Published in final edited form as:
AIDS Behav. 2008 November ; 12(6): 964–973. doi:10.1007/s10461-007-9352-2.

Identifying Multiple Submissions in Internet Research: Preserving Data Integrity

Anne M. Bowen✉,
Department of Psychology, University of Wyoming, Laramie, WY 82071, USA

Candice M. Daniel,
Department of Psychology, University of Wyoming, Laramie, WY 82071, USA

Mark L. Williams, and
Center for Health Promotion and Prevention Research, UT-H SPH, Houston, TX, USA

Grayson L. Baird
Department of Psychology, University of Wyoming, Laramie, WY 82071, USA

Abstract

Internet-based sexuality research with hidden populations has become increasingly popular. Respondent anonymity may encourage participation and lower social desirability, but associated disinhibition may promote multiple submissions, especially when incentives are offered. The goal of this study was to identify the usefulness of different variables for detecting multiple submissions from repeat responders and to explore incentive effects. The data included 1,900 submissions from a three-session Internet intervention with a pretest and three post-test questionnaires. Participants were men who have sex with men and incentives were offered to rural participants for completing each questionnaire. The final number of submissions included 1,273 "unique", 132 first submissions by "repeat responders" and 495 additional submissions by the "repeat responders" (*N* = 1,900). Four categories of repeat responders were identified: "infrequent" (2–5 submissions), "persistent" (6–10 submissions), "very persistent" (11–30 submissions), and "hackers" (more than 30 submissions). Internet Provider (IP) addresses, user names, and passwords were the most useful for identifying "infrequent" repeat responders. "Hackers" often varied their IP address and identifying information to prevent easy identification, but investigating the data for small variations in IP, using reverse telephone look up, and patterns across usernames and passwords were helpful. Incentives appeared to play a role in stimulating multiple submissions, especially from the more sophisticated "hackers". Finally, the web is ever evolving and it will be necessary to have good programmers and staff who evolve as fast as "hackers".

Keywords

Internet research; Multiple submissions; Repeat responders; Validity; Rural; MSM

Internet-based research with high HIV risk and hidden populations, such as men who have sex with men (MSM), is rapidly expanding (Pequegnat et al. 2006). The increase in Internet research has led a number of authors to call for examination of methodological issues related to internal and external validity (Kraut et al. 2004; Pequegnat et al. 2006). The anonymity of the Internet provides advantages and disadvantages in terms of internal validity (Kraut et al.

Department of Psychology, University of Wyoming, Dept. 3415, 1000 E. University, Laramie, WY 82071, USA, e-mail: abowen@uwyo.edu.

2004). Respondent anonymity may reduce pressure to respond in socially desirable ways, but may lead to unintended outcomes. Catania (1999) suggests that lowered presentation bias may result in more honest responses to sensitive questions. However, anonymity in online studies may lower self-regulation, increasing the possibility of developing multiple identities (Joinson 1998). If significant numbers of participants in online studies change their identity and enroll multiple times, the integrity of data will be severely compromised (Birnbaum 2004; Mustanski 2001; Pequegnat et al. 2006; Reips 2002a, b).

Early in the Internet research revolution, it was suggested that the prevalence of repeat responders (the number of people who enroll multiple times) and the frequency of multiple submissions (the number of times a single individual enrolls) were “rare” and “easy to detect” (Birnbaum 2000, 2004; Musch and Reips 2000; Reips 2002a, b). In these early studies, incentives were seldom used to encourage enrollment or increase retention (Birnbaum 2004). However, researchers acknowledged that incentives might increase the number of repeat responders and multiple submissions (Mustanski 2001). Despite the potential threats to internal validity, to our knowledge, only one assessment study has reported on the frequency of repeat responders and multiple submissions (Konstan et al. 2005). Konstan et al. identified 10% of the submissions in their online study as multiple submissions, and 55% of these came from one person. Clearly, additional research is needed to clarify the frequency of repeat responders and multiple submissions on Internet research.

Researchers have identified several variables for detecting multiple submissions from repeat responders (Birnbaum 2004; Nosek et al. 2002; Pequegnat et al. 2006; Reips 2002a, b). Recommendations include recording participants' Internet protocol (IP) addresses, eliciting personally generated passwords and usernames, and requesting personal data, such as names, phone numbers, and e-mail addresses. Another suggestion was to ask participants if they had previously participated in a study (Nosek et al. 2002) or to ask that those who have participated previously in the online study not to do so again (Mustanski 2001). Each of these recommendations has limitations. Although, IP address was found to be effective, many services have “floating” IP addresses, public computers may have the same IP, and participants may use multiple computers. User provided information may identify people who resubmit on a whim or by accident, but e-mail addresses, user names, and passwords can be easily changed. Asking repeat responders to identify themselves and to rely on their honesty may be naïve.

Incentives may increase both the number of repeat responders and frequency of multiple submissions because they appeal to individuals motivated by the offer of money rather than an interest in participating in the project. Unfortunately, as the novelty of online studies wanes and requirements on participants' time increase, incentives appear to be needed to encourage participation and retain participants across multiple sessions. Bowen (2005) found that rural MSM were more likely to complete an online survey if incentives were provided. Further, participants reported that monetary incentives motivated them to participate in online research. While researchers recommend providing participants with incentives to increase participation and retention in Internet-based studies (Michalak and Szabo 1998), little is known about the effects incentives have on promoting multiple submissions.

The goals of this manuscript were to: (1) identify the utility of several previously recommended variables and procedures for identifying multiple submissions, (2) identify demographic factors associated with multiple submissions, and (3) explore the effects of incentives on the tendency to provide multiple submissions. The data for this study were taken from the WRAPP study, an Internet delivered HIV risk reduction intervention project targeting rural MSM.

Methods

Procedures

WRAPP Study Procedures (Fig. 1)—Participants were recruited for the study using Internet banner ads displayed on a popular gay dating web site from June 2005 to June 2006. Men who screened as eligible were asked to read a consent form and to indicate informed consent by providing a unique username, password, and a valid e-mail address. They were also asked to provide contact information (i.e., name, phone number, alternate e-mail address), although these data were not required. Prior to activation, study personnel examined eligible participant account information for duplicates, including IP addresses, usernames, passwords, e-mail addresses, and telephone numbers.

Men were eligible to participate in the study if they were male, 18 years or older, identified as gay or bisexual or had had sex with a man in the 12 months before enrolling in the study. Eligible men were then routed into either the “paid” or “unpaid” group. The “paid” group consisted of men who lived in a rural area (i.e., a town of less than 75,000 and <60 min from an urban center; Bowen et al. 2004) and were identified as unique participants. The remaining eligible men (i.e., urban and/or duplicate) were routed to the “unpaid” group”.

Participants were sent an activation e-mail providing a link to the study (Pequegnat et al. 2006) that also informed them as to their payment eligibility. Participants then joined the project by accessing their account using their username and password. Complete study participation included a pre-test, three interventions, and a post-test after each intervention (Post-tests 1, 2, and 3). The design and timeline for study participation can be seen in Fig. 1. Incentives (e-mail gift certificates to a popular shopping web site) were provided within 24 h of completion of a questionnaire to compensate participants for their time. Incentive ranged from $15 to $30 per questionnaire, and a maximum of $90 could be earned for completing all questionnaires.

The accounts of men who met screening criteria and were identified as repeat responders were activated, but told they would not receive an incentive to participate in the study. The rationale for this approach included: (1) people could participate more than once if they wished, without reimbursement, (2) project personnel might have been mistaken about the submission being a duplicate, and (3) there was less chance that a participant might react negatively because of an accusation of multiple submissions. Accounts identified after activation as multiple submissions were “locked” and sent a message to contact the project due to a problem with their account. It was assumed that anyone who contacted the project could be reactivated. No ‘locked’ participants contacted the project. The study was approve by the University committee for the protection of human subjects.

Data Used for This Manuscript—The data used in this manuscript represent a *post-hoc analysis* of all respondents who completed the screening questionnaire, consent form, and provided a username, password, optional personal identifiers, and valid e-mail. These data were supplied prior to the project staff's activation of an account. Many multiple responders were identified prior to, or just after activation. Few participants in the non-reimbursement track completed the project and none of the ‘locked’ accounts were reactivated. As a result, we were unable to do a within-subject comparison of responses by the repeat responders. Informal examination of the completed questionnaires by multiple responders indicated that many questions were skipped and little usable data was generated, so these data were not used.

Terms Used in This Manuscript

Detection Variable—Variables used to identify multiple submissions.

Submission—A discrete set of information provided by an eligible participant, such as consent to study procedures, agreement with ‘terms of participation, and provision of a username, password, and valid e-mail address.

Multiple Submission—A submission provided by a participant who has provided a previous submission.

Suspect Submission—A submission with one variable in common with a prior submission (e.g., IP address, telephone number, username, etc).

Repeat Responder—A participant who joins the project more than one time by submitting multiple times.

Detection Variables

Internet Protocol Address (IP)—The IP is a computer address that includes four sets of numbers with one to three numbers per set (e.g., 204.191.66.148). It is used to identify and communicate with other computers on a network. An IP address is similar to a street address or phone number that uniquely identifies a building or telephone. The analogy to a telephone system would be the use of a long-distance phone card. IP addresses can be shared by a number of persons because proxy servers can act as an intermediary agent between home computers and servers. Additionally, two or more people in the same household may share an IP address. The use of the IP alone as a method for identifying duplicate accounts is the most conservative method, but it may eliminate many people in rural areas who use a shared intermediary server.

Web Browser—A Web browser is a software application that enables a user to display and interact with text, images, and other information typically located on a web page. Browser information was collected automatically.

E-mail—Eligible participants were required to provide a valid e-mail address. Participants could use an existing e-mail address or generate a new one using any service provider. If a submission shared the same e-mail address or a similar distinct e-mail address (i.e., kulguy84@somewhere.com and kulguy85@somewhere.com) with another submission, it was identified as a multiple submission.

Username—Each participant generated a personal username as part of the consent process. There were no restrictions on usernames in terms of the use of numbers, letters, or characters. If a submission shared or had a similar distinct username with another submission (i.e., nycguy84 and nycguy85), it was identified as a multiple submission. If a submission shared a common username (i.e., 123456), it was not considered a multiple submission.

Password—All participants were required to generate a personal password, and there were no restrictions on passwords in terms of the use of numbers, letters, or characters. If a submission shared a common password (i.e., 123456) with another submission, it was not considered a multiple submission. If a submission shared a distinct password (i.e., coolguy84) or a similar distinct password (i.e., coolguy85) with another submission, it was considered a multiple submission.

Telephone Number—Participants were asked to provide a telephone number for purposes of follow-up and retention, but they were not required. We used telephone numbers to check the authenticity of suspicious submissions through “reverse look-up” on the Internet. For example, if a large group of submissions shared a common IP address and the user provided telephone numbers which shared a common theme (i.e., veterinary clinics, sandwich shops) they were identified as multiple submissions from the same responder. If a submission shared the same telephone number with two or more submissions, it was considered a multiple submission.

Additional Variables Used to Explore Multiple Submissions

Zip Code—Zip code was not coded independently as a detection variable, because a number of participants from a small town might share a zip code. Zip code was used with suspect submission to look for correspondence with telephone area code.

Last Name—If a suspect submission reported a unique last name that was the same as other suspect submissions it was coded as a multiple submission.

Incentives—The paid/unpaid designation (i.e., rural/unique) was computed automatically by information presented in the screening questions. Questionnaire completion and reimbursement for a questionnaire were automatically coded as “1” for yes and “0” for no.

Analyses

Identification of Multiple Submissions—The data set consisted of 1,900 submissions. The data used for this manuscript represent a post-hoc analysis of all submissions, although detection of potential fraudulent accounts was ongoing during the study. The rationale for the post-hoc analysis is that data from all participants who met study inclusion criteria, completed the consent and provided contact information could be examined in more depth. This would allow identification of the most useful detection variables for future studies.

Two independent coders examined the data set for repeat responders. Each coder began by sorting the data set by an individual “fraud detection” variable and labeling identical variables with a unique number to identify them as a duplicate. For example, the data were first sorted by IP address and the first set of identical IP addresses were coded “1”. The second set of identical IPs was coded “2” and this continued through all 1,900 submissions. The data were then resorted by the next identifying variable and coded in the same way until all “detection” variables had been coded independently.

Final identification of multiple submissions began by resorting the data by IP and examining it for duplications across all detection variables. Submissions that had at least two identical detection variables were coded as multiple submissions. Submissions with only a duplicate IP or telephone number or IP that varied by one or two digits were labeled as “suspect.” “Suspect” submissions were then examined in depth by using a reverse telephone lookup, correspondence of zip code and area code, similar usernames and passwords. Finally, browser was examined within a set of multiple submissions and a unique browser might result in removing the submission from the “duplicate” group.

Agreement between coders was computed prior to resolution of suspect submissions. The Kappa statistic was chosen because it compares agreements to what might be expected by chance. The overall Kappa was .75 (s.e. = .01), which is considered substantial. After resolution of suspect submissions, the data used in the manuscript is from Coder 2.

Participant Typologies—After resolution of the suspect submissions, all 1,900 submissions were coded into three categories. The “unique” category included respondents who set up only one account (N = 1,273). The second category included the first submission by a repeat responder and was labeled “repeat responder” (N = 132). The third category included the additional submissions by a “repeat responder” and was labeled “multiple submission” (N = 495).

"Multiple submissions" were coded into four categories based upon the number of submissions. First, the "Infrequent" category included those with 2–5 submissions. The "Persistent" category included those with 6–10 multiple submissions. The "Very Persistent" category included those with 11–15 submissions. Finally, the "Hacker" category included those with more than 15 submissions.

Demographics—Participant demographic characteristics included age, ethnicity, relationship status, education, income, and work status. Age was collected as a continuous variable. Participants checked their self-identified ethnicity using the following categories: African American, Asian or Asian Pacific Islander, Caucasian, Hispanic/Latino, and "other". These were then reduced to three groups, Caucasian, Hispanic, and other due to low numbers of minority participants. Relationship status included single, never married; living with same sex partner, and living with opposite sex partner or divorced. Education included two categories: high school diploma or less and some college or more. Income was reported as annual income and included 4 categories: <$15,000; $15,000–$24,999; $25,000–$49,999; and ≥$50,000. Finally, work status included full-time, part-time, or unemployed/retired. Screening questions were used to designate a participant as rural or urban. The rural designation required the participant to live in a town of 75,000 or less and live at least 60 minutes drive from an urban center.

Demographic Comparison of "Unique" and "Repeat Responders"—A forward stepwise conditional logistic regression was used to identify demographic characteristics that might differential "unique" and "repeat" responders (Table 3). The demographic variables included age as a continuous variable and six categorical variables (income, work situation, relationship status, ethnicity, education, and urban/rural designation). Urban or rural residence was determined by responses to screening questions and participants who were identified as "rural" and "unique" were offered incentives and those identified as "urban" and/or "repeat responders" were told they would not receive an incentive. Age was also queried in the screening questions, but all other demographic characteristics were queried in the pretest questionnaire, resulting in valid data for 701 "unique" responders and 104 "repeat" responders.

Results

Utility of detection variables (Table 1)

The following section examines each detection variable individually in terms of unique and duplicate values. Each detection variable is discussed independently of the other variables. It should be noted that in the following paragraphs and Table 1, the sum of the unique and duplicate cases do not total the number of cases in the data set. The reason for this is that the numbers of duplicates represent a unique individual, not the total number of submissions for each individual.

IP addresses and web browser information were collected automatically with each submission. Of the submissions, 1,317 had unique IP addresses and 152 included multiple cases from the same address. Eighty five percent of the duplicate IP addresses occurred between two and five times. Four IP addresses were responsible for 17, 18, 20, and 27 submissions, respectively. Web browser had the least number of unique submissions with only 211 and the second most frequent number of duplicates with 128. Sixty percent of the duplicate browsers were represented twice, while 20 browser addresses were represented more than 15 times, with two occurring 177 and 361 times, respectively. The limited number of different browsers led us to reject the Web Browser as a useful independent detection variable. Browser was used to provide information about a potential duplicate account if a group of submissions shared one IP address, but one of theses had a unique browser. This submission would be considered unique and different from the group.

Participants were required to provide a username, password, and e-mail. There were 1,827 unique usernames and 26 duplicated names. Twenty-five of the duplicates occurred between two and five times. There were 1,560 unique passwords and 112 duplicates. One hundred-five of the duplicates occurred two to five times, and only one password occurred more than 15 times. E-mail addresses fell between username and password with 1,704 unique e-mails and 79 duplicates. No individual e-mail address occurred more than 11 times.

Participants were asked, but not required to provide a contact telephone number. Only 32 telephone numbers were clearly invalid, with non-existent area codes or too few numbers. There were 1,700 unique numbers and only 84 were duplicated. Of the 84 duplicates, all but one occurred two to five times.

Examination of Repeat Responders (Table 2)

Overall, there were 1,900 submissions, 1,273 of which were identified as unique submissions. There were 132 repeat responders who attempted to join the project 2 or more times. The repeat responders provided 495 submissions in addition to their first submission. The “infrequent” category accounted for 36% of the multiple submissions and included 114 individuals who joined the project between two and five times. The majority of this group tried two (53.8%) or three times (18.9%). As seen in Table 2, more than 68% of individuals in the “infrequent” category had the same IP address. Approximately 80% provided identical information in terms of one or more personal identifiers (e.g., e-mail, username, password, or telephone number).

The “persistent” category consisted of 10 men who joined the project between 6 and 10 times, providing 64 submissions in addition to their first submission. They accounted for 13% of the multiple submissions. Eight of these men had the same IP address for each submission, and four provided between one and three identical personal identifiers.

The “very persistent” category included five men who joined the project between 11 and 30 times, producing 86 submissions in addition to their first submission. This group of five individuals account for 17% of the multiple submissions. Each of these men had the same IP address for each submission and three provided three or four duplicate variables. Examination of the personal identifiers often showed patterns that would not be detected as identical, but were clearly provided by the same user. For example, using an e-mail address that included an identical name plus a number that increased incrementally with each submission (e.g., name34@somewhere.com and name35@somewhere.com). Other patterns included using variations on state or city names or text in an e-mail address that matched another username. For example, participant one would have the username “foxyman” and participant two's e-mail address would be foxyman@somewhere.com.

The “hacker” category included only three participants, but these joined the project 46, 59, and 67 times, respectively, and provided 169 (34%) multiple submissions. Each “hacker” had identical or nearly identical IP addresses, but two had no additional identical detector variables. The participant with 67 submissions was identified using a combination of detection variables, including two passwords that were used frequently during his earliest submissions, telephone numbers that were linked to mini-marts and large multipurpose stores, and e-mails that were similar to other usernames. The remaining two “hackers” were the most difficult to identify. Many of their IP addresses were the same, but there were also slight variations with single digit changes. The repeat responder with 46 multiple submissions used a similar password or username only twice, but many of these multiple submissions occurred on one day often within minutes of the previous submission. Reverse telephone lookup for this individual was not particularly useful in that most numbers appeared legitimate, such as unlisted cell phones and private telephones from a variety of states and reported area codes and zip codes were congruent. The participant who supplied 59 submissions had a number of different, although similar IPs. All the telephone, e-mail, password, zip codes and usernames were unique, but reverse telephone look-up revealed that all numbers were to veterinary hospitals and humane associations across the country.

Demographic Characteristics that Differentiate “Unique” and “Repeat Responders”

The “unique” or “repeat responder” designation of the first submission of all eligible participants was used as the dependent variable in a logistic regression. Results of the logistic regression can be seen in Table 3 including regression coefficients, Wald statistics, odds ratios, and 95% confidence intervals for significant predictors. “Paid” or “Unpaid” group entered first with an improvement X^2 (df = 1) = 56.27, P = .000. Age entered second with an improvement X^2 (df = 1) = 4.40, P = .036 and relationship status entered third with an improvement X^2 (df = 2) = 6.98, P = .03, for an overall Model X^2 (df = 4) = 67.65, P = .000. The Nagelkerke pseudo R^2 suggests that the full model accounts for only 15% of the variance, 12.6% of which was accounted for by the paid/unpaid variable. The prediction success did not improve with the addition of the three significant predictors, resulting in none of the repeat responders being correctly classified. The odds ratio for predicting the “unpaid” designation indicates that men who were in the “paid” group more likely to be repeat responders. The odds ratio for age was .96, indicating that younger men were slightly more likely to submit multiple times. In terms of relationship status, the odds ratios for men who submitted multiple accounts was 2.36 for those living with opposite sex partners or divorced and 1.70 for those living with a same sex partner indicating that men who had partners were more likely to be multiple submitters than men who were single or never married to be repeat responders.

Incentive Effects

Examining 783 unique participants who were designated as “unpaid”, only 2.9% were in the repeat responder category. In contrast, there were 622 submissions that were eligible for reimbursement and 17.4% of those were repeat responders. The logistic regression above supports this finding in that the paid designation increases the likelihood that a participant will submit additional multiple accounts as much as six times.

Second, we examined the likelihood that the overall frequency of questionnaire reimbursement would be greater for those repeat responders who submitted more submissions than for those who gave up after two or three submissions. Lack of reimbursement may be due to failure to complete a questionnaire, designation as “unpaid” (i.e., urban or identified by project staff as a multiple submission and having their account locked). Table 4 shows the four categories of repeat responders in terms of the percentage of questionnaires for which repeat responders received compensation. Each row and column is independent, thus none sum to 100%. Examination of row 1 indicates that of the 176 “infrequent” submissions, 11.9% were reimbursed for completion of the pre-test, and the percent reimbursed declined across the next three points in time to 5.1%. Rows 2 and 3 show that nearly three times the percentage of the “persistent” and “repeater” submissions were reimbursed at pre-test than the “infrequent” submissions. While the percent reimbursement declines across time for both groups, a much higher percent continue to be reimbursed than those in the “infrequent” category. Finally the submissions by the “hacker” category indicate that 84.1% of the submissions by the three repeat responders were reimbursed at pre-test and, while this declines across time, 31.4% still received an incentive for post-test 3.

Discussion

In the last 10 years, research using the Internet has exploded. The Internet has been touted as an excellent venue for reaching hidden populations because it is easily accessible, affordable, and participation can be anonymous. While anonymity may be useful, it may also inhibit researchers' ability to identify participants who attempt to join a study multiple times. Early studies suggested that the frequency of multiple submissions in Internet-based sexuality research was rare (Birnbaum 2004; Reips 2002a, b) and probably related to the use of incentives to encourage participation.

The use of incentives in this study appears to have affected both the number of repeat responders and the number of submissions by an individual repeat responder and, like the Konstan et al. (2005) study; we found a high frequency of multiple submissions. The men designated as eligible for compensation were six times more likely to be repeat responders than the ‘unpaid’ group. While the incentives in this study do not appear especially high, the frequency of repeat responders and the high rate of reimbursement for those with the most submissions suggest that the promise of even $15 may increase the rate of spurious submissions. Additionally, as the sophistication of the repeat responders increased (i.e., “hackers”), so did the likelihood of reimbursement. These participants are the most difficult to detect, likely to make the most money, and to jeopardize the internal validity of the study. Early identification of repeat responders, prior to activation may reduce the incentive to try again, as seen by the low multiple submission rates of unpaid participants and the low reimbursement rate for the “infrequent” group. Failure to detect multiple submissions early can increase the cost of an Internet study, both in terms of dollars and valid data. Finally, in this study, failure to identify the repeat responders could have resulted in as much as 25% fraudulent data.

Identification of repeat responders requires a combination of programming and human interaction with detection variables. Our data suggest that procedures for identifying multiple submissions from repeat responders cannot be fully automated. Internet studies will need considerable involvement by well-trained personnel for both *a priori* and *post-hoc* examination of the data. The simplest means of reducing multiple submissions would be to rely on the honesty of participants, as suggested by Reips (2002a, b) and Mustanski (2001). In the consent section of this study, participants were provided a “terms of participation” agreement that mentioned participating only once and, like software licensing agreements, they were required to click “I Agree” to continue with the study. This did not deter the repeat responders. Further, the frequency and timing of multiple submissions supports the notion that these men were intentionally providing fraudulent data. If the study by Ross et al. (2006) is any indication, lying online may be fairly common, and it should not be expected that individuals attempting to gain access to a study to earn an incentive would be motivated by a plea for honesty.

While the published studies suggest that floating IP addresses may be increasingly common, similar and identical IP addresses was the most useful detection variable. Programs should include a system that not only alerts project staff to submissions from identical IP addresses, but also to ones that are nearly identical. IP address are composed of four sets of numbers and the last two sets varied most often in the “hacker” group. Similarly, if possible, programming might include identification of user-supplied information that varies in systematic ways such as adding a number to a single stem, such as bob1, bob2, bob3, etc. Once similar detection variables are identified, staff should exam user supplied information for patterns across usernames, passwords and e-mail addresses. Prior to activation, suspect cases may require reverse telephone look-up and examination of resubmission intervals as in the Konstan et al. (2005) study. In studies with multiple incentives, as in this study, it would be advisable to “flag” suspect submissions and have them checked regularly for data anomalies against other submissions with similar personal information.

Although a profile of repeat responders would be helpful, the age and relationship differences between repeat and unique responders were not sufficient to create a profile. Nevertheless, giving more scrutiny to submissions by younger persons with partners may provide some benefit. Younger men may be more computer savvy and thus think about the possibility that online sites might be tracking IP addresses and reviewing e-mail addresses. Particularly relevant for this study, some MSM, especially those who wish to maintain their anonymity, may change their names and e-mail addresses when accessing online dating websites. Using similar tactics when participating in online research studies would not be unusual. Additionally, the men with partners may be working together or possibly have more time to access research studies when their partners are absent. Hacking into web sites for money may provide amusement as a group or possibly relief from boredom.

The study has a number of limitations. The sample is restricted to men who report sex with other men who responded to an advertisement for an HIV risk reduction intervention with payment for those who qualified. First, rural residence is confounded with qualifying for payment. While this criterion was not made clear to participants, it makes data interpretation more difficult in terms of incentive effects. It is possible that living in a rural area was the incentive for multiple submissions. Other studies (Bowen 2005) as well as the analyses of incentive effects in this study support the effect of money on multiple submission, but further research should examine the potential for a “rural” effect. It is also unclear how different samples for other types of studies would respond, both in terms of sample characteristics and to the payment issue. There is a need for studies that examine different effects of incentives on multiple submissions, including different schedules for delivering the incentives, and different types of incentives. Finally, while our system was programmed to automatically identify a number of variables that were identical across participants, clearly the hackers were finding new ways to get around the system. Therefore, while the recommendations in this study should be helpful for identifying repeat responders, researchers should remember that the web is ever evolving and it will be necessary to have good programmers who are evolving as fast as hackers.

Acknowledgements

The project was supported in part by a grant from the National Institute of Mental Health (MH-63667); The Wyoming Rural AIDS Prevention Projects. All work and opinions herein are solely those of the authors. I would also like to thank reviewers for providing very useful and easy to follow suggestions that have resulted in significant improvements in the manuscript.

References

Birnbaum, MH. Psychological experiments on the Internet. San Diego: Academic; 2000.

Birnbaum MH. Human research and data collection via the Internet. Annual Review of Psychology 2004;55:803–832.

Bowen AM. Internet sexuality research with rural MSM: Can we recruit and retain them? Journal Sex Research 2005;42(4):317–323.

Bowen AM, Williams MW, Horvath K. Using the Internet to recruit rural MSM for HIV risk assessment: Sampling issues. AIDS & Behavior 2004;8(3):311–319. [PubMed: 15475678]

Catania J. A comment on advancing the frontiers of sexological methods. The Journal of Sex Research 1999;36:1–2.

Joinson, AN., editor. Causes and implications of disinhibited behavior on the Net. NY: Academic Press; 1998.

Konstan, J.; Rosser, BRS.; Ross, M.; Stanton, J.; Edwards, W. The story of subject naught: A cautionary but optimistic tale of Internet survey research [Electronic Version]. Journal of Computer-Mediated Communication. 2005. Retrieved May 28, 2007 from http://jcmc.indiana.edu/vol10/issue2/konstan.html

Kraut R, Olson J, Banaji M, Bruckman A, Cohen J, Couper M. Psychological research online. American Psychologist 2004;59(2):105–117. [PubMed: 14992637]

Michalak E, Szabo A. Guidelines for Internet research: An update. European Psychologist 1998;3:70–75.

Musch, J.; Reips, UD. A brief history of Web experimentation. San Diego: Academic; 2000.

Mustanski BS. Getting wired: Exploiting the Internet for the collection of valid sexuality data. The Journal of Sex Research 2001;38(4):292–301.

Nosek BA, Banaji MR, Greenwalk AG. E-research: Ethics, security, design and control in psychological research on the Internet. Journal of Social Issues 2002;58(1):161–176.

Pequegnat W, Rosser BRS, Bowen AM, Bull SS, DiClemente R, Bockting WO, et al. Conducting Internet-based HIV/STD prevention survey research: Considerations in design and evaluation. AIDS & Behavior 2006;11:505–521. [PubMed: 17053853]

Reips, UD., editor. The Web experiment method: Advantages, disadvantages, and solutions. San Diego: Academic; 2000.

Reips UD. Internet-based psychological experimenting: Five dos and five don'ts. Social Science Computer Review 2002a;20(3):241–249.

Reips UD. Standards for Internet-based experimenting. Experimental Psychology 2002b;49(4):243–256. [PubMed: 12455331]

Ross M, Rosser B, Coleman E, Mazin R. Misrepresentation on the Internet and in real life about sex and HIV: A study of Latino men who have sex with men. Culture, Health & Sexuality 2006;8:133–144.

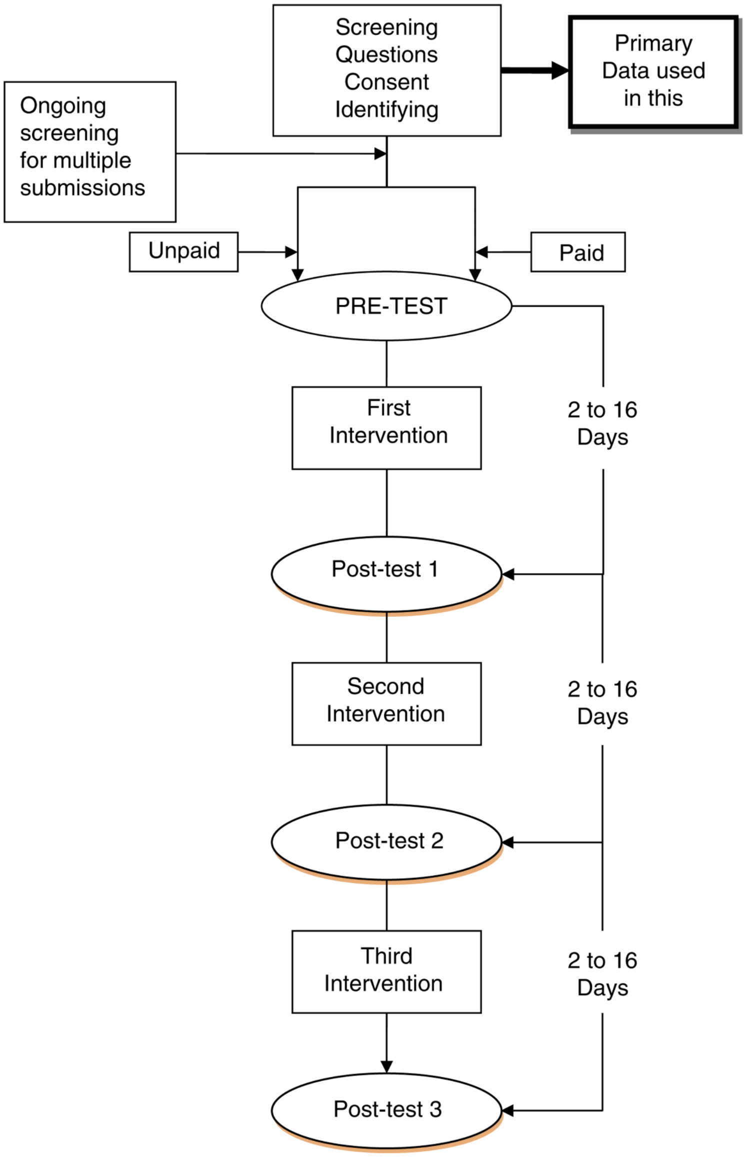

Fig. 1.

Flow chart of the overall HOPE study design. Screening for multiple submissions was ongoing during the study. Many were identified prior to activation and routed to non-payment group. If a multiple submission was identified after activation, the account was locked with a request to contact the project. Data used for this manuscript represents a post hoc analysis of all 1,900 participants who completed the screening, consent, and identifying information page

Table 1

Utility of detection variables

Detection variable	# Unique	# Duplicates	Number of times a duplicate detection variable appeared				*M* (*SD*)	Maximum[a]
			2–5, %	6–10, %	11–15, %	>15, %		
Automatically recorded variables								
IP address	1,317	152	84.9	8.3	4.2	2.8	3.84 (3.76)	27
Browser	211	128	59.8	18.6	6.6	14	13.20 (36.59)	361
Required user determined variables								
Password	1,560	112	92.9	4.5	1.85	.9	3.04 (2.30)	17
Username	1,827	26	96.6	0	3.4	0	2.81 (1.98)	12
E-mail	1,704	79	97.4	2.6	1.3	0	2.23 (1.26)	11
Optional user provided variables								
Phone	1,699	84	98.9	0	1.1	0	2.39 (1.24)	12

[a] Maximum number of times a specific variable was duplicated

Table 2

Categorization of repeat responders by number of submissions and number of identical detection variables

Participant typology	Number of submissions	Percent of all multiple submissions (*N* = 627)	Number of all duplicate responders (*N* = 132)	Percent identical IP, %	Percent personal identifiers duplicated				
					0[a], %	1[a], %	2[a], %	3[a], %	4[a], %
Infrequent[b]	2–5	36	86.3	68.4	20.2	36.8	25.4	16.7	.9
Persistent[b]	6–10	13	7.7	80.0	50.0	10.0	20.0	20.0	0
Repeater[b]	11–30	17	3.9	100	40.0	0	0	20.0	40.0
Hackers[b]	45–67	34	2.4	100	66.7	0	0	33.3	0

[a] Sum of identical detection variables (password, username, telephone, e-mail) after IP address within each fraudulent account. For example, “0” means none of the personal identifiers matched and “2” means there were two personal identifier that were the same within a multiple submission

[b] Rows represent independent analyses for each participant typology

Table 3

Stepwise logistic regression of demographics predicting unique versus repeat responder status

Demographic variables	Regression coefficient	Wald	Odds ratio	95% Confidence interval
Age (continuous)	−.04	7.46**	.96	.94–.99
Ethnicity (Caucasian, Hispanic, Other)	n.s.			
Relationship		7.26*		
Single, Never married	Referent			
Living with same sex partner	.53	3.14	1.70	.95–3.05
Living with opposite sex partner/divorced	.86	5.65*	2.36	1.16–4.81
Education (high school, college)	n.s.			
Work status (full time, part time, unemployed)	n.s.			
Income (<15,000, 15,000–24,999, 25,000–49,999, ≥50,000)	n.s.			
Paid versus unpaid	1.80	37.72***	6.04	3.40–10.73
	X^2 (df = 4) = 67.65, P<.000, Nagelkerke R^2 = .15			

* $P < .05$,

** $P < .01$,

*** $P < .000$

Table 4

Longitudinal reimbursement history for each category of repeat responder, including pre-test and three post-intervention questionnaires

Category of repeat responder	Total number of multiple submissions	% Of multiple submissions reimbursed for completion of each questionnaire			
		Pre-test	Post-test 1	Post-test 2	Post-test 3
Infrequent	176	11.9	7.8	6.8	5.1
Persistent	64	31.2	28.1	28.1	25.0
Repeater	86	30.2	20.9	16.3	15.1
Hacker	169	84.1	47.3	37.9	31.4