

Field Methods

<http://fmx.sagepub.com/>

Data Quality in HIV/AIDS Web-Based Surveys: Handling Invalid and Suspicious Data

Jose A. Bauermeister, Emily Pingel, Marc Zimmerman, Mick Couper, Alex Carballo-Diéguez and Victor J. Strecher

Field Methods 2012 24: 272 originally published online 26 April 2012

DOI: 10.1177/1525822X12443097

The online version of this article can be found at:

<http://fmx.sagepub.com/content/24/3/272>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Field Methods* can be found at:

Email Alerts: <http://fmx.sagepub.com/cgi/alerts>

Subscriptions: <http://fmx.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations: <http://fmx.sagepub.com/content/24/3/272.refs.html>

>> [Version of Record](#) - Aug 10, 2012

[OnlineFirst Version of Record](#) - Apr 26, 2012

[What is This?](#)

Data Quality in HIV/ AIDS Web-Based Surveys: Handling Invalid and Suspicious Data

Jose A. Bauermeister¹, Emily Pingel¹, Marc Zimmerman¹, Mick Couper², Alex Carballo-Diéguez³, and Victor J. Strecher¹

Abstract

Invalid data may compromise data quality. We examined how decisions made to handle these data may affect the relationship between Internet use and HIV risk behaviors in a sample of young men who have sex with men (YMSM). We recorded 548 entries during the 3-month period and created six analytic groups (i.e., full sample, entries initially tagged as valid, suspicious entries, valid cases mislabeled as suspicious, fraudulent data, and total valid cases) using data quality decisions. We compared these groups on the sample's composition and their bivariate relationships. Forty-one cases were

¹ Department of Health Behavior & Health Education, School of Public Health, University of Michigan, Ann Arbor, MI, USA

² Institute for Social Research, University of Michigan, Ann Arbor, MI, USA

³ HIV Center for Clinical and Behavioral Studies, Columbia University and New York State Psychiatric Institute, New York, NY, USA

Corresponding Author:

José A. Bauermeister, Department of Health Behavior & Health Education, School of Public Health, University of Michigan, 1415 Washington Heights, SPH I Room 3822, Ann Arbor, MI 48109, USA

Email: jbauerme@umich.edu

marked as invalid, affecting the statistical precision of our estimates but not the relationships between variables. Sixty-two additional cases were flagged as suspicious entries and found to contribute to the sample's diversity and observed relationships. Using our final analytic sample, we found that very conservative criteria regarding data exclusion may prevent researchers from observing true associations. We discuss the implications of data quality decisions and its implications for the design of future HIV/AIDS web surveys.

Keywords

Internet methodology, web-survey design, HIV/AIDS, bias, data quality

Internet interactions may result in the development of social relationships such as romantic partnerships and sexual encounters. Researchers have taken advantage of these social interactions to examine the relationship between Internet use and HIV/AIDS risk. Men who have sex with men (MSM) who meet sexual partners over the Internet may engage in HIV risk behaviors, including having unprotected anal intercourse with one or more partners of unknown or serodiscordant HIV status (Garofalo et al. 2007). These associations are not surprising, given the Internet's affordability, acceptability, anonymity, and accessibility for seeking partners (Ross 2005). However, researchers have suggested that data collection methods need to be critically considered for measuring relationships between Internet use and HIV risk outcomes (Pequegnat et al. 2007). Given the increasing adoption of web-based data collection in HIV/AIDS research, it is becoming increasingly important to make sure data collected through this modality are valid and reliable. As a contribution to this literature, we examine how decisions regarding handling multiple entries in a web survey may influence data quality. We look at how the relationship between young MSM's sexual behaviors and their Internet use to meet sex partners varies, depending on what data quality decisions are made.

Data Quality Considerations in HIV/AIDS Web Research

The Internet is a rich environment to study HIV/AIDS risk behaviors among hard-to-reach populations (Pequegnat et al. 2007). Web-based data collection is faster and cheaper, reduces participant burden, provides greater flexibility in measuring responses with complex skip patterns than face-to-face or telephone surveys, and may elicit greater variability in responses (Bowen

et al. 2008; Couper 2008). Several methodological concerns, however, may threaten the internal and external validity of these web surveys. While some concerns reflect broader theoretical and methodological issues pervasive in HIV/AIDS research among MSM (i.e., low response rates) and inability to elicit representative samples due to the absence of a population-sampling frame (Pequegnat et al. 2007; Ross et al. 2004), others are specific to Internet data collection (i.e., ensuring adequate security protocols to protect responses during data collection and minimizing duplicate and fraudulent entries). In a web survey of HIV risks among Latino MSM, for example, Konstan and colleagues (2005) found that 11% of entries in their sample ($N = 1,150$) were duplicate entries from participants. Similarly, Bowen and colleagues (2008) found that in a prospective web-based intervention study focused on HIV prevention among MSM living in rural areas, approximately one-third of the 1,900 total submissions were multiple entries. Taken together, these findings suggest that invalid entries, whether due to deliberate falsification or misrepresentations or to inadvertent (duplicate) entries, occur commonly in web-based research.

Based on Wang and Strong's (1996) multidimensional framework to assess data quality, we focus on the intrinsic component of data quality; that is, the concern that data collected is accurate and believable and therefore lead to reputable results. When the high incidence of multiple entries is high, for example, these data may threaten the internal validity of a web survey and the inferences derived from them (Bowen et al. 2008; Konstan et al. 2005; Kraut et al. 2004; Pequegnat et al. 2007). The presence of these data may artificially change the variability within variables of interest and lead to biased relationships between study variables and to incorrect programmatic recommendations. Even when these data may be identified and excluded, the final sample size may be smaller than planned, which may reduce statistical power. Invalid data may also increase research costs if incentives are disbursed to individuals who were ineligible or who had already participated and if personnel time must be allocated to monitor and identify errors. In this study, we illustrate how invalid data and the strategies taken to remove these data may alter observed findings.

Automated and manual procedures have been suggested to minimize invalid web-survey data. Automated procedures may serve to identify invalid data a priori. Automated procedures may include data checks that can be programmed into web surveys. These procedures include requiring participants to complete an eligibility screener, providing eligible participants with a password to access the questionnaire, accepting only one submission

from an Internet Protocol (IP) address, cross-checking submissions entered from a similar IP address using personal indicators (e.g., e-mail address, telephone number), and/or flagging cases that provide erratic patterns in their answers (e.g., missing or out-of-range responses). Manual approaches may also be employed post hoc. These approaches may include cross-checking patterns within submitted e-mail addresses or other personal information, verifying flagged cases through online databases such as social networking sites and IP reverse look-up applications, and/or assessing irregular variations in measures of central tendency.

None of these approaches is without flaws and may affect a study's costs, including hiring additional study personnel (e.g., programmers) or expansion of existing staff's work responsibilities (e.g., manual data checks), purchasing a dynamic web-survey platform for customized programming, and subscribing to a reverse IP address service. Furthermore, as the techniques to identify invalid entries become more sophisticated, so do the methods that circumvent these safeguard techniques. Consequently, we sought to examine whether the association between Internet use to meet sex partners and HIV risk outcomes would vary based on quality checks performed in a sample of young MSM who completed a web survey examining their online dating experiences.

Study Objectives

This study had two objectives. First, we examined the relationship between YMSM's (young men having sex with men) Internet use to seek casual partners and their sexual behaviors in the past 2 months. Consistent with previous literature, we hypothesized that greater Internet use to meet sex partners would be associated with greater number of partners and unprotected receptive anal intercourse (URAI) occasions, respectively. Second, we tested how data quality procedures commonly employed in online questionnaires could influence these observed relationships. We hypothesized that the associations between Internet use and HIV risk behaviors would vary once invalid data were removed from the analyses.

Methods

Sample

Data for this article come from a cross-sectional study examining the association between YMSM's online partner-seeking behaviors and HIV/AIDS risk behaviors (Bauermeister et al. 2011). Four criteria for

eligibility included being 18 and 24 years old, having used a dating website in the past 3 months, having been sexually active with a male partner met on a dating website in the past 6 months, and being single. Participants were primarily recruited through online advertisements, participant referrals, and flyers posted at local venues frequented by YMSM. Social network advertisements were viewable only to men who fit our age range and who lived in the United States. Promotional materials displayed a synopsis of eligibility criteria, incentives, and the survey's website.

Procedures

Study data were protected with a 128-bit SSL encryption and kept within a University of Michigan firewalled server. Upon entering the study site, participants were asked to enter a valid and private e-mail address, which served as their username. This allowed participants to save their answers and, if unable to complete the questionnaire at one sitting, continue the questionnaire at a later time. Participants were then asked to answer four questions to determine their eligibility. If eligible, participants were presented with a detailed consent form that explained the purpose of the study and their rights as participants. YMSM were asked to acknowledge that they read and understood each section of the consent form, respectively.

Consented participants then answered a 30- to 45-minute questionnaire that covered assessments regarding their sociodemographic characteristics, HIV status, Internet use, relationship ideals, sexual and substance use behaviors, and general mood over the last few months. Participants with incomplete questionnaires were sent two reminder e-mails, one a week after they had commenced the questionnaire and another a week before the study was scheduled to close, encouraging them to complete the survey. Participants were compensated with \$15 in iTunes gift cards via e-mail upon completion of the questionnaire. We acquired a Certificate of Confidentiality to protect study data. The Institutional Review Board of the University of Michigan approved all procedures.

Measures

Sexual behavior. Respondents were asked to report their sexual behavior with men and women during the previous 2 months using the Sexual Practices Assessment Schedule (Bauermeister et al. 2011). Questions were

posed both in formal language and vernacular (in italics) to increase comprehension. Questions regarding the number of male partners and occasions of URAI with men (e.g., “How many times did a male partner put his penis in your rectum without a condom?”) were most relevant for this study. We assigned a value of zero to participants who did not have URAI.

Internet use to meet sex partners. Using an open-ended response, participants were asked: “On average, how many hours a week do you spend online looking for a sexual partner with whom to hookup?”

Demographic characteristics. Respondents were asked to report their age (in years), highest level of education completed (1 = *eighth grade or less*, 2 = *some high school*, 3 = *graduated high school*, 4 = *technical school*, 5 = *associate degree*, 6 = *some college*, 7 = *college*, 8 = *some graduate school*, 9 = *graduate school*), and current occupation. Respondents were asked to report if they considered themselves of Latino or Hispanic ethnicity and if they were African American or black, Asian or Pacific Islander, white or European American, Native American, or Other. We combined the Native American and Other race categories given the limited number of observations, and created dummy variables for each race/ethnicity group. Whites served as the referent group in our analyses.

Data Analytic Strategy

We first created four different analytic subsamples. To create the subsamples, we identified suspicious cases based on whether more than one entry came from the same IP address and/or a similar e-mail address (Bowen et al. 2008). Given that shared IP addresses may reflect communal spaces (e.g., library or coffee shop) and households where more than one participant was eligible (e.g., roommates), we identified valid and invalid cases using additional criteria: (1) comparing e-mail addresses (e.g., duplicate@gmail.com and duplicate@hotmail.com); (2) verifying suspicious response patterns within the questionnaire (e.g., participant reports “Don’t Want To Answer” to all questions); (3) assessing whether participants had reported living alone or with others; (4) cross-checking their e-mail address with profiles in public social network accounts to verify authenticity; and/or (5) determining whether a questionnaire was completed in a less than realistic time frame (i.e., <15 minutes). From the

548 entries recorded in our survey, we excluded data from 60 cases that created an account but did not begin the questionnaire (i.e., full missing data), as these data may be attributable to participants' nonresponse or decision to complete the questionnaire using a different account (e.g., a personal account instead of a school or work e-mail address). Of the 488 participants who consented, 31 had partial missing data (26 cases subsequently identified as "valid"); most often, partially missing data occurred once participants approached the end of the survey. Consequently, the overall completion rate was 89.05% ($N = 488$).

We then computed descriptive statistics for study measures (see Table 1). To examine how invalid data may alter the overall description of the sample, we compared the data across the analytic subsamples identified (Table 1). We used t -tests and chi-square statistics, as appropriate. Prior to conducting t -tests, we performed Levene's test of equal variance to account for differences across groups. Furthermore, we computed the log10 transformation for skewed variables in our analyses: time spent online looking for sex partners, number of partners, and URAI. We present the nontransformed means for interpretation, yet analyses were done using the transformed variables.

Finally, we compared the magnitude of the observed relationships across the different analytic subsamples. Comparisons were based on mutually exclusive groups to avoid violations of the independence assumption in statistical analyses. Given that variance is closely linked with the estimation of linear relationships, estimation of these parameters with invalid data may result in incorrect correlations, both in magnitude and in direction, and statistical significance (Schraepfer and Wagner 2005). We used Fisher's r -to- Z conversion formula prior to testing whether the magnitude of the observed correlations was statistically significant between the analytic subsamples with independent sample t -tests. We used a Z critical value of 1.96 (two-tailed) to test whether the associations between Internet and sex behavior differed (e.g., $r(\text{Invalid Cases}) - r(\text{Unique Cases}) = 0$). We assigned a Z -score of 0 to correlations that were not statistically significant (i.e., a nonsignificant association reflects the inability to reject the null hypothesis; consequently $r = 0$). These analyses allowed us to examine how different data quality decisions influenced the composition of the sample. Finally, based on these comparisons, we tested the relationship between Internet use and sexual behavior in a multivariate regression model with our final analytic sample (i.e., only valid cases). We performed Bonferroni post hoc adjustments to avoid inflating the Type I error rate.

Table 1. Descriptive Statistics across Analytic Samples

	Full Sample (N = 488)	Initial Valid Sample (Group A; N = 385)	Pooled Valid Sample (Group E; N = 447)	Invalid Cases (Group D; N = 41)
Age (in years)	21.41 (1.95)	21.56 (1.98)	21.48 (1.97)	20.67 (1.69)
Education	6.04 (1.60)	6.17 (1.58)	6.11 (1.57)	5.33 (1.75)
Race/ethnicity				
White	346 (70.9%)	281 (73.0%)	321 (71.8%)	25 (61.0%)
Black	32 (6.6%)	25 (6.5%)	29 (6.5%)	3 (7.3%)
Hispanic	58 (11.9%)	34 (8.8%)	48 (10.7%)	10 (24.4%)
Asian or Pacific Islander	35 (7.2%)	32 (8.3%)	34 (7.6%)	1 (2.4%)
Other	17 (3.5%)	13 (3.4%)	15 (3.4%)	2 (4.9%)
Occupation				
School	196 (40.2%)	150 (39.0%)	176 (39.4%)	20 (48.8%)
Part-time work	28 (5.7%)	24 (6.2%)	27 (6.0%)	1 (2.4%)
Full-time work	136 (27.9%)	106 (27.5%)	121 (27.1%)	15 (36.6%)
School and work	123 (25.2%)	101 (26.2%)	119 (26.6%)	4 (9.8%)
Unemployed	5 (1.0%)	4 (1.0%)	4 (0.9%)	1 (2.4%)
Time spent online looking for sex ^a	4.54 (7.62)	3.92 (5.71)	4.26 (7.01)	7.76 (12.33)
Number of partners ^a	3.98 (7.18)	3.55 (5.54)	3.68 (5.66)	7.60 (16.62)
URAI occasions ^a	0.98 (3.00)	0.75 (2.16)	0.87 (2.63)	2.26 (5.70)

Note: ^aUntransformed values are presented; however, differences across groups are estimated using the log₁₀ transformation to alleviate skewness.

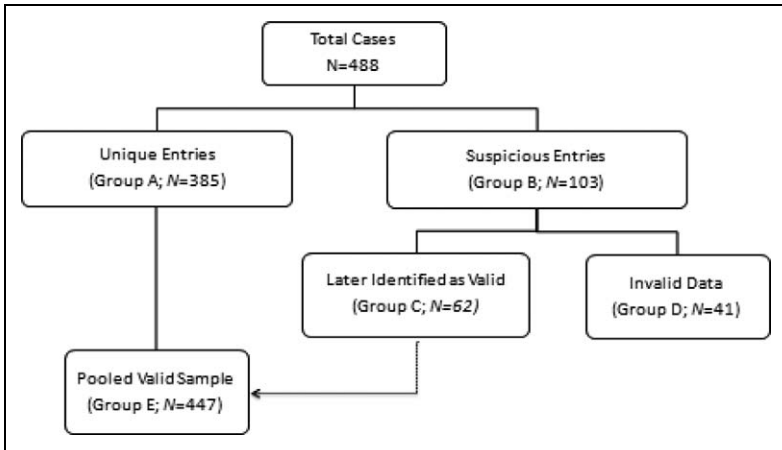


Figure 1. Decomposition of the sample to ensure adequate final analytic sample. Note. Adjusted completion rates are calculated taking full missing data out of the estimation. Completion rate = $[488/548] = 89.05\%$. Adjusted completion rate using Group A = $[385/488] = 78.89\%$. Adjusted completion rate using Group E = $[447/488] = 91.60\%$.

Results

Creating Analytic Subsamples

To examine how invalid data influenced our sample, we divided the 488 entries into two groups (see Figure 1). The first group (i.e., Group A; $n = 385$) included valid cases identified by having a unique IP or e-mail address, or by being the first completed entry from a shared IP address. The second group (i.e., Group B; $n = 103$) included the remaining cases, which were then marked as suspicious entries.

We inspected the 103 suspicious cases by examining whether these entries originated from a shared computer (i.e., entries having the same IP address and operating system and/or browser; $n = 71$) or a different computer (i.e., entries share IP address but use different computers, as identified by a different operating system and/or browser; $n = 32$). We marked a case as valid if the entry had at least two data quality indicators (i.e., a unique e-mail address; indicated living alone in questionnaire, an e-mail address associated to a unique Facebook and MySpace account, and/or completion of the questionnaire in a realistic time frame). We identified 62 cases as valid ($n = 36$ using a shared computer; $n = 26$ using a different computer)

Table 2. Correlations by Total Sample (below diagonal) and Initial Unique Subsample (above diagonal; Group A)

	1	2	3	4
1. Age		.14*	.11*	-0.31
2. Time spent online looking for sex ^a	0.07		.32*	0.1
3. Number of partners ^a	0.04	.36*		.30*
4. URAI occasions ^a	-0.06	.15*	.34*	

Note: ^aAssociations estimated using the log10 transformation to alleviate skewness.
 * $p < .05$.

and included them in a new group (i.e., Group C). Finally, we labeled the remaining 41 cases as invalid data and included in their own group (i.e., Group D) if they did not have at least two quality indicators. We used these groups to compare the sample’s characteristics and the observed associations between Internet use and HIV risk outcomes.

Comparing Mean Differences across Analytic Subsamples

We examined the sample’s composition regardless of invalid data (see Table 1). The full sample had an average age of 21 years ($M = 21.41$, $SD = 1.95$ years) and some college education ($M = 6.04$, $SD = 1.60$). The sample was predominantly white (70.9%), followed by Hispanic (11.9%), Asian Pacific Islander (7.2%), black or African American (6.6%), or reporting another race (3.5%). Over half of the sample reported being in school exclusively (40.2%) or concurrently with work (25.2%), with the remainder of the sample reporting having a part-time (5.7%) or full-time job (27.9%), or being unemployed or on disability (1.0%). Participants reported seeking sexual partners online more than 4 hours per week ($M = 4.54$, $SD = 7.62$), having four male partners in the past 2 months ($M = 3.98$, $SD = 7.18$) and at least one URAI ($M = 0.98$, $SD = 3.00$).

Table 2 reports the correlational analysis examining whether the observed associations between YMSM’s Internet use and sexual behaviors varied if we used cases from the entire sample (lower diagonal of the correlation matrix) or cases initially identified as valid (i.e., Group A; upper diagonal of the correlation matrix), respectively. Comparisons between the diagonals indicate that different associations emerge depending on which group is used in the analyses. In the lower diagonal (entire sample), we noted associations between Internet use to meet casual partners and number of partners ($r = .36$) and URAI($r = .15$), respectively, and number of

partners and URAI ($r = .34$). In the upper diagonal (initially valid sample), we noted similar associations between Internet use and URAI with number of partners; however, we also noted different associations as age was correlated both with Internet use to meet partners ($r = .14$) and number of partners ($r = .11$). Furthermore, we found no association between Internet use to meet casual partners and URAI. These findings suggest that the associations between age and Internet use and sexual behaviors, respectively, would not be significant if we included all cases, whereas they would be significant if we excluded the suspicious data.

We were unable to test whether the differences in sample composition or correlations were significant, given that some cases contributed to both estimates. Consequently, to avoid violations to the independence assumption, we used mutually exclusive groups to examine whether misestimation occurred depending on different data quality criteria (see Table 3).

How Do Data Quality Decisions Influence the Observed Relationships?

Inclusion of invalid data may influence the overall variation in the measures and statistical precision. The presence of group differences would indicate that the variance from invalid data is not randomly distributed and would bias the true relationship between study variables. Conversely, failing to observe differences between the groups would suggest that the influence of invalid data is distributed at random and, as a result, would not dramatically bias the relationships among variables. Consequently, we carried out three comparisons to examine how data quality decisions may influence the relationships among the study variables.

Should we only use cases initially identified as valid? (Contrast 1: Group A Versus B). Our first contrast examined whether exclusion of Group B cases (i.e., suspicious cases) altered the composition of the sample and the underlying relationships among the variables. We found cases in Group B were younger ($M = 20.85$, $SD = 1.75$; $t(176.65) = 3.53$, $p < .001$) and less educated ($M = 5.55$, $SD = 1.57$; $t(482) = 3.55$, $p < .001$) than cases in Group A (see Table 1). We found Whites contributed the greatest number of cases in Group B (all invalidly defined cases). Hispanic/Latino participants were the largest ethnic group (41.4%) represented in Group B ($\chi^2(488, 4) = 18.81$, $p < .001$). Compared to Group A, cases in Group B reported more time seeking partners online ($M = 6.83$, $SD = 12.12$; $t(137.83) = -2.06$, $p < .05$), more partners ($M = 5.56$, $SD = 11.23$; $t(131.59) = -2.19$, $p < .05$), and

Table 3. Correlations and Z-Scores across Contrasted Samples

	Group B (Suspicious)	Z-Score	Group A (Unique Entries)	Z-Score	Test Value
Contrast 1: Should we only use cases initially identified as valid?					
r(Age, Time)	0	0.00	0.14	0.14	-1.19
r(Age, Partners)	0	0.00	0.11	0.11	-0.93
r(Age, URAl)	0	0.00	0	0.00	0.00
r(Time, Partners)	0.45	0.48	0.32	0.33	1.29
r(Time, URAl)	0.25	0.26	0	0.00	2.15*
r(Partners, URAl)	0.4	0.42	0.3	0.31	0.96
Contrast 2: Does the inclusion of entries later identified as valid influence our analyses?	Group A (Unique Entries)	Z-Score	Group C (Later Identified as Valid)	Z-Score	Test Value
r(Age, Time)	0	0.00	0.14	0.14	-0.99
r(Age, Partners)	0	0.00	0.11	0.11	-0.78
r(Age, URAl)	0	0.00	0	0.00	0.00
r(Time, Partners)	0.56	0.63	0.32	0.33	2.12*
r(Time, URAl)	0.29	0.30	0	0.00	2.10*
r(Partners, URAl)	0.35	0.37	0.3	0.31	0.39
Contrast 3: Does the inclusion of invalid data lead to different conclusions in our analyses?	Group D (Invalid Data)	Z-Score	Group E (Pooled Valid Data)	Z-Score	Test Value
r(Age, Time)	0	0.00	0.11	0.11	-0.57
r(Age, Partners)	0	0.00	0	0.00	0.00
r(Age, URAl)	0	0.00	0	0.00	0.00
r(Time, Partners)	0	0.00	0.36	0.38	-1.93
r(Time, URAl)	0	0.00	0.15	0.15	-0.77
r(Partners, URAl)	0.57	0.65	0.32	0.33	1.22

*p < .05 in a two-sided test (i.e., Z-score > 1.965).

more URAI ($M = 1.80$, $SD = 4.92$; $t(127.51) = -2.58$, $p < .01$). We found no differences by occupation.

We then compared the correlation matrix for both groups (see Table 3) and found two instances where statistical significance differed. The correlation between participants' time spent online seeking partners and number of URAI partners was null among Group B participants, but present for Group A ($r = .24$). Conversely, the correlation between participants' time spent online seeking partners and URAI was significant in Group B ($r = .25$), but null in Group A. These findings suggested that difference in the variance between cases initially identified as valid (Group A) and those marked as suspicious (Group B) affected the composition of the sample and the relationship between study variables.

Given that overly conservative exclusion criteria may lead to some true variability being lost, however, we ran two additional contrasts. The first contrast examined whether we would lose important variability if we excluded cases later identified as valid (i.e., Group C) from cases included initially identified as valid (i.e., Group A). The second examined whether invalid data (i.e., Group D) would bias the relationships in the data, even after pooling all valid cases (Groups A and C) together (i.e., Group E).

Does the inclusion of entries later identified as valid influence our analyses? (Contrast 2: Group A Versus C). This contrast examined whether the inclusion of Group C (i.e., valid cases within the suspicious group) increased the variability found in Group A (i.e., initially valid group). Group C cases were younger ($M = 20.97$, $SD = 1.79$; $t(87.06) = -2.38$, $p < .05$) and had lower education attainment ($M = 5.69$, $SD = 1.44$; $t(442) = -2.24$, $p < .05$) than Group A. Compared to other race/ethnicity groups, a greater proportion of white and Hispanic cases were identified in Group C than Group A; $\chi^2(447, 4) = 11.76$, $p < .05$. Furthermore, Group C cases reported more URAI ($M = 1.54$, $SD = 4.45$) than Group A, $t(72.78) = 2.36$, $p < .05$. We found no differences by occupation, time spent online looking for partners, or number of partners in the past 2 months.

To understand whether the different correlations observed in Contrast 1 were attributable to variation from valid cases in Group C (i.e., true variation), we compared the correlation matrices of Groups A and C. Significant differences would suggest that inclusion of valid cases in Group C would contribute to the overall estimation of associations of interest, and therefore should be included in the total valid sample. We found the correlation between participants' time spent online seeking out partners and number of URAI partners was greater for Group C ($r = .56$) than for Group A

($r = .32$). Furthermore, the correlation between participants' time spent online seeking partners and URAI was also found to differ across groups. While the correlation was null for Group A, we found a linear relationship between these variables for Group C ($r = .29$). These findings suggested that cases in Group C contributed unique variation to the data. Consequently, we pooled cases from Groups A and C together (i.e., Group E).

Does the inclusion of invalid data lead to different conclusions in our analyses?(Contrast 3: Group D vs. E). Finally, we examined whether invalid cases (Group D) would change the estimated relationships found in the pooled valid sample (Group E). Compared to Group E, Group D cases were younger ($M = 20.67$, $SD = 1.69$; $t(482) = -2.50$, $p < .05$), had lower education attainment ($M = 5.33$, $SD = 1.75$; $t(482) = -2.99$, $p < .01$), were more likely to report spending more time seeking partners online ($M = 7.76$, $SD = 12.33$; $t(470) = 2.70$, $p < .05$), and to indicate a greater number of partners in the past 2 months ($M = 7.60$, $SD = 16.62$; $t(448) = 2.59$, $p < .01$). We found no differences by race/ethnicity, occupation, or URAI.

Significant differences within this comparison would suggest that we would be adding nonrandom invalid variability to our final analytic sample if we included these invalid data. We found no evidence to suggest that there were significant differences across the correlations between invalid and the valid data (see Table 3); however, to maximize statistical precision, we used Group E (i.e., all valid cases identified in the sample; $N = 447$) as our study's final analytic sample. The final completion rate using Group E was 91.60%. Final correlations using Group E are included in Table 3.

What Is the Relationship between Internet Use and HIV Risk Behaviors in Our Sample?

Once the final analytic sample was identified, we ran a multivariate regression model to test the association between Internet use and number of partners ($R^2 = 13.8\%$; $F(7, 408) = 9.19$, $p < .001$) and URAI ($R^2 = 4.8\%$; $F(7, 408) = 2.88$, $p < .01$), controlling for age, race/ethnicity, and education.

Participants' number of partners in the past 2 months was associated with the log-transformed number of hours spent online looking for sexual partners ($b = 0.28$, $SE = .04$; $\beta = 0.37$, $p < .001$). We found no association between number of partners and age, race/ethnicity, or education.

Participants' URAI was related to the log-transformed number of hours spent online looking for sexual partners to hook-up ($b = 0.14$, $SE = .04$; $\beta = 0.16$, $p < .01$), and negatively associated with education ($b = -0.03$,

$SE = .01$; $\beta = -0.12, p < .05$). We also found African American participants reported less URAI than Whites ($b = -0.13, SE = .06$; $\beta = -0.10, p < .05$). We found no additional relationships between URAI and age, or other race/ethnicity groups.

Discussion

The Internet's accessibility, affordability, anonymity, and acceptability has broadened the public discourse of sexuality, increased our ability to study HIV/AIDS risk behaviors, and created a new venue to develop Internet-based prevention (Pequegnat et al. 2007; Ybarra and Bull 2007). The very anonymity that makes this an attractive modality, however, makes it difficult to verify the identity of research participants and ensure that only valid entries are included in the analyses (Bowen et al. 2008; Konstan et al. 2005). In this study, we examined how different definitions of invalid data may influence the observed relationships in a confidential web survey. Although these issues are not *specific* to web surveys, as they may also occur in other data collection modalities such as telephone interviews, they may have a greater propensity to occur in a web-based format. The web also provides unique opportunities to check for data quality. Our findings support past research suggesting that YMSM who spend a greater number of hours online looking for casual partners may increase their HIV risks (Liau et al. 2006). As expected, however, we found that decisions to handle suspicious and invalid data led to different associations among study variables.

Data quality is vital to ensure the correct estimation of a sample's composition across the study measures. While the calculated completion rate did not vary when compared across the three major analytic groups (i.e., Groups A, C, and E), the overall composition of the sample was affected depending on data quality decisions. Decisions regarding how to handle suspicious entries, in particular, seemed to have the largest influence on the overall sample composition. One approach that some researchers may consider is the disposal of all suspicious cases (e.g., entries with similar IP and/or e-mail addresses, erratic survey patterns) as the quickest approach to avoid bias (i.e., exclude cases in Group B). We caution researchers who wish to employ this approach for data management; we found that 60% of them to be valid cases (i.e., Group C) and increased the sample's diversity once pooled with the other valid entries (i.e., Group E). Group C cases may occur if YMSM access the web survey using a shared IP connection, as commonly found in libraries, coffee shops, and shared living spaces.

Consequently, the removal of all suspicious entries is an inadequate data quality strategy as it may exclude valid data, decrease statistical power, and bias conclusions drawn from the study. Future research that includes a question assessing whether participants are completing their survey on a shared computer may facilitate the identification of valid cases. Furthermore, although we used multiple data checks beyond IP address, we assumed that the first case from a shared IP address was a true case. Future research examining whether this assumption holds across studies may provide further insights regarding data quality decisions.

Data quality is also vital to ensure that we estimate correct associations, draw accurate conclusions, and make informed recommendations from our findings. Some researchers may argue that removing all suspicious cases provides conservative estimates of the relationships between variables in a study. In other words, if one is able to detect an association between a set of variables with a sample that excludes valid cases marked as suspicious (i.e., Group A), then the identified relationship may be greater, had we included the valid cases included in the suspicious group (i.e., Group C). While a seemingly logical argument, it assumes that the variance in Group A is equivalent to that provided by Group C. From our analyses, however, we found this argument to be incorrect. When we compared the relationships between Groups A and C, we found that Group C contributed unique variation to study measures and influenced the observed relationships between Internet use and HIV risk outcomes. If we excluded these cases, for example, we would not have detected a positive association between time spent online seeking casual partners and URAI. Thus, taking an overly conservative approach to handle suspicious data may change the relationship among variables. These findings underscore the importance of testing whether different analytic groups have comparable variance prior to their exclusion.

Undoubtedly, invalid cases (i.e., false entries, duplicates; Group D) should be removed from the analytic sample to maximize the veracity and precision of estimated relationships. Invalid cases altered the overall composition of the sample. When we examined whether the inclusion of these invalid cases (i.e., Group D) would bias the relationships between study variables, we found that their variation was not different from valid cases and could be considered random error in our sample. Even though these data did not bias the direction or statistical significance of the correlations in our sample, their inclusion did change the precision of the estimated relationships. Nevertheless, care should be taken in future studies, as invalid

data may not always vary at random and may pose additional threats to the internal and external validity of the study.

Participants may have different reasons for duplicating or falsifying data. The presence of study incentives may be one explanation. Some participants might seek to increase their chances of obtaining incentives if a study raffles rewards, whereas others may pursue additional compensation if the payment system is automated. Future research examining how to provide participant incentives that do not encourage multiple submissions is required. Alternatively, invalid data may be attributed to participants' curiosity about the study; some participants may decide to participate in a survey more than once to see if different questions are asked depending on their answers (i.e., skip patterns). Finally, some participants may just be malicious and desire to corrupt the data. Given that some of these issues may arise due to lack of education regarding the research process, we sent a follow-up e-mail to cases marked as suspect to debrief them on the potentially harmful effects of duplicate or falsified data. Two participants who had several duplicate entries apologized for their behavior.

Future research examining whether educating prospective participants before they complete the questionnaire about the research process (i.e., above and beyond the study details provided in a consent form) and appealing to their collectivistic and altruistic attitudes may diminish the presence of invalid data in web surveys may be useful. Similarly, research examining whether falsification decreases if respondents are notified that the web survey is detecting irregular data patterns may be warranted, particularly if these approaches do not evoke social desirability among respondents.

Strategies to decrease invalid data have been developed to ensure that participants fill out the survey only once (Bowen et al. 2008; Kraut et al. 2004). Unfortunately, these strategies may not be compatible with some web surveys (e.g., anonymous) and affect participation rates. Some participants, for example, may not participate in a study if strategies to handle invalid data require that their anonymity be foregone (e.g., providing personal information), particularly if the survey's content focuses on sensitive information (e.g., sexual history, substance use). Consequently, researchers need to carefully consider whether foregoing anonymity is a suitable approach to decrease invalid data depending on the study's topic, population, and available staff resources. Other strategies may require that a population list of all eligible participants be available for verification purposes. In our study, for example, the use of Facebook and MySpace to cross-check suspicious cases was most useful to identify valid cases. Some participants, however, may not be listed in these sites, have a different e-mail address

associated with their account, have multiple profiles, or have privacy restrictions. Taken together, these issues underscore the importance of considering how to prevent and handle invalid data when designing and executing a study. Future research examining the acceptability and efficacy of these strategies, as well as contrasting the net gains of each data quality approach, is warranted.

Study Strengths and Limitations

Our study has several strengths and limitations deserving mention. First, we focused our attention on invalid data due to duplicate entries, but these are only one type of problem (Wang and Strong 1996). Other problems include quick click-throughs (i.e., satisficing or suboptimal responding of a survey to expedite its completion; Krosnick 1991) and misrepresentation (i.e., qualifying for a questionnaire by lying about age or other characteristics), which also deserve attention in future studies.

All *apparent* instances of satisficing corresponded to cases marked as invalid cases in our data, but it is possible that less apparent satisficing patterns were present in valid cases without being identified. Furthermore, we sought to decrease misrepresentation by focusing study advertisements in social network sites for our target population (i.e., single, male, ages 18 and 24, interested in men). Nevertheless, it is possible that individuals referred by other respondents could have misrepresented their information (e.g., age, relationship status) in order to be eligible; however, we do not know if these participants also lied about their survey answers. In other words, some participants may have falsified their eligibility information (e.g., a 28-year-old posing online as a 23-year-old), yet answered the survey honestly.

Finally, participants were recruited as a convenience sample, as there are no population frames from which to select a randomly representative sample of YMSM. Nonetheless, the fact that our findings replicate past research examining the relationship between Internet use and HIV risk, had a high completion rate, and recruited a diverse sample of YMSM through social networks sites all lend support to the validity of our findings.

Conclusions and Implications

Web-survey data collection is an efficient and timely modality to carryout HIV/AIDS research with hard-to-reach populations. Our findings suggest that *pre and post hoc* decisions regarding how to handle suspicious data are warranted. A data quality approach that is too conservative (e.g.,

indiscriminate removal of cases from same IP), however, may harm the validity of a study. Similar to attrition analyses, we encourage researchers using web surveys to mention whether data exclusion criteria were set prior to analyses, explicitly stating whether the presence of invalid data was examined and how it was handled. This information may be a useful quality criterion for web-survey research in HIV/AIDS and other topics. Furthermore, this criterion may be an important covariate in meta-analyses seeking to pool studies and find overall relationships among study variables. Researchers may want to carefully consider how suspect data will be handled and budget for quality checks throughout the data collection process.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by an award from the University of Michigan's Office of the Vice-President for Research (UM-OVPR 5996) and an NIH Career Development Award (K01-MH087242) to Dr. Bauermeister.

References

- Bauermeister, J., M. Leslie-Santana, M. Johns, E. Pingel, and A. Eisenberg. 2011. Mr. Right and Mr. Right now: Romantic and casual partner-seeking online among young men who have sex with men. *AIDS and Behavior* 15:261–72.
- Bowen, A., C. Daniel, M. Williams, and G. Baird. 2008. Identifying multiple submissions in Internet research: Preserving data integrity. *AIDS & Behavior* 12:964–73.
- Couper, M. 2008. *Designing effective web surveys*. Cambridge, MA: Cambridge University Press.
- Garofalo, R., A. Herrick, B. Mustanski, and G. Donenberg. 2007. Tip of the iceberg: Young men who have sex with men, the Internet, and HIV risk. *American Journal of Public Health* 97:1113–17.
- Konstan, J., B. Simon-Rosser, M. Ross, J. Stanton, and W. Edwards. 2005. The story of subject naught: A cautionary but optimistic tale of Internet survey research. *Journal of Computer-Mediated Communication* 10: article 11.

- Kraut, R., J. Olson, M. Banaji, A. Bruckman, J. Cohen, and M. Couper. 2004. Psychological research online: Report of board of scientific affairs' advisory group on the conduct of research on the internet. *American Psychologist* 59:105–17.
- Krosnick, J. 1991. Response strategies for coping with the coping demands of attitude measures in surveys. *Applied Cognitive Psychology* 5:213–36.
- Liau, A., G. Millet, and G. Marks. 2006. Meta-analytic examination of online sex-seeking and sexual risk behavior among men who have sex with men. *Sexually Transmitted Diseases* 33:576–84.
- Pequegnat, W., B. Simon-Rosser, A. Bowen, S. Bull, R. DiClemente, W. Bockting, and R. Zimmerman. 2007. Conducting Internet-based HIV/STD prevention survey research: Considerations in design and evaluation. *AIDS & Behavior* 11:505–21.
- Ross, M. 2005. Typing, doing, and being: Sexuality and the Internet. *Journal of Sex Research* 42:342–52.
- Ross, M., B. Simon-Rosser, and J. Stanton. 2004. Beliefs about cybersex and Internet-mediated sex of Latino men who have Internet sex with men: Relationships with sexual practices in cybersex and in real life. *AIDS Care* 16:1002–11.
- Schraepfer, J., and G. Wagner. 2005. Characteristics and impact of interviews in surveys—An analysis of genuine fakes in the raw data of SOEP. *Allgemeines Statistisches Archiv* 89:7–20.
- Wang, R., and D. Strong. 1996. Beyond accuracy: What data quality means to consumers. *Journal of Management Information Systems* 12:5–34.
- Ybarra, M., and S. Bull. 2007. Current trends in Internet and cell phone-based HIV prevention and intervention programs. *Current HIV/AIDS Reports* 4:201–207.