Methods

2.1 Overview

The current study included two samples recruited sequentially to avoid participant overlap. For Sample 1, the Cyborg Method was applied after data collection and all participants were compensated. For Sample 2, the Cyborg Method was enforced during data collection. The budget for Sample 1 was $4,900 and for Sample 2 was $1,331. The budget for Sample 1 was larger to allow for sufficient valid and invalid respondents to be collected. Data collection continued until the budget for each sample was exhausted. All study procedures were approved by the local Institutional Review Board.

2.2 Recruitment

Recruitment occurred via Amazon’s Mechanical Turk. Sample 1 data was collected from January-March 2021. Sample 2 data was collected from April-May 2021. For both waves, Human Intelligence Tasks (HITs) were made available to those with a HIT approval rate >95% and United States residency. HITs advertised a research study that examined reactions to stressful events. Inclusion criteria for both samples were having experienced directly or witnessed in person a traumatic event that met Criterion A for a posttraumatic stress disorder (PTSD) and residing in the United States (American Psychiatric Association, 2013). The requirement of a traumatic event that met Criterion A was used to collect a personalized writing sample about a traumatic event. Participants were compensated $4.00 for successful completion of the HIT. Amazon charged $0.80 per participant as a service fee. Thus, the base cost per participant was $4.80.

2.3 Measures

2.3.1 Patient Health Questionnaire-8 (PHQ-8; Kroenke, Strine, Spitzer, Williams, Berry, Mokdad, 2009): The PHQ-8 is an 8-item self-report measure that assesses depression symptoms experienced over the past two weeks on a 0–3 point Likert scale. The PHQ-8 is adapted from the PHQ-9 and removes an item on suicidal ideation. Total scores are obtained by summing all items such that scores range from 0–24 with higher scores indicating greater depression. The PHQ-8 has excellent psychometric properties in a wide range of populations, including those with significant trauma histories (Christensen et al., 2017). The PHQ-8 has been shown to be reliable, have strong correlations with other measures of depression (construct validity), weak relations to theoretically distinct constructs (discriminant validity), and excellent sensitivity to change in samples of those with a range of comorbid conditions (Beard et al., 2016).

2.3.2 Life Events Checklist for the DSM 5 (LEC-5; Weathers, Blake, Schnurr, Kaloupek, Marx, & Keane, 2013): The LEC-5 is a 17-item self-report measure that assesses exposure to potentially traumatic events across one's life span. The extended version of the LEC-5 was used in the current study, which asked for a written description of the worst event in part 2. Items from the LEC are used to describe the presence/absence of trauma history with possible scores ranging from 1 to 17 indicating the experience a single type to every type of trauma. Participants in the current study were those who endorsed at least one event as happening to them directly. The LEC has been shown to be temporally stable with repeated assessment, associated with other validated measures of trauma history, and associated with measures of psychological distress that are commonly associated with elevated trauma history (Gray et al., 2004).

2.3.3 PTSD Checklist for the DSM-5 (PCL-5; Weathers, Litz, Keane, Palmieri, Marx, & Schnurr, 2013): The PCL-5 is a 20-item self-report measure that assesses PTSD symptoms experienced over the last month on a 0–4 point Likert scale. The PCL-5 was anchored to the traumatic event specified in the LEC-5. Total scores are obtained by summing all items such that scores range from 0–80 with higher scores indicating greater PTSD. The PCL-5 has excellent psychometrics with a recent systematic review suggested that it has excellent internal consistency, test-retest reliability, and construct validity (Forkus et al., 2022).

2.3.4 Quality Control Measures: Three quality control measures implemented: (1) Automated IP Evaluation, (2) Attention Items, (3) Short Answer Review. Information on how to implement these methods is detailed here: https://www.crestresearch.org/research/cyborg-method.

2.3.4.1 Automated IP Evaluation: Automated IP evaluation services determined if the IP address used a VPN or a BOT network. Two external services were used - IPhub (https://iphub.info/) and IPQualityscore (https://www.ipqualityscore.com/). These services maintain an active blacklist of IP addresses associated with problematic internet behavior. They were selected because they performed well in prior studies (Dennis et al., 2019), have API's that

can be integrated into Qualtrics for automatic IP evaluation (see supplemental materials), and offer free accounts.

2.3.4.2 Attention-Check Items: Six attention items were placed at equidistant points that required a specific response (e.g., For this question, please select “Quite a bit.”). A valid response based on the attention check items required at least 5 correct responses.

2.3.4.3 Short Answer Review: The written trauma description in the LEC-5 Part 2 was reviewed by trained research assistants. Instructions were to provide 2–3 sentences describing the most distressing traumatic event while not including identifying information.

2.4 Procedure

Please see the supplemental materials for detailed information on how to implement these methods (Figure 1).

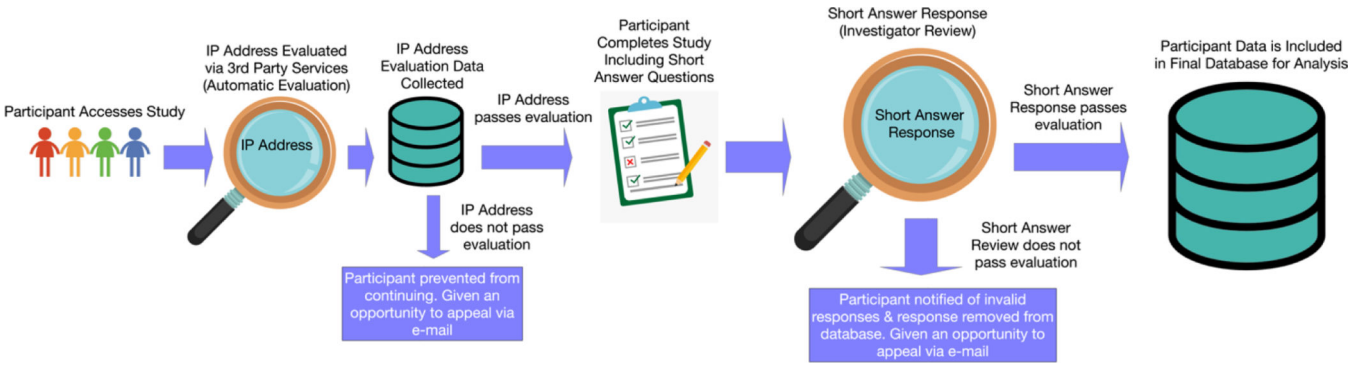

Figure 1.

2.4.1 Cyborg Method Overview: After selecting the HIT, participants were informed that their IP address would be reviewed for the presence of a VPN. IP information was sent to IPhub and IPqualityscore via an integrated API. Information about the likelihood of bots, VPNs, and other VPN behavior were collected automatically. Those who failed the IP evaluation were discontinued. Participants who were discontinued were given the opportunity to appeal this decision. The appeal process involved contacting the research team via e-mail and providing information as to their physical location and confirming that they did not use a VPN or Bot. Participants who passed the IP evaluation were allowed to continue and complete all remaining surveys, including the LEC that contains the Short Answer Response for review. After completing all surveys, the Short Answer Review was conducted by the investigators. Responses that pass this review were considered valid. Those that did not were considered invalid.

2.4.2 Sample 1.—For sample 1, all participants were allowed to complete the study and were compensated. That is, participants were not excluded for failing to pass the IP evaluation or the Short Answer Review. However, all IP-related data were collected.

2.4.3 Sample 2.—Sample 2 data collection proceeded similar to that of Sample 1, except that quality control measures were enforced during data collection. That is, those with an invalid IP were discontinued. Upon completion, two trained research assistants conducted the Short Answer Review of the LEC-5 and determined the validity of the short answer response. Agreement among the research assistants was 100%. Only valid responses were compensated. Participants who were identified as invalid were notified and given the opportunity to appeal this decision. Seven individuals appealed successfully and were included in the final dataset for Sample 2.

2.5 Data Analysis

Analyses for the current study aimed to describe the proportion of invalid responses at each stage of the Cyborg Method. Proportions of the sample at each stage of review were documented for Sample 1 and 2. For Sample 1, comparisons were conducted to determine if there were differences in PCL-5, PHQ-8, and LEC Trauma types for each component of the Cyborg approach using a 2 × 2 (IP Evaluation x Short Answer Review) ANOVA. Additional 2 ×2 (Cyborg x Attention Check) ANOVAs were conducted to determine if scores on the PCL-5, PHQ-8, and LEC Trauma types differed between responses identified as valid on the Cyborg Method and Attention Check method. Finally, the economic impact of using the Cyborg Method was evaluated by cost per response for Sample 1 and Sample 2. This value was obtained by dividing the total "yield", defined as the number of valid responses, by the total budget available for each method of recruitment.

3. Results

3.1 Proportion of Individuals Identified by Each Part of the Cyborg Method

For Sample 1, $n = 4427$ users accessed the HIT (Figure 2). Of these responses, $n = 2942$ did not endorse an event on the LEC-5 and thus were discontinued for not meeting the inclusion criteria. Of note, a substantial portion of these responses ($n = 1815$, 61.7%) had invalid IP's. Of the 1485 remaining, approximately a third discontinued voluntarily ($n = 474$, 31.9%). The 1011 responses from participants who endorsed experiencing a traumatic event on the LEC-5 and who completed the study were evaluated via the Cyborg Method. Among this group, $n = 324$, 32.0% were invalid based on their IP address and their short answer response, $n = 431$, 42.6% had valid IP addresses, and $n = 520$, 51.4% provided valid short answer responses. A subset, $n = 264$, 26.1%, had a valid IP address and passed the Short Answer Review (Supplemental Table 1).

3.2 Comparison of Valid and Invalid Responses According to the Cyborg Method

Follow-up analyses were conducted to determine if there were differences in PCL-5, PHQ-8, and LEC Trauma types for each component of the Cyborg approach (Table 1). On the PCL-5, there was a significant IP evaluation x Short Answer Review interaction for the PCL-5, $F(1, 1006) = 19.77$, $p < .001$. Post-hoc tests suggested that the responses of those who were valid based on IP evaluation and Short Answer Review had the lowest PCL-5 scores compared to the other groups ($p's < .001$; Supplemental Table 2; Figure 2a). Furthermore, the mean score of the invalid responses was close to the center of the possible range of scores on the PCL-5 ($M = 43.64$, Center = 40). This mean score is notable as it

is likely to be obtained when choosing answers on the measure at random. A similar set of findings emerged for the PHQ-8. There was a significant IP evaluation x Short Answer Review interaction, $F(1, 1006) = 6.15$, $p = .013$ (Supplemental Table 2; Figure 2b). The mean PHQ-8 score of the invalid responses was also at the center of the range of scores ($M = 13.21$, Center = 13.50). Finally, the number of traumatic event types that were endorsed were compared across the evaluation methods. There was a significant IP evaluation x Short Answer Review interaction, $F(1, 1006) = 9.08$, $p = .003$. Those who were determined to be valid by both methods endorsed fewer traumatic event types than the other groups (p's < .001; Supplemental Table 2; Figure 2c). Of note, 221 invalid responses endorsed every type of traumatic event on the LEC-5.

Short answer responses for each of the four groups were reviewed (Supplemental Table 1). Those who passed IP evaluation but not Short Answer Review had responses that were nonsensical or copied from social media. Those who passed Short Answer Review but not IP evaluation had short answer responses that were overly vague (e.g., "I broke my wrist") or were entered with exactly the same text by multiple participants, which indicates they were invalid. These data suggest that both components of the Cyborg method, IP evaluation and Short Answer Review, were necessary (Supplemental Table 3).

3.3 Comparison of the Cyborg Method to the Attention Check Method

Responses were then compared between the Attention Check and Cyborg Methods. Of the 1011 responses who endorsed an event on the LEC, $n = 890$ (88.03%) completed all 6 attention check items correctly. A substantial portion of this subset ($n = 655$, 64.8%) completed all 6 attention check items but did not pass the validity checks of the Cyborg Method. Alternatively, only 18 responses (1.8%) were considered valid by the Cyborg Method, but not the Attention Check method. A 2 × 2 ANOVA (Cyborg x Attention Check) comparing PCL responses showed a significant interaction, $F(1, 1006) = 5.72$ $p = .017$. Post-hoc comparisons suggested that within the Attention Check group, those with valid Cyborg Method responses had significantly lower scores (diff = 17.28, $p < .001$, Figure 2). For PHQ-8 scores, there was a main effect for the Cyborg Method, $F(1, 1006) = 4.06$, $p = .044$. Post hoc tests suggested that those who were valid according to the Cyborg Method were significantly lower than those who were not (diff = 3.57, $p < .001$). There was no main effect for the Attention Check group and no significant interaction. There was a significant main effect for the Cyborg Method for the number of trauma types, $F(1, 1006) = 45.56$, $p < .001$ (Figure 3). Again, this group endorsed fewer event types than those in the other groups. Taken together, these results suggest that the Cyborg Method was more likely to detect valid responses than the Attention Check method.

3.4 Economic Impact

The economic impact of the Cyborg Method was then evaluated. For Sample 1, 264 participants were determined to be valid by the Cyborg Method. A subset ($n = 32$, 12.1%) provided a valid short answer response that did not describe a traumatic event that would meet Criterion A for PTSD. Thus, a valid sample of 232 participants was obtained for $4,900, resulting in a cost per participant of $21.12. This is an excess cost of ($21.12 - $4.80) $16.82 per participant. Sample 2 was then recruited using a budget of $1,331

and proactively applying the Cyborg Method (Table 2; Figure 4). Invalid responses were rejected when they were identified and not provided compensation. For this sample, 301 were considered valid by the Cyborg Method. Of these responses, 30 reported a stressful event that did not meet Criterion A for PTSD. The final sample of 271 participants had a cost of $4.91 per participant, resulting in an excess cost of $0.11 per participant. Thus, the cyborg method substantially reduce cost per participant. The distributions for the Sample 1 and Sample 2 responses that passed the Cyborg Method had a high degree of overlap (Supplemental Figure 1).

4. Discussion

4.1 Advantages of Multimethod Quality Controls

The current study demonstrated the utility of a multimethod strategy to improve the quality of crowdsourced data that we are referring to as the Cyborg Method for its use of automated and human review of data. Prior work has shown that, when effective quality control measures are used, data collected via crowdsourcing is comparable to that of clinical samples collected via traditional means (Arch & Carr, 2017; Engle et al., 2020; Morgan & Desmarais, 2017; Shapiro et al., 2013; van Stolk-Cooke et al., 2018). The Cyborg Method is efficient in that it reduces investigator effort and financial costs, while increasing data quality. It also can be applied to previously collected data in which IP addresses and written responses were collected. Thus, it is a viable strategy for improving the quality of previously collected data as well as future data collection.

The results of the present study highlight the need for multimethod evaluation of responses from crowdsourced data. Despite the high agreement between the IP evaluation and Short Answer Review, there were large subsets of responses that passed one evaluation but not the other. Those that passed the IP evaluation but failed the Short Answer Response Review may have used IP addresses that were newly acquired. Thus, the IP evaluation services were not aware of their fraud potential. Alternatively, several seemingly valid written responses came from invalid IP addresses. These written responses were vague descriptions of a trauma that appeared multiple times in the dataset. Detection of such responses by a researcher is challenging, but the IP evaluation service flagged them as problematic. Thus, it is highly recommended that the multimethod used in the present study be applied in future research using crowdsourced data.

4.2 Benefit of the Cyborg Method

The proposed Cyborg Method has several advantages. First, it reduces researcher burden such that the do not need to scour large, crowdsourced datasets for invalid responses in that only a subset of responses with valid IP addresses needs to be considered for the Short Answer Review. The data from the Sample 1 showed that a substantial portion of participants with invalid IP addresses also provided invalid written responses. Second, the approach is flexible in its implementation. This method can be applied as needed during data collection or after study completion if all data needed to implement the method were collected. Thus, there is reason to believe that continued use of this method would result in similar samples, especially among trauma-exposed individuals.

The results of the present study also demonstrate the economic benefit of the Cyborg Method. The cost to recruit participants in Sample 1, where quality control measures were implemented after the data collection was completed, was four times that of Sample 2. This increased cost highlights the challenges of conducting crowdsourced research without proper quality control measures. For example, a study that was able to recruit 25% of its intended sample would likely draw erroneous conclusions from analyses due to the lack of power for the planned analyses. As crowdsourcing is likely to be used by early career investigators and students who may have limited financial resources to collect data, this budget efficiency is critical (Agley et al., 2021).

4.2.1 Comparison of Cyborg to Attention Check Methods.—The present study also compared the results of the "attention" check method of validating responses to the Cyborg Method. The majority of cases classified as valid by the Cyborg method were also classified as valid via the attention check method. However, a substantial number of responses that were invalid according to the Cyborg Method also answered all attention questions correctly, as was found in prior work (Dennis et al., 2019). This finding highlights the potential limitation of the attention check method for data validation. Respondents who submit invalid responses are likely aware of this approach and have strategies to navigate them (Chandler et al., 2020). Thus, relying on this strategy alone may still result in the collection of invalid data. Based on these results and the concern that the inclusion of such items may alienate certain users, researchers are recommended to adopt alternative strategies than attention check items (Vannette, 2017).

4.3 Limitations

The study had several limitations. First, the validity of each participant could not be verified independently. That is, we were unable to independently confirm that each response labeled valid was in fact valid and vice-versa. As such, this method may still result in errors during data collection. It is recommended that an appeal process be included in such a data collection method to allow for incorrectly identified responses to be appealed to the study team and the outcomes rectified. Second, participants in the current study were paid $4 for completing the study, which was above the $2 market rate. This was done to incentivize participation, but it may have attracted more invalid responses. Finally, we were focused on collecting data from a trauma-exposed sample and thus the short answer questions were specific to trauma exposure. Future work should determine the extent that other short answer questions can be used as effectively as those about a traumatic event.

4.4 Conclusions

In conclusion, the Cyborg Method provides an efficient method for obtaining high quality crowdsourced data. Methods that improve the quality of such data are needed to maintain the viability of this approach. Future work is needed to evaluate this method's utility and viability over time with a range of samples. This work should include novel methods to evaluate the qualitative Short Answer Review expedite that process. 'Bad actors' are likely to devise additional strategies to navigate existing quality control measure and further protections will be needed. For example, the rise of sophisticated AI chatbots may require

components of this method to be revisited. If researchers and manuscript reviewers remain diligent, crowdsourcing will remain a viable method to collect data efficiently.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Agley J, Xiao Y, Nolan R, & Golzarri-Arroyo L. (2021). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. Behavior Research Methods. 10.3758/s13428-021-01665-8

Aguinis H, Villamor I, & Ramani RS (2021). MTurk Research: Review and Recommendations. Journal of Management, 47(4), 823–837. 10.1177/0149206320969787

American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (5th ed.). American Psychiatric Association.

Arch JJ, & Carr AL (2017). Using Mechanical Turk for research on cancer survivors. Psycho-Oncology, 26(10), 1593–1603. 10.1002/pon.4173 [PubMed: 27283906]

Arditte KA., Çek D., Shaw AM., & Timpano KR. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. Psychological Assessment, 28(6), 684–691. 10.1037/pas0000217 [PubMed: 26302105]

Beard C, Hsu KJ, Rifkin LS, Busch AB, & Björgvinsson T. (2016). Validation of the PHQ-9 in a psychiatric sample. Journal of Affective Disorders, 193, 267–273. 10.1016/j.jad.2015.12.075 [PubMed: 26774513]

Boas TC, Christenson DP, & Glick DM (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. Political Science Research and Methods, 8(2), 232–250. 10.1017/psrm.2018.28

Buhrmester MD, Talaifar S, & Gosling SD (2018). An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use. Perspectives on Psychological Science, 13(2), 149–154. 10.1177/1745691617706516 [PubMed: 29928846]

Chandler J, Sisso I, & Shapiro D. (2020). Participant carelessness and fraud: Consequences for clinical research and potential solutions. Journal of Abnormal Psychology, 129(1), 49–55. 10.1037/abn0000479 [PubMed: 31868387]

Christensen KS, Oernboel E, Zatzick D, & Russo J. (2017). Screening for depression: Rasch analysis of the structural validity of the PHQ-9 in acutely injured trauma survivors. Journal of Psychosomatic Research, 97, 18–22. 10.1016/j.jpsychores.2017.03.117 [PubMed: 28606494]

Dennis SA, Goodson BM, & Pearson CA (2019). Online Worker Fraud and Evolving Threats to the Integrity of MTurk Data: A Discussion of Virtual Private Servers and the Limitations of IP-Based Screening Procedures. Behavioral Research in Accounting, 32(1), 119–134. 10.2308/bria-18-044

Dolan M., Contractor AA., Ryals AJ., & Weiss NH. (2020). Trauma, posttraumatic stress disorder severity, and positive memories. Memory (Hove, England), 28(8), 998–1013. 10.1080/09658211.2020.1809679 [PubMed: 32840463]

Engle K, Talbot M, & Samuelson KW (2020). Is Amazon's Mechanical Turk (MTurk) a comparable recruitment source for trauma studies? Psychological Trauma: Theory, Research, Practice, and Policy, 12(4), 381–388. 10.1037/tra0000502 [PubMed: 31380674]

Forkus SR, Raudales AM, Rafiuddin HS, Weiss NH, Messman BA, & Contractor AA (2022). The Posttraumatic Stress Disorder (PTSD) Checklist for DSM–5: A systematic review of existing psychometric evidence. Clinical Psychology: Science and Practice, No Pagination Specified-No Pagination Specified. 10.1037/cps0000111

Gray MJ, Litz BT, Hsu JL, & Lombardo TW (2004). Psychometric properties of the life events checklist. Assessment, 11(4), 330–341. 10.1177/1073191104269954 [PubMed: 15486169]

MacInnis CC, Boss HCD, & Bourdage JS (2020). More evidence of participant misrepresentation on Mturk and investigating who misrepresents. Personality and Individual Differences, 152, 109603. 10.1016/j.paid.2019.109603

Mellis AM, & Bickel WK (2020). Mechanical Turk data collection in addiction research: Utility, concerns and best practices. Addiction, 115(10), 1960–1968. 10.1111/add.15032 [PubMed: 32135574]

Morgan JK., & Desmarais SL. (2017). Associations Between Time Since Event and Posttraumatic Growth Among Military Veterans. Military Psychology, 29(5), 456–463. 10.1037/mil0000170

Natesan Batley P, Contractor AA, Weiss NH, Compton SE, & Price M. (2021). Psychometric Evaluation of the Posttrauma Risky Behaviors Questionnaire: Item Response Theory Analyses. Assessment, 10731911211036760. 10.1177/10731911211036760

Qualtrics XM. (2021). Fraud Detection. https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/

Shapiro DN, Chandler J, & Mueller PA (2013). Using Mechanical Turk to Study Clinical Populations. Clinical Psychological Science, 1(2), 213–220. 10.1177/2167702612469015

Strickland JC, & Stoops WW (2019). The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. Experimental and Clinical Psychopharmacology, 27(1), 1–18. 10.1037/pha0000235 [PubMed: 30489114]

van Stolk-Cooke K, Brown A, Maheux A, Parent J, Forehand R, & Price M. (2018). Crowdsourcing Trauma: Psychopathology in a Trauma-Exposed Sample Recruited via Mechanical Turk. Journal of Traumatic Stress, 31(4), 549–557. 10.1002/jts.22303 [PubMed: 30025175]

Vannette D. (2017, June 29). Using Attention Checks in Your Surveys May Harm Data Quality. Qualtrics https://www.qualtrics.com/blog/using-attention-checks-in-your-surveys-may-harm-data-quality/

Weathers FW, Blake DD, Schnurr PP, Kaloupek DG, Marx BP, & Keane TM (2013). The Life Events Checklist for DSM-5 (LEC-5). Instrument Available from the National Center for PTSD at Www.Ptsd.va.Gov. http://www.ptsd.va.gov/professional/assessment/te-measures/life_events_checklist.asp

Weber S. (2021). A Step-by-Step Procedure to Implement Discrete Choice Experiments in Qualtrics. Social Science Computer Review, 39(5), 903–921. 10.1177/0894439319885317

Weiss NH, Schick MR, Contractor AA, Goncharenko S, Raudales AM, & Forkus SR (2021). Posttraumatic stress disorder symptom severity modulates avoidance of positive emotions among trauma-exposed military veterans in the community. Psychological Trauma: Theory, Research, Practice, and Policy, No Pagination Specified-No Pagination Specified. 10.1037/tra0001048

Highlights

- A substantial portion of crowdsourced data will contain responses that are invalid.
- Automated evaluation of a user's IP address can identify a portion of invalid responses.
- Reviewing responses to short answer questions can also identify a portion of invalid responses.
- IP evaluation and short answer review, a Cyborg Method, are needed to identify valid responses.
- The Cyborg Method can be applied proactively to reduce study costs.

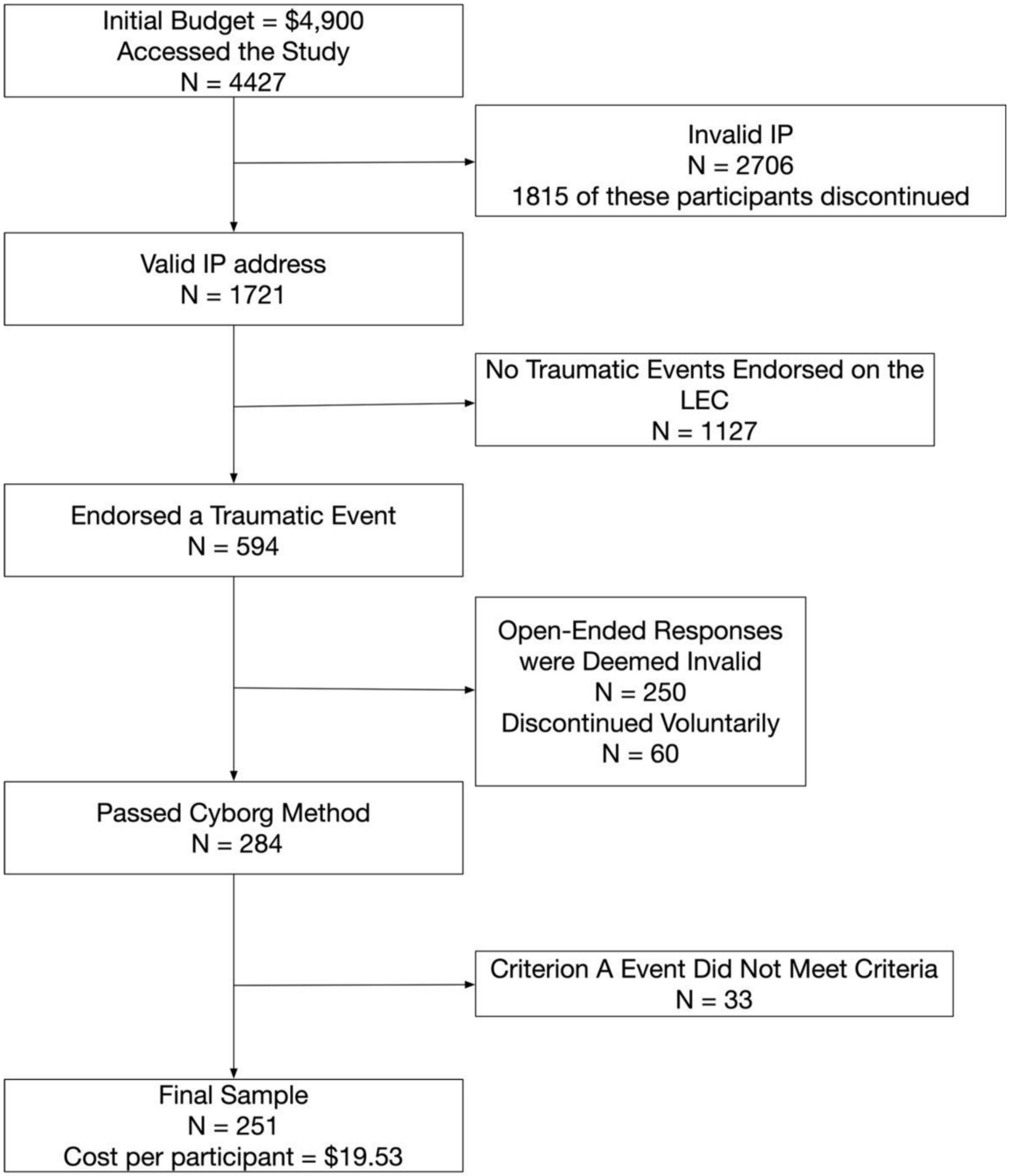

Figure 2.
Flow diagram illustrating the number of participants screened at each stage of the screening procedures for Sample 1.

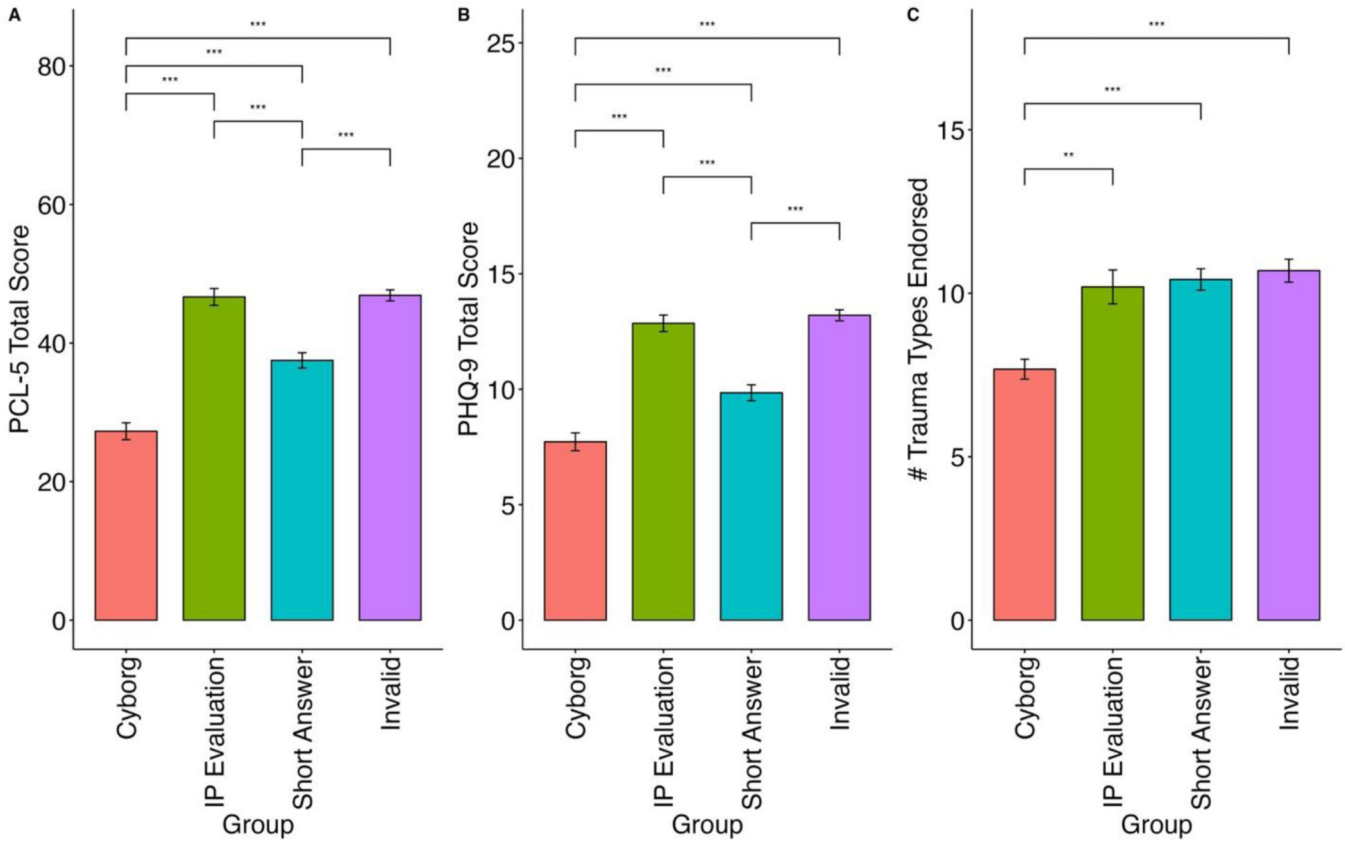

Figure 3.

Comparisons of Scores on the PCL-5 and PHQ-8 across the different categories of validation. Panel A: PCL-5 scores. Panel B: PHQ-8 Scores, Panel C: Number of Trauma Types Endorsed on the Life Events Checklist. Cyborg N = 232. IP Evaluation N = 167. Short Answer N = 256. Invalid N = 324. *** = $p < .001$. ** = $p < .01$.

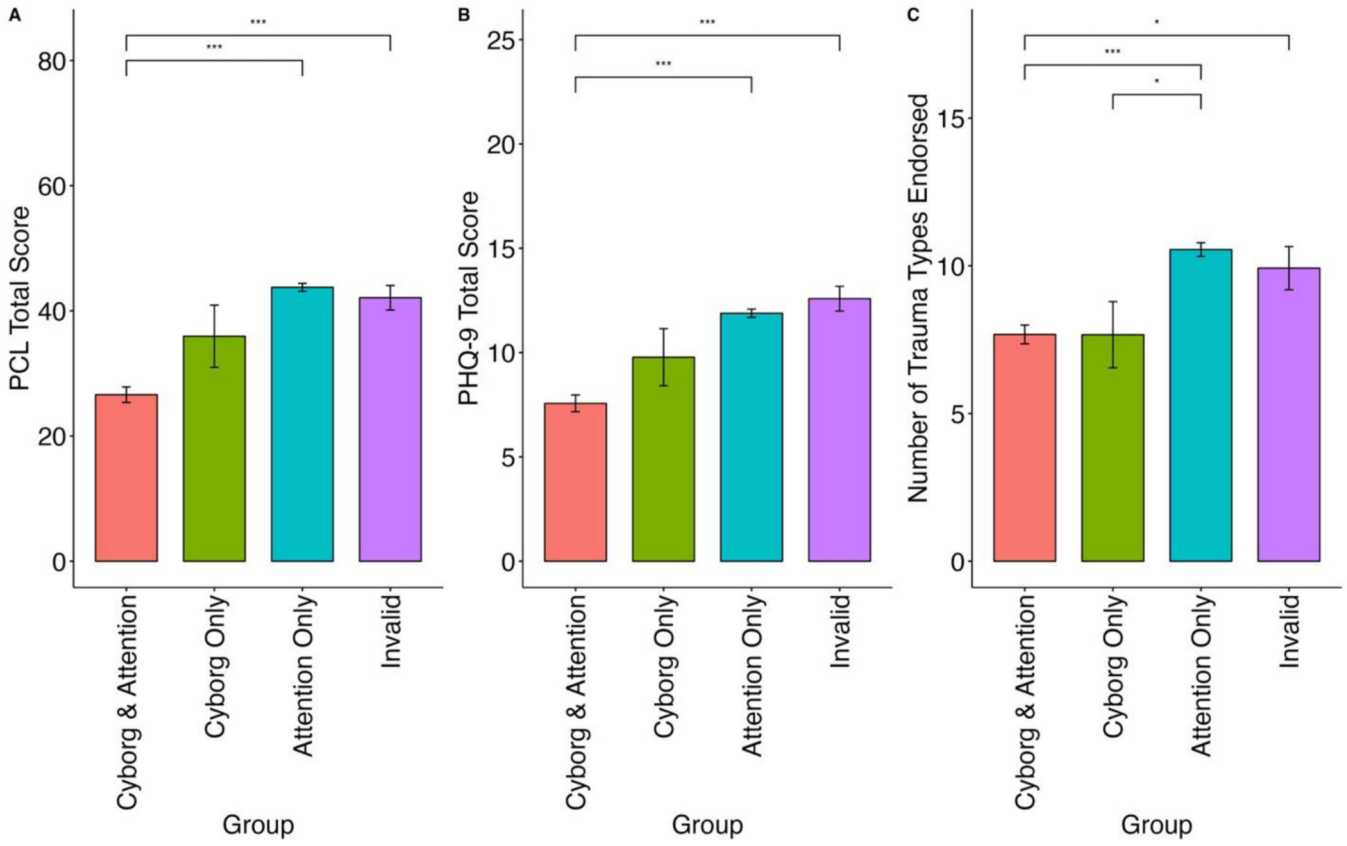

Figure 4.

Comparisons of Scores on the PCL-5 and PHQ-8 across the Cyborg and Attention Check Methods. Panel A: PCL-5 scores. Cyborg & Attention N = 235. Cyborg Only N = 18. Attention Only N = 655. Invalid N = 235. Panel B: PHQ-8 Scores, Panel C: Number of Trauma Types Endorsed on the Life Events Checklist. *** = $p < .001$. * = $p < .05$.

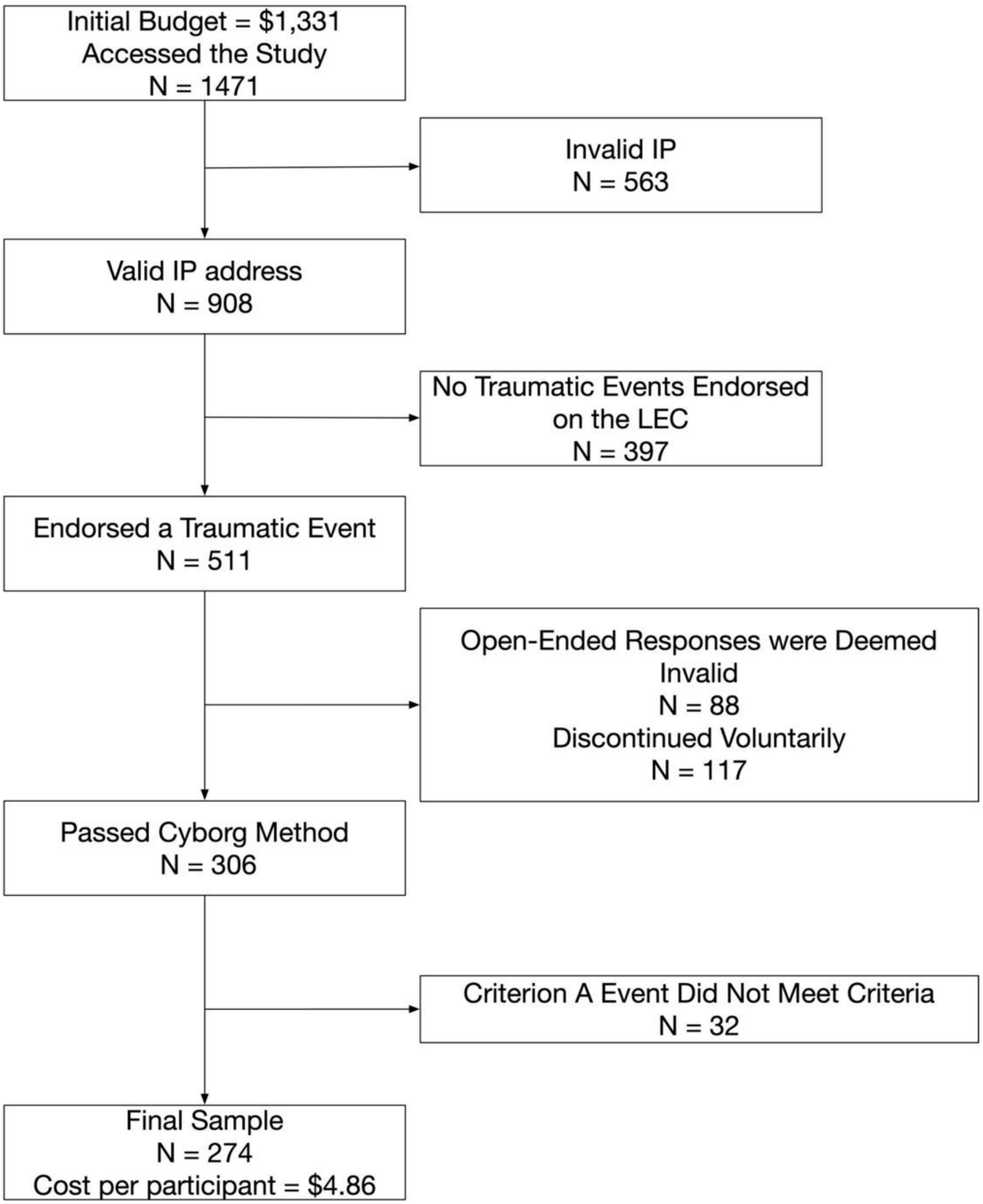

Figure 5.
Flow diagram illustrating the number of participants screened at each stage of the screening procedures for Sample 2.

Table 1.

Descriptive statistics for the valid and invalid responses from Sample 1.

Variable	IP + / SA + (Cyborg) N = 232		IP + / SA - N = 167		IP - / SA + N = 256		IP - / SA - N = 324	
	M	SD	M	SD	M	SD	M	SD
PCL-5	27.15	19.27	46.59	15.63	37.50	17.44	46.96	14.38
PHQ-8	7.60	6.08	12.86	4.57	9.84	5.50	13.21	4.25
Trauma Types								
Experienced	7.73	4.78	10.17	6.67	10.40	5.22	10.58	6.25
Age	36.09	9.89	33.47	8.44	31.82	7.75	32.30	8.10
	N	%	N	%	N	%	N	%
Gender (Male)	113	48.71	91	54.49	127	49.61	182	56.17
Latinx	22	9.48	56	33.53	42	16.41	96	29.63
Race								
White	188	81.03	108	64.67	162	63.28	176	54.32
African American	15	6.47	40	23.95	22	8.59	41	12.65
American Indian	4	1.72	11	6.59	7	2.73	6	1.85
Asian American	11	4.74	6	3.59	57	22.27	76	23.46
Pacific Islander	0	0.00	0	0.00	0	0.00	8	2.47
Bi-racial	7	3.02	0	0.00	6	2.34	0	0.00
Other	7	3.02	2	1.20	2	0.78	17	5.25

Note: PCL-5 = PTSD Checklist for the DSM 5. PHQ-8 = Patient Health Questionnaire 9.

Table 2.

Descriptive Statistics for Sample 2.

	Sample 2 Cyborg N = 271	
Variable	**M**	**SD**
PCL-5	21.12	19.83
PHQ-8	6.56	5.59
Trauma Types Experienced	8.52	5.06
Age	40.07	11.71
	N	**%**
Gender (Male)	142	52.4
Latinx	37	13.7
Race		
White	207	76.38
African American	28	10.33
American Indian	3	1.11
Asian American	18	6.64
Pacific Islander	1	0.37
Bi-racial	7	2.58
Other	7	2.58