Published in final edited form as:
AIDS Behav. 2015 October ; 19(10): 1928–1937. doi:10.1007/s10461-015-1033-y.

An Updated Protocol to Detect Invalid Entries in an Online Survey of Men Who Have Sex with Men (MSM): How Do Valid and Invalid Submissions Compare?

Jeremy A. Grey[1], **Joseph Konstan**[2], **Alex Iantaffi**[3], **J. Michael Wilkerson**[4], **Dylan Galos**[5], and **B. R. Simon Rosser**[5]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Rd. NE, Atlanta, GA 30322, USA

[2]Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA

[3]Department of Family Medicine & Community Health, University of Minnesota Medical School, Minneapolis, MN, USA

[4]Division of Health Promotion & Behavioral Sciences, The University of Texas Health Sciences Center (UTHealth) at Houston, Houston, TX, USA

[5]Division of Epidemiology & Community Health, University of Minnesota School of Public Health, Minneapolis, MN, USA

Abstract

Researchers use protocols to screen for suspicious survey submissions in online studies. We evaluated how well a de-duplication and cross-validation process detected invalid entries. Data were from the Sexually Explicit Media Study, an Internet-based HIV prevention survey of men who have sex with men. Using our protocol, 146 (11.6 %) of 1254 entries were identified as invalid. Most indicated changes to the screening questionnaire to gain entry (n = 109, 74.7 %), matched other submissions' payment profiles (n = 56, 41.8 %), or featured an IP address that was recorded previously (n = 43, 29.5 %). We found few demographic or behavioral differences between valid and invalid samples, however. Invalid submissions had lower odds of reporting HIV testing in the past year (OR 0.63), and higher odds of requesting no payment compared to check payments (OR 2.75). Thus, rates of HIV testing would have been underestimated if invalid submissions had not been removed, and payment may not be the only incentive for invalid participation.

Keywords

Survey methods; Questionnaires; Bias; HIV; Validity

Correspondence to: Jeremy A. Grey.

Introduction

Internet-based research has created new opportunities for public health researchers [1]. With proper design and appropriate rigor, researchers can improve data quality, decrease participant burden, and better protect participant confidentiality [1]. Internet-based research may also be an effective way to recruit individuals from vulnerable populations, who might be more geographically dispersed. These qualities make Internet-based research a popular option for studies of HIV risk and prevention. Researchers have used Internet-based methods to recruit hard-to-reach populations such as men who have sex with men (MSM) [2–4], transgender individuals [5, 6], and illicit drug users [7, 8].

The ease of access and relative anonymity of Internet-based surveys pose unique challenges to researchers. Investigators do not typically meet the participants in online studies, making it more difficult to detect submissions from ineligible enrollees, such as those who misrepresent themselves in order to participate or who complete multiple surveys. As Skitka and Sargis [9] point out, “[T]he tendency to take on false identities on the Web poses a problem for those whose research depends on successfully identifying specific personal characteristics of research participants” (p. 548).

In order to reduce threats to data integrity from invalid submissions, researchers may implement protocols for de-duplication and validation. De-duplication refers to identifying multiple submissions from the same person, while cross-validation is the process of confirming that participants meet the study’s eligibility criteria [10–13]. Studies that have focused on de-duplication use methods such as recording Internet protocol (IP) addresses and may require personal information (e.g., email address or telephone number) for registration or payment [10, 14]. To confirm validity, researchers have also examined the internal consistency of responses by including “cross-validation” items periodically throughout the survey. An example of this is requiring a birth date at the beginning of the survey and asking age at the end [12, 15].

Our team was among the first to publish a protocol for identifying potentially invalid survey entries in online HIV prevention research [12]. Since then, several studies of MSM have adapted this protocol [10, 16]. Our original de-duplication and cross-validation protocol included the following elements: cross-checking eligibility criteria such as MSM status, United States (US) residence, and age with answers to survey items and payment information; identifying duplicate survey submissions by detecting duplicate IP addresses (full and partial), e-mail addresses, names, and payment information (check address or e-payment receipts); and noting short completion times (i.e., 12 min or less) [12]. Bowen et al. [10] used similar detection variables but also examined similarities in usernames and passwords created for the study’s registration process. In a recent study, Bauermeister et al. [16] used two new methods of determining validity: using user-submitted data to profiles on public social network accounts and asking questions that helped explain reasons why multiple entries might be submitted from same residence/IP address (e.g., whether the respondent reported having roommates or a partner) [16].

Published studies have demonstrated the utility of these protocols. As we noted in Konstan et al. [12], our protocol enabled us to detect one individual who completed the study 65 times. Similarly, Bauermeister et al. [16] identified 41 of 548 (7.4 %) of their cases as invalid and a further 62 (11.3 % or 18.7 % total) as suspicious entries. The associations of interest in their study of young MSM would have differed if they had not excluded data from suspicious submissions or checked for valid cases among those flagged [16].

Some researchers question whether protocols for detecting invalid data are ethical or necessary. There is debate in human subjects' research over whether it is ethical to collect and utilize IP addresses, which some researchers consider a critical component of such protocols, without informed consent [17]. In brief, IP addresses are identification numbers for devices that connect to the Internet. They are comprised of quadrants that are separated by decimal markers. The first three quadrants identify a network, while the fourth is unique to the device. Thus, multiple devices (e.g., desktop computers or tablets) that connect on the same network will share the first three quadrants and differ on the fourth. IP addresses can also be dynamic or static. With a dynamic IP address, Internet service providers (ISPs) assign a new IP address to a device every time it connects to the Internet [18, 19]. Static IP addresses, however, do not change between sessions. For this reason, some researchers caution that the collection of IP address can be considered personally identifying information [19, 20]. Others argue it identifies a geographic location or ISP, not a person. For this location to become identifying, the ISP would need to provide the additional information. Consequently, those who maintain that an IP address is not identifying information reason that the identity of individuals generally cannot be discovered from an IP address alone [19–21].

Investigators who do not wish to implement study validation protocols, particularly those that use IP addresses, might choose to reduce the incentive for individuals to submit invalid data. Researchers implementing similar protocols have found that reducing monetary incentives or implementing additional identity checks resulted in reduced rates of invalid participation [22]. Nosen and Woody conclude that reducing motivation using minimal incentives and automated participation may be an easy alternative to implementing a full de-duplication and cross-validation protocol [22].

We found no studies since Konstan et al. [12] that evaluated specific components of a full protocol (i.e., not just multiple submissions, as Bowen et al. [10] reported). Hence, this manuscript had three aims. The first aim was to identify variables to include in a de-duplication and cross-validation protocol. Self-report and automated data (e.g., IP addresses) were inspected according to an extended adaptation of our original protocol [12]. The second aim was to examine differences between invalid and valid subsamples, particularly on key HIV risk and prevention variables [e.g., sexual risk behavior and sexually transmitted infection (STI) testing], to identify potential differences in study conclusions between the full sample and the validated sample (e.g., using or not using the protocol). The third aim was to compare the rate of invalid participation between participants who requested and declined compensation, since payment has been identified as a main motivator of multiple and invalid participation in studies [22]. For this aim, a priori, we hypothesized that the odds of invalid responding would be significantly higher among those completing the study for payment than those who requested no payment.

Methods

Study Design

The Sexually Explicit Media (SEM) Study was a cross-sectional survey that examined the relationship between SEM (e.g., “pornography”) consumption and sexual behavior in MSM. The primary research question was to examine whether exposure to and consumption of SEM is related to HIV risk behavior. Recruitment occurred between May and August, 2011. Banner advertisements appeared 7,939,758 times across 148 websites during the collection period, which resulted in a click-through rate of 0.16 %. These rates were comparable or better than those seen in similar studies [4]. Clicking on a banner ad led potential participants to a website with an eligibility screening questionnaire. Individuals were eligible if they were males aged 18 and older who lived in the US and had at least one male sex partner in the 5 years prior to the study. Those deemed eligible advanced to screens detailing the purpose of the study and consent protocols. These web pages clearly stated that men who had previously participated were ineligible. To ensure racial/ethnic diversity, the study also implemented a block recruitment strategy of a maximum of 400 persons per racial/ethnic category. Because of this, non-Hispanic white men became ineligible after the target recruitment of 400 was reached. No other racial or ethnic groups reached the ceiling for this criterion.

After completing the screening questionnaire, individuals viewed a summary of their responses, including that they had not completed this study previously, and either confirmed that they were accurate or changed them if there was an error. Only those who met criteria after verifying their age, sex, country of residence, a male sex partner in the previous 5 years, and no prior survey submissions could proceed to the consent process. Individuals who were ineligible received a message thanking them for their interest and exited the survey automatically.

The screening questionnaire and study survey were implemented with LimeSurvey version 1.0 [23]. A total of 1254 completed surveys were submitted. The median completion time was 40 min.

Measures

The protocol made use of paradata, sometimes called metadata, and survey responses. Paradata are those that are collected automatically by survey software in conjunction with survey responses [24]. These include IP addresses and timestamps, described in the following sections. In addition, we collected demographic data and information about sexual behavior and testing for HIV and other STIs.

IP Address—The survey software logged an IP address each time an individual accessed the eligibility screening questionnaire. For the current study, we also looked up the country and, when available, city and state of the IP address. All data were treated as protected health information (PHI) and were stored on a secure server in a password-protected file with other identifying information about the participant, which was kept separate from the de-identified file containing participants' responses.

Timestamp—The software logged two timestamps during each visit to the online survey. One recorded the first time the survey was accessed and the other recorded the last activity on the survey. We computed the time taken to complete the survey as the difference between the two times. An acknowledged limitation of this estimate is that some participants could save their progress and complete the survey in multiple sessions or at a later time, resulting in times that appear to be many hours or days in duration. Therefore, we caution that the *time taken* variable is only valid for detecting short to moderate completion times; we did not examine unusually long completion times. For the SEM study, we determined a priori that a completion time of under 20 min be flagged as suspicious. This amount of time represented approximately 10 s per question under the most minimal scenario according to our skip logic.

Demographics—Individuals were asked to provide their age, race, and ethnicity as part of the eligibility screener for the survey. In the screener, response options for age were categorical according to years: *under 16*, *16–17*, *18–24*, *25–34*, *35–44*, *45–54*, and *55 or older*. Hispanic/Latino ethnicity was asked as a *yes* or *no* question, and race was asked as discrete categories: *white*, *black or African American*, *American Indian or Alaska Native*, *Asian or Pacific Islander*, and *multiracial or other*. Each question was asked again as part of the survey, with age asked as an open-ended numeric item, Latino/Hispanic identity further specified (*Mexican*, *Mexican–American*, *or Chicano; Puerto Rican; Cuban*; or *another Hispanic*, *Latino*, *or Spanish origin*), and race asked as a multiple-answer item (e.g., *check all that apply*). In order to have sufficient groups for analysis, and due to low participation of some races in the sample, participants' race and ethnicity was re-categorized as *white* (*non-Hispanic*), *black* (*non-Hispanic*), *Latino/Hispanic*, and *multiracial or other* (*non-Hispanic*) for analysis. Age was also categorized as *17–24*, *25–34*, *35–44*, and *45 and over* (the two highest age categories were collapsed due to small sample sizes). Although ineligible, two subjects reported being 17 at the time of the survey, so the lowest category was extended for reporting the demographics of the invalid sample.

In addition to cross-validity items, participants were asked their sexual identity (*gay/ homosexual*, *bisexual*, *straight*, *same-gender-loving*, *queer*, and *other*), level of education according to highest degree completed (*up to 11th*, *high school diploma/GED*, *some college but no degree*, *associate's degree*, *bachelor's degree*, and *graduate degree*), and annual income (open-ended). As with race and ethnicity, *other* responses for sexual orientation were recoded and sexual orientation was collapsed as *gay* and *not gay* due to low percentages of all other identity categories. Finally, income was categorized as *$0–24,999*, *$25,000–$49,999*, *$50,000–$74,999*, and *$75,000 or more.*

Sexual Behavior—Participants reported the number of male partners they had in the 90 days prior to the survey. Follow-up items asked how many of this number were primary partners, defined as "a regular sex partner such as a boyfriend, husband, domestic partner that you have been in a relationship with for at least 3 months." The number of casual male partners was derived by subtracting the number of primary male partners from the total number of male partners. Participants who had casual partners were asked the number of partners with whom they engaged in protected and unprotected insertive and receptive anal sex.

HIV/STI Testing—Individuals were asked their HIV status. Response options were *HIV-positive; I'm not sure, but I think HIV-positive; I don't know; I'm not sure, but I think HIV-negative; HIV-negative*; and *refuse to answer*. For the current study, *I'm not sure…* and *I don't know* responses were collapsed into a single *HIV-unsure* category. *Refuse to answer* responses were coded as missing.

Men who were HIV-negative or HIV-unsure were asked how recently they were tested for HIV. Response options were *in the last 3 months*, *in the last year*, *1–2 years ago*, *more than 2 years ago*, *I have never been tested for HIV*, *I can't remember*, *not applicable*, and *refuse to answer*. Men who reported being tested either in the last 3 months or in the last year were categorized as having been tested in the previous year. All other responses, including *I can't remember* and *not applicable* were coded as not having been tested in the previous year. *Refuse to answer* was coded as missing.

Men were asked if they were tested for seven STIs in the past 12 months: syphilis, gonorrhea, chlamydia, human papilloma virus (HPV; genital or anal warts), genital herpes, hepatitis A, and hepatitis B. Response options for each were *yes, and I got the results; yes, but I did not get the results; I don't know; no*; and *refuse to answer*. Men in the first category were categorized as having received an STI result (whether negative or positive), while all others (except *refuse to answer*) were categorized as not receiving an STI result. As with other variables, *refuse to answer* responses were coded as missing.

De-duplication and Cross-Validation Protocol

For all analyses, we considered duplicate and invalid submissions to be one *invalid* group. Survey submissions were identified as potentially invalid according to an extended version of the de-duplication and cross-validation protocol described by Konstan et al. [12] and Rosser et al. [25]. For de-duplication, we examined submissions for common IP addresses (full and first three quadrants), payment names, payment addresses, and payment emails. We considered two or more entries that matched on any variable to be possible repeat submissions. Identical matches of payment information such as payment name and address or email handles, even on different domains, were considered multiple submissions. Entries without matching payment information but with matching IP addresses—including the first three quadrants of an IP address, which indicate the same network but not a specific machine —were examined using the remaining de-duplication procedures. Following Konstan et al. [12], the first entry of a repeat responder was considered valid as long as there was no indication of changed eligibility status.

To cross-validate surveys, we compared participants' responses to the eligibility screening questionnaire to similar items asked in the body of the survey. We extracted the relevant items from the survey and stored them in a separate file. These items were age, asked as an open-ended numeric item and as a categorical, multiple-choice item in the screener; US residence, asked as ZIP code in both the screener and the survey and confirmed by IP address; and status as a man who has sex with men, asked categorically in the screening questionnaire and derived from lifetime male sexual partner frequencies in the body of the survey. Submissions were flagged automatically for incongruity between these items and then checked manually. Responses that were near-matches, such as nearby ZIP codes, were counted as valid. Submissions in which individuals reported being 18–24 years old in the screener but subsequently reported being 17 (an ineligible age) were counted as invalid.

Finally, we examined the data for evidence of ineligibility. Completed submissions were compared with earlier attempts from the same IP address. Since potential participants were unable to change their responses to the screening questionnaire after confirming them, individuals who were ineligible had to begin a new session if they wanted to change their answers. Thus, each attempt at the survey was recorded independently. By examining timestamps and IP addresses, it was possible to determine whether multiple screener entries were submitted from the same computer or network and if eligibility status changed within a short time frame. Eligible entries that were submitted from an IP address within 30 min of an ineligible entry were interpreted as attempts to determine the eligibility criteria in order to manipulate entry to the study. These were then deemed invalid.

Analyses

We evaluated the utility of each component in two ways. First, we summarized the number and percentage of submissions that were flagged by the criterion, followed by the number and percentage that were deemed invalid. Next, we reported the percentage of all invalid submissions that were flagged by each component to demonstrate how influential that component was in identifying the invalid submissions.

We used unadjusted logistic regression models to compare valid and invalid submissions on key demographic variables. These variables included age, race/ethnicity, sexual identity, urban or rural locality, level of education, and income category. HIV status (i.e., positive, negative, and unknown) and payment preference (i.e., check, PayPal, or no payment) were also compared between subsamples. Following this, we constructed a multivariable logistic regression model using statistically significant predictors.

We used a series of negative binomial regression models to examine differences between the validity statuses in counts of total, primary, and casual sexual partners during the 90 days prior to the survey. We further specified casual sexual partners according to anal sex role and condom use. Negative binomial regression models were more appropriate than Poisson regression models based on indices of overdispersion in the data and model fit improvement. These models yielded unadjusted rate ratios (RRs) as well as RRs that were adjusted for age and race/ethnicity.

Finally, we used a series of logistic regression models to determine if validity status predicted self-reported HIV and STI testing behavior. Data were analyzed using the Stata statistical package, version 12.1 [26]. Statistical significance was evaluated at $\alpha = 0.05$.

Results

Using the de-duplication and cross-validation protocol to screen 1254 submissions, we identified 25 (2.0 %) as repeat submissions and 125 (10.0 %) as being from ineligible individuals. Table 1 reports the number and percent of invalid submissions out of the full sample by each criterion. For each component of the protocol, Table 1 also lists the percentage of all invalid entries (n = 146) that contained that evidence. Since the protocol used all criteria, they were not mutually exclusive. Thus, frequencies and percentages of invalid submissions with each feature may sum to more than 146 and 100, respectively.

Changes to the eligibility screener (74.7 % of invalid entries), duplicate payment name (38.4 % of invalid entries), and duplicate IP address—both the entire address (31.2 % of invalid entries) and the first three quadrants (34.4 % of invalid)—identified the highest percentage of invalid submissions. Of the 109 enrollees who changed their responses to the screener, a majority (56.9 %) changed more than one response (Table 2).

Age, ethnicity, HIV status, and payment preference were significantly associated with invalidity (Table 3). Individuals in the oldest category (45 and over) had lower odds of being identified as invalid compared to those in the lowest age category (17–24; adjusted OR 0.46, 95 % CI 0.23–0.90). Participants who identified as Hispanic or Latino had greater odds of being invalid compared to non-Hispanic White participants (aOR 2.26, 95 % CI 1.50–3.39). Those who reported being HIV-positive were at lower odds of being identified as invalid compared to those who reported being HIV-negative (OR 0.30, 95 % CI 0.12–0.75). However, after adjusting for other significant covariates, HIV status was not significantly different by valid/invalid status (aOR 0.40, 95 % CI 0.16–1.03). Those who requested either payment via PayPal or no payment had greater odds of being deemed invalid than those who requested check payments (aOR 1.75, 95 % CI 1.19–2.57; aOR 2.75, 95 % CI 1.59–4.75).

Negative binomial regression models indicated no statistically significant differences between valid and invalid submissions in terms of reported number of sexual partners in the 3 months prior to the survey (Table 4). This was true of casual anal sex partners as well, regardless of anal sex role or condom use.

Crude and adjusted models indicated that invalid participants who reported being HIV-negative or uncertain of their HIV status were at significantly lower odds of self-reported HIV testing in the previous year (aOR 0.62, 95 % CI 0.42–0.90; Table 5). However, valid and invalid respondents did not differ significantly in reported STI testing for any STIs in either the unadjusted or adjusted logistic regression models (Table 5).

Discussion

This manuscript had three aims. The first aim was to evaluate components of a de-duplication and cross-validation protocol. The results highlight several design elements that researchers may incorporate into online surveys. The second aim was to determine whether submissions identified as invalid differed significantly from those accepted as valid. Age, ethnicity, and payment preference were associated with invalid submissions. Furthermore, among HIV-negative and HIV-unsure individuals, invalid submissions had lower odds of reporting an HIV test in the past year. The third aim postulated that invalidity would be higher among participants who chose to complete the study with compensation than those completing it without compensation; however, when we compared those who requested check payments and those who requested no payment, we found higher odds of invalidity among those who did not request payment.

Using the full de-duplication and cross-validation protocol, including the extended cross-validation, 146 (11.6 %) of all survey entries were identified as invalid, which is approximately the same as Konstan et al. [12]. The three components that were most effective for identifying invalid participants were tracking multiple attempts at the screening questionnaire, identifying repeated payment last names, and examining repeated IP addresses. In general, components derived from IP address identified the most invalid submissions. IP addresses were used both to detect multiple submissions and as an indicator of geographic location. When combined with timestamps and eligibility confirmation, they also made it possible to determine when ineligible and eligible submissions were submitted from the same network.

We recommend researchers adopt the practice of confirming eligibility responses with enrollees. By displaying participants' responses back to them and allowing them to make changes to their answers prior to submission, researchers can be more certain that subsequent changes to screening questionnaires are not corrections. Such changes, particularly within a short time frame, are most likely to be repeated attempts to discover the correct set of criteria in order to gain entry to the study. This extended cross-validation component was an important addition to the SEM Study's protocol. While Konstan et al.'s [12], Bowen et al.'s [10], and Bauermeister et al.'s [16] protocols all involved the detection of multiple *completed* entries, the SEM Study also took into account evidence of possible eligibility changes from prior attempts at the survey. This resulted in the detection of many individuals (n = 109, 8.7 % of the study sample) who would not have been identified using the earlier protocols.

As reported in Konstan et al. [12] and confirmed by Bauermeister et al. [16], manual inspection is a necessary follow-up to automatic flagging based on algorithms. With proper coding, statistical software can indicate when values are out of acceptable ranges (e.g., ages under 18) or text appears in multiple submissions (e.g., IP addresses) [12]. However, some of the components that identified the highest percentage of invalid entries required additional, manual inspection for a final valid or invalid classification (e.g., duplicate payment information or IP address). This was true of Bauermeister et al.'s [16] study as well, in which 60 % of entries initially flagged as potentially invalid were later determined to be valid.

The second aim of this study was to determine whether submissions that were determined to be invalid differed from those that were accepted as valid. In addition to age and ethnicity differences between valid and invalid samples, invalid submissions from individuals who were HIV-negative or HIV-unsure had lower odds of reporting an HIV test in the last year. This highlights the potential for non-validated samples to distort estimates. In this case, inclusion of invalid submissions would have resulted in a lower observed percentage of HIV testing among the study sample. Depending on the reasons for asking this question (e.g., as part of an intervention or needs assessment), such bias may have profound consequences.

Associations between invalidity and the demographic composition of the sample were consistent with previous research and with what would be expected from changes to the eligibility screening questionnaire. For example, Bauermeister et al. [16] also found that invalid cases were younger and more likely to be Hispanic or Latino than valid cases. In the current study, the lowest age category may have been the easiest one to switch to after individuals realized that they were too young. Likewise, the association between invalidity and Latino/Hispanic identity may be due to the quota being reached on non-Hispanic white participants; a change from white, non-Hispanic, to Latino/Hispanic identity was common when re-attempting the eligibility screening questionnaire. Consequently, researchers who wish to improve racial and ethnic diversity in their sample might consider methods other than quotas in online surveys, since this may encourage individuals to misreport race or ethnicity in order to participate. For studies that require sufficient diversity for statistical comparisons, implementing a registration process to separate the screening survey from the research study would make repeated attempts at the screening questionnaire more difficult and would allow researchers more time to vet submissions.

The association between invalid submissions and payment preferences was complex. PayPal payments were preferred over check payments in invalid submissions, which may be unsurprising; it is easier to remain anonymous using PayPal. Checks require both a name and address, and it is easier to create multiple payment profiles on a website than it is to have multiple legal names and addresses. The finding that *no payment* was selected more in the invalid group than in the valid group is unexpected, however. It appears to suggest that there are reasons for invalid participation that are not related to compensation. This is interesting in light of the findings of a previous study, which concluded that lowering or eliminating incentives resulted in reduced rates of invalid participation [22]. Given the intriguing subject matter of the current study (e.g., pornography), it may also be that individuals wanted to see the content of the survey (e.g., voyeuristic but ineligible persons) or to influence its findings (e.g., individuals who want to promote or discourage research on sexual minority populations). Regardless, motivations for participating multiple times or despite being ineligible deserve further investigation.

The study had several limitations. First, as advocated in Konstan et al. [12], investigators should adapt their de-duplication and cross-validation protocol to the particular research in question. Consequently, there were differences between the protocol used in this manuscript and those used in other published work [12, 16], which may drive the discrepancies in the studies' conclusions. Second, although a proxy for relationship status was asked as part of the survey (e.g., "how many primary partners do you have?"), there were no specific questions regarding cohabitation, either with a partner or roommate. Thus, we could not determine whether multiple entries from the same IP address reflected repeat submissions, or submissions from two different people who happened to be using the same computer. Third, in order to maintain confidentiality and, to the highest degree possible, anonymity of the study sample, we did not use social networking sites to verify eligibility. This was a limiting factor for detecting repeat respondents; we relied on matching payment information, and algorithms did not automatically detect when individuals varied the way their names or addresses were presented for checks, such as using first or last initials in place of names. Names were also not always available for unverified PayPal accounts. Finally, the findings of this study may not extend to other online survey research, or to research focusing on populations other than MSM. Few online studies publish the outcomes of their validation procedures, and other methods of recruitment, screening, and validation (e.g., roommate data [16] or additional identifying information [10]) might yield different rates of invalid participation. Thus, as more de-duplication and cross-validation protocols are modified and tailored, research groups should publish the outcomes of their validation procedures in order to expand the options available to investigators who use Internet-based methods.

Future research should examine ways to improve de-duplication and cross validation efforts. In particular, two issues warrant further study. First, protocols could clarify for participants how many surveys will be accepted from the same IP address. This way, researchers could prevent multiple submissions made in good faith (e.g., by multiple participants completing a survey from the same computer) or by accident (e.g., two or more people independently completing the survey without knowledge of the other's involvement). However, finding ways to measure and enforce this raises confidentiality concerns. Although software programs can prevent multiple submissions from the same IP address, notifying participants that a prior submission was received from the IP address could lead one individual to deduce that another network user (e.g., roommate, partner, family member, employee) participated in the study. Furthermore, revealing that only one submission per IP address will be accepted might lead individuals who wish to participate multiple times to use more sophisticated methods, such as IP proxies or different locations, to gain access.

Second, the human subjects concerns regarding using IP addresses (or social networks, as in Bauermeister et al. [16]) deserve further consideration. Do researchers have an ethical responsibility to inform individuals that their IP address is collected and/or that their payment information may be used to verify their responses? When and under what circumstances is it ethical to track such data on enrollees? Will informing subjects that their IP address is being tracked deter both ineligible individuals and eligible individuals with privacy concerns from participating? Given the current political climate regarding data privacy, collecting individuals' data and using it without their knowledge raises practical considerations over and above the ethical concerns. Research regarding participants' perceptions of IP address as sensitive data, and commentary from ethicists and human subjects policy experts, would help to inform this area.

Conclusion

Protocols for identifying potentially ineligible or repeat submissions to online surveys should be considered for all Internet-based studies. Since the protocols themselves are relatively easy to implement with proper planning, the potential to identify approximately ten percent of a sample as invalid further justifies their use. The current study demonstrated that the components that identified the most potentially invalid entries from the de-duplication and cross-validation protocols were derived from IP addresses or from payment information across payment modes (e.g., check and PayPal). Invalid and valid subsamples differed in ways that might have biased the findings of research questions focused on age, race/ethnicity, education, or HIV testing. In addition, contrary to the widely held belief that payment promotes ineligible and repeat participation, those not requesting payment had higher odds of invalid surveys than those requesting check payment in this study. While the utility of these protocols was reaffirmed, additional research regarding methods of improving protocols while protecting human subjects is recommended.

Acknowledgments

The Sexually Explicit Media (SEM) Study was funded by the National Institute of Mental Health (NIMH), Grant #5R01MH087231-02. The authors are also grateful for the support of Barbara Lea in preparing this manuscript.

References

1. van Gelder MMHJ, Bretveld RW, Roeleveld N. Web-based questionnaires: the future in epidemiology? Am J Epidemiol. 2010; 172(11):1292. [PubMed: 20880962]
2. Bowen A. Internet sexuality research with rural men who have sex with men: can we recruit and retain them? J Sex Res. 2005; 42(4):317–23. [PubMed: 19827236]
3. Rosser, BRS., Konstan, J., Gurak, L., et al. Recruiting high risk Latino MSM for Internet prevention research. STD/HIV Prevention and the Internet Conference; Washington. 2001.
4. Salyers Bull S, Lloyd L, Rietmeijer C, McFarlane M. Recruitment and retention of an online sample for an HIV prevention intervention targeting men who have sex with men: The Smart Sex Quest Project. AIDS Care. 2004; 16(8):931–43. [PubMed: 15511725]
5. Bockting, WO., Miner, MH., Robinson, BE., et al. Use of the Internet to reach the US transgender population for HIV/STD prevention research. STD/HIV Prevention and the Internet Conference; Washington. 2003.
6. Rosser BRS. Capturing the social demographics of hidden sexual minorities: an Internet study of the transgender population in the United States. Sex Res Soc Policy. 2007; 4(2):50–64. DOI: 10.1525/srsp.2007.4.2.50
7. Duncan D. The general well-being of recreational drug users: a survey on the WWW. Int J Drug Policy. 2000; 11(5):315–23. [PubMed: 10967514]
8. Nicholson T. A survey of adult recreational drug use via the World Wide Web: the Drugnet Study. J Psychoact Drug. 1999; 31(4):415–22.
9. Skitka LJ, Sargis EG. The Internet as psychological laboratory. Annu Rev Psychol. 2006; 57:529–55. [PubMed: 16318606]
10. Bowen AM, Daniel CM, Williams ML, Baird GL. Identifying multiple submissions in Internet research: preserving data integrity. AIDS Behav. 2008; 12(6):964–73. [PubMed: 18240015]
11. Gosling SD, Vazire S, Srivastava S, John OP. Should we trust Web-based studies? a comparative analysis of six preconceptions about Internet questionnaires. Am Psychol. 2004; 59(2):93–104. [PubMed: 14992636]
12. Konstan JA, Rosser BRS, Ross MW, Stanton J, Edwards WM. The story of subject naught: a cautionary but optimistic tale of Internet survey research. J Comput Mediat Commun. 2005; 10(2):11.
13. Pequegnat W, Rosser BRS, Bowen AM, et al. Conducting Internet-based HIV/STD prevention survey research: considerations in design and evaluation. AIDS Behav. 2007; 11(4):505–21. [PubMed: 17053853]
14. Birnbaum MH. Human research and data collection via the Internet. Annu Rev Psychol. 2004; 55:803–32. [PubMed: 14744235]
15. Rosser BRS, Miner MH, Bockting WO, et al. HIV risk and the Internet: results of the Men's Internet Study (MINTS). AIDS Behav. 2009; 13(4):746–56. [PubMed: 18512143]

16. Bauermeister JA, Pingel E, Zimmerman M, Couper M, Carballo-Diéguez A, Strecher VJ. Data quality in HIV/AIDS Web-based surveys: handling invalid and suspicious data. Field Methods. 2012; 24(3):272–91. [PubMed: 23180978]

17. Lundevall-Unger P, Tranvik T. IP addresses—just a number? Int J Law Inf Technol. 2011; 19(1): 53–73. DOI: 10.1093/ijlit/eaq013

18. Chiasson MA, Parsons JT, Tesoriero JM, Carballo-Dieguez A, Hirshfield A, Remien RH. HIV behavioral research online. J Urban Health. 2006; 83(1):73–85. [PubMed: 16736356]

19. Nosek BA, Banaji MR, Greenwald AG. E-research: ethics, security, design, and control in psychological research on the Internet. J Soc Issue. 2002; 58(1):161–76. DOI: 10.1111/1540-4560.00254

20. Konstan, JA., Rosser, BRS., Horvath, KJ., Gurak, L., Edwards, W. Protecting subject data privacy in Internet-based HIV/STI prevention survey research. In: Conrad, FG., Schober, MF., editors. Envisioning the survey interview of the future. John Wiley & Sons; New York: 2008.

21. Im EO, Chee W. Issues in protection of human subjects in Internet research. Nurs Res. 2002; 51(4): 266. [PubMed: 12131239]

22. Nosen E, Woody SR. Online surveys: effect of research design decision on rates of invalid participation and data credibility. Grad Stud J Psychol. 2008; 10:3–14.

23. Schmitz, C. LimeSurvey Project Team. LimeSurvey: an open source survey tool. LimeSurvey Project; Hamburg: 2011.

24. Couper MP. Technology trends in survey data collection. Soc Sci Comput Rev. 2005; 23(4):486–501. DOI: 10.1177/0894439305278972

25. Rosser, BRS. Online HIV prevention and the Men's Internet Study-II (MINT-II): a glimpse into the world of designing highly-interactive sexual health Internet-based interventions. Canada: World Congress for Sexology; Montreal. 2005;

26. StataCorp. Stata Statistical Software: Release 12. StataCorp LP: College Station; 2011.

Table 1

Number of submissions flagged and deemed invalid by each component of the de-duplication and cross-validation protocol in the SEM Study (N = 1254)

Protocol component	**Full sample**		**Invalid sample (n = 146)**
	Flagged suspicious n (%)	**Deemed invalid[a] n (%)**	**Flagged by component n (%)**
De-duplication			
Payment last name duplicate	293 (23.4)	48 (3.8)	48 (32.9)
Check address duplicate	20 (1.6)	13 (1.0)	13 (8.9)
Payment e-mail address duplicate	8 (0.6)	8 (0.6)	8 (5.5)
IP address			
Complete	90 (7.2)	39 (3.1)	39 (26.7)
First 3 quadrants	103 (8.2)	43 (3.4)	43 (29.5)
Cross-validation			
Age invalid (<18)	2 (0.2)	2 (0.2)	2 (1.4)
Age mismatch (>1 year difference)	32 (2.6)	19 (1.5)	19 (13.0)
Changed screener to be eligible	109 (8.7)	109 (8.7)	109 (74.7)
Payment e-mail address invalid	0	–	–
IP address not US	35 (2.8)	35 (2.8)	35 (24.0)
No lifetime male partners	0	–	–
Time taken ≤20 min	38 (3.0)	20 (1.6)	20 (13.7)
Zip code			
Not valid US	8 (0.6)	5 (0.4)	5 (3.4)
Mismatch	69 (5.5)	12 (1.0)	12 (8.2)

Protocol components are not mutually exclusive. Totals may sum to more than 146

[a] *Invalid* refers to participants determined to be ineligible after they participated

Table 2

Changes to the SEM Study screening questionnaire to meet eligibility (n = 109)

Change	n	%
Age	2	1.8
Race/ethnicity	16	14.7
Gender	1	0.9
US residence	14	12.8
Sex with men	14	12.8
Multiple	62	56.9

Table 3

Demographic characteristics and payment preferences of valid and invalid submissions in the SEM Study (N = 1254)

	Valid		Invalid		OR[a] (95 % CI)	AOR[b] (95 % CI)
	n	%	n	%		
Total	1108	86.8	146	11.64		
Age						
17–24	379	34.2	66	45.2	Ref.	Ref.
25–34	345	31.1	49	33.6	0.82 (0.55, 1.21)	0.88 (0.59, 1.32)
35–44	185	16.7	19	13.0	0.59 (0.34, 1.01)	0.67 (0.38, 1.16)
45+	199	18.0	12	8.2	0.35 (0.18, 0.66)**	0.46 (0.23, 0.90)*
Race/ethnicity						
White	481	43.4	45	30.8	Ref.	Ref.
Black	131	11.8	10	6.8	0.82 (0.40, 1.66)	0.75 (0.37, 1.56)
Hispanic/Latino	303	27.3	64	43.8	2.26 (1.50, 3.39)***	2.10 (1.38, 3.21)***
Other/multi	193	17.4	27	18.5	1.50 (0.90, 2.48)	1.32 (0.78, 2.21)
Sexual identity						
Gay	178	16.1	31	21.4	Ref.	
Not gay	927	83.9	114	78.6	1.42 (0.92, 2.17)	
Education						
Less than 4-year degree	558	50.4	83	56.8	Ref.	
4-Year degree or higher	550	49.6	63	43.2	0.77 (0.54, 1.09)	
Annual income						
$0–$24,999	444	44.1	51	41.8	Ref.	
$25,999–$49,999	298	29.6	40	32.8	1.17 (0.75, 1.81)	
$50,000–$74,999	136	13.5	19	15.6	1.22 (0.69, 2.13)	
$75,000+	128	12.7	12	9.8	0.82 (0.42, 1.58)	
HIV status						
HIV+	114	10.3	5	3.4	0.30 (0.12, 0.75)*	0.40 (0.16, 1.03)
HIV−	839	75.8	122	83.6	Ref.	Ref.
Unsure	154	13.9	19	13.0	0.85 (0.51, 1.42)	0.73 (0.44, 1.24)

	Valid		Invalid		OR[a] (95 % CI)	AOR[b] (95 % CI)
	n	%	n	%		
Payment						
Check	322	29.1	56	38.4	Ref.	Ref.
Paypal	696	62.8	68	46.6	1.78 (1.22, 2.60)**	1.75 (1.19, 2.57)**
None	90	8.1	22	15.1	2.50 (1.47, 4.24)***	2.75 (1.59, 4.75)***

Frequencies may not add up to total (N = 1254) due to *refuse to answer* responses

[a] *OR* refers to the crude/unadjusted *odds ratio*

[b] *AOR* refers to the *adjusted odds ratio*, or the odds ratio after adjusting for other significant covariates

* $P < 0.05$,

** $P < 0.01$,

*** $P < 0.001$

Table 4

Rate ratios (*invalid* vs *valid*) of reported number of sexual partners (last 90 days) in the SEM Study, by type of partner, sex role, and condom use

	N[a]	RR[b] (95 % CI)	ARR[c] (95 % CI)
Total partners	1219	1.04 (0.62, 1.74)	0.97 (0.58, 1.64)
Primary partners	1048	1.24 (0.83, 1.86)	1.22 (0.81, 1.84)
Casual partners	1025	1.05 (0.68, 1.18)	1.04 (0.66, 1.62)
Insertive anal sex			
With condom	751	1.21 (0.77, 1.91)	1.17 (0.77, 1.91)
Without condom	751	1.08 (0.68, 1.73)	1.04 (0.64, 1.68)
Receptive anal sex			
With condom	750	0.87 (0.55, 1.38)	0.79 (0.50, 1.27)
Without condom	750	1.08 (0.67, 1.74)	0.96 (0.59, 1.56)

For these analyses, *valid* is the reference group

[a] Sample sizes are the number who responded to each item

[b] *RR* refers to the *rate ratio*

[c] *ARR* refers to the *adjusted rate ratio*, which was adjusted for age and race/ethnicity

[*] *P* < 0.05,

[**] *P* < 0.01,

[***] *P* < 0.001

Table 5

Odds of reporting HIV or STI test (last year) among invalid submissions in the SEM Study

	N	Tested n (%)	OR (95 % CI)	AOR[a] (95 % CI)
HIV[b]	1122	746 (66.5)	0.63 (0.44, 0.92)*	0.62 (0.42, 0.90)*
Chlamydia	1240	593 (47.8)	0.95 (0.67, 1.36)	0.89 (0.62, 1.28)
Gonorrhea	1241	614 (49.5)	0.94 (0.66, 1.34)	0.91 (0.63, 1.30)
Hepatitis A	1240	542 (43.7)	0.88 (0.62, 1.26)	0.88 (0.61, 1.27)
Hepatitis B	1240	549 (44.3)	0.92 (0.64, 1.31)	0.92 (0.64, 1.32)
Herpes	1240	481 (38.8)	1.00 (0.70, 1.44)	0.95 (0.66, 1.38)
HPV	1240	472 (38.1)	0.84 (0.58, 1.22)	0.79 (0.54, 1.16)
Syphilis	1241	616 (49.6)	0.94 (0.66, 1.33)	0.92 (0.64, 1.32)

For these analyses, *valid* is the reference group

[a] Model adjusted for age and race/ethnicity

[b] Individuals who reported being HIV-positive were not asked this question and are not included in this analysis

* $P < 0.05$,

** $P < 0.01$,

*** $P < 0.001$