Published in final edited form as:

J Sex Marital Ther. 2002 ; 28(4): 331–338.

Reliability of Retrospective Self-Reports of Sexual and Non-Sexual Health Behaviors Among Women

Lauren E. Durant, M.S. and Michael P. Carey, Ph.D.

Department of Psychology and Center for Health and Behavior, Syracuse University

Abstract

The accuracy of self-reports regarding sexual health behavior has been questioned. To investigate whether assessment of sexual health behaviors are uniquely difficult to report, 185 college women were asked to answer behavioral frequency questions about sexual and non-sexual health behaviors for an 8-week interval. Women took part in either a face-to-face interview or completed a self-administered questionnaire. One week later the women returned and responded to the same questions in the same mode of assessment conditions. The test-retest intraclass correlations showed that all health behaviors, sexual and non-sexual were reported reliably. There was a trend for lower frequency reports to yield more stable estimates of behavioral frequency. These findings converge with other methodological investigations to indicate that socially sensitive health behaviors are not more difficult to assess reliably.

Keywords

Assessment; Self-Report; Sexual Behavior; Interview; Self-Administered questionnaire

Obtaining reliable and accurate self-reports of sexual behavior is essential for both public health and clinical research on reproductive and sexual health. Despite the concern that self-reports may not accurately reflect actual behavior, investigators rarely assess the accuracy (validity) of their data¹. This is often due to the lack of objective measures available and the absence of a “gold standard” of frequency of behavior with which to compare self-reports.

An alternative to assessing the veracity of self-reports is to present evidence of the stability or consistency of the data. Generally, studies of sexual behavior have used test-retest of their measures to assess reliability of self-reports². These studies have investigated a range of sexual health behaviors across diverse populations. To date, investigations have assessed stability of sexual behavior reports with gay men³, heterosexual men and women^{4,5}, drug users^{6,7} and African-Americans⁸. These studies have found test-retest correlations that range from .3 to .9 across the populations studied¹.

With such a range of test-retest correlation values across samples and different types of sexual activity the question arises: Is the assessment of sexual behavior uniquely difficult to report, or are stable estimates of the frequency of most health-related behavior difficult to elicit? To our knowledge, no study has examined whether type of health behavior reported (sexual versus non-sexual) affects the stability of the self-reports. Therefore, our investigation examined whether type of behavior reported (i.e., sexual versus non-sexual) influenced the stability of self-reported estimates. We predicted that both types of behavior would be reported reliably, given confidentiality assurances. Also, we expected that there would be a trend for stronger

reliability coefficients for low frequency behavior reports (regardless of whether these behaviors were sexual or non-sexual).

Method

Participants

One-hundred and ninety women were recruited from psychology courses at Syracuse University. Participants tended to be young ($M=19$ yrs, $SD=1.4$), Caucasian (75%), and in their first or second year of college (87%). Informed consent was obtained from all women, who received course credit for participating.

Procedure

During a regularly scheduled lecture time, women from introductory level psychology courses ($N = 190$) received a brief presentation regarding the study's purpose and rationale during a regularly scheduled lecture time. The study was introduced as a study of "College Women's Health." Participants were not told the entire purpose of the study in an effort to minimize potential biases that might influence how participants responded in the experimental condition. Interested women signed up for an initial overview session, during which they received (a) an explanation of the procedures; (b) provided written consent; and (c) completed the demographic questionnaire (i.e., a 9-item questionnaire that requested information regarding the participant's age, ethnicity, living, and relationship status) and some exploratory measures. After completing the questionnaires, participants were randomly assigned to receive either a self-administered questionnaire (SAQ) or a face-to-face interview (FTFI) and asked to generate code names to insure the anonymity of their data.

At the second meeting, women either (a) completed a 13-item SAQ that requested information regarding sexual behaviors (e.g., masturbation, oral, vaginal, and anal sex) and non-sexual behaviors (e.g., smoking, caffeine); or (b) participated in a FTFI that was conducted by a trained assistant who was blind to the study's hypotheses. Both assessment modes (i.e., SAQ and FTFI) contained the same questions.

Participants returned in one week for a third meeting and an identical assessment. At the end of their session, participants were debriefed and thanked for their participation.

Results

Tables 1 and 2 provide the means (M), standard deviations (SD), and ranges for the non-sexual and sexual health behaviors, respectively. Because distributions of sexual behaviors were positively skewed (i.e., for each behavior, there was an overrepresentation of participant's reporting 0's and 1's) violating the assumption of normality upon which tests of significance of correlation coefficients are based, these data were transformed for correlations and linear regression using bootstrapping and a Log transformation.

We computed the test-retest reliability of participants' responses on each behavior over the 8-week period using the intraclass correlation (r). When computing the intraclass correlation, we included all participants in the sample to account for participants who reported no behaviors at time 1 but who reported behaviors at time 2 or visa versa. Table 3 reports test-retest correlation coefficients and reliability comparisons for non-sexual behaviors by mode of assessment. Table 4 reports the test-retest reliability intraclass correlation coefficients for sexual behaviors by mode of assessment.

Analysis of reliability by assessment mode for each behavior yielded moderately strong reliability coefficients for reports of non-sexual behaviors (range $r = .47$ to $.98$, $M = .77$; see

Table 3) and strong reliability coefficients for sexual behaviors (range $r = .84$ to $.98$, $M = .92$; see Table 4).

To determine whether the modes differed by behavior, we calculated confidence intervals. These analyses revealed no differences between modes by type of behavior (sexual or non-sexual), that is, all confidence intervals for the mode comparisons contained zero.

To examine whether stronger reliability coefficients were related to low frequency behavior reports, we used linear regression analyses to test each behavior for O slope (i.e., the average response of time 1 and time 2 reports were regressed on the log transform of the difference between time 1 and time 2 reports). All sexual behaviors exhibited the predicted trend of lower frequency reports being more stable than higher frequency reports. For non-sexual behaviors, all behaviors except reports of cigarettes smoked per day and alcoholic beverages consumed (in the FTFI condition) yielded more stable estimates at lower frequencies of behavior.

Discussion

There were 2 primary findings that emerged from this investigation: (a) reports of sexual and non-sexual behaviors can be obtained reliably from both FTFIs and SAQs, and (b) behavioral frequency reports tend to be less stable as the number of events to report increase, regardless of type of behavior reported. These findings converge with findings from Kalichman et al.⁷, who found the SAQ and FTFI were the most reliable methods of obtaining self-reports of sexual behavior, and with findings from Downey et al.'s⁹ investigation where accuracy declined as frequency of behavior increased.

Although there were no significant differences in reliability found across behaviors or modes of assessment, there was a non-significant decline in stability between reports of sexual and non-sexual behaviors. Part of this decline may be attributed to the differences in wording of the behavioral frequency questions. The sexual behaviors were worded "how many times" whereas the non-sexual behaviors were worded "On average, how many..." or "how many beverages have you consumed on a typical day?" The latter questions may have led respondents to provide a less exact response. This effect has been found when respondents are presented with pre-coded response formats compared to open response formats¹⁰. An alternative explanation may be related to the frequency at which most non-sexual behaviors are engaged in. For example, participants who smoke need to remember many more smoking occasions than sexual events.

The findings of this study need to be interpreted in light of its limitations. First, reliability was measured over a relatively brief interval of time with questions that were not worded identically. Second, the generalizability of these findings may be limited by use of a college-aged sample. Finally, the sample size (for each behavior engaged in) was modest and requires that these results be replicated.

In summary, our results indicate that self-reported sexual and non-sexual behaviors can be assessed with moderately strong test-retest stability. As there were no differences by mode of assessment, investigators can choose either mode of administration (i.e., face-to-face interviews or self-administered questionnaires) without declines in stability. Future investigations of the validity, as well as the stability, of self-reported sexual health behaviors is encouraged.

Acknowledgements

This research was supported by grants from the National Institute of Mental Health to Michael P. Carey (K02-MH01582 and RO1-MH54929). The authors thank John R. Gleason for statistical consultation; Jennifer Alvarez, Jennifer Chernowski, Nikia Hearst, Lori Mothersell, and LaToya Shakes who helped to collect and enter data.

References

1. Catania J, Binson D, van der Straten A, Stone V. Methodological research on sexual behavior in the AIDS era. *Ann Rev Sex Res* 1995;6:77.
2. Weinhardt LS, Forsyth AD, Carey MP, Jaworski BA, Durant LE. Reliability and validity of self-report measures of HIV-related sexual behavior: Progress since 1990 and recommendations for research and practice. *Arch Sex Behav* 1998;27:155. [PubMed: 9562899]
3. McLaws M, Oldenburg B, Ross MW, Cooper DA. Sexual behaviour in AIDS-related research: Reliability validity of recall and diary measures. *J Sex Res* 1990;27:265.
4. Weinhardt LS, Carey MP, Maisto SA, Carey KB, Cohen MM, Wickramasinghe SM. Reliability of the timeline followback sexual behavior interview. *Ann Behav Med* 1998;20:25. [PubMed: 9755348]
5. Taylor JF, Rosen RC, Lieblum SL. Self-report assessment of female sexual function: Psychometric evaluation of the brief index of sexual functioning for women. *Arch Sex Behav* 1994;23:627. [PubMed: 7872859]
6. Darke S, Hall W, Heather W, Ward J, Wodak A. The reliability and validity of a scale to measure HIV risk-taking behavior among intravenous drug users. *AIDS* 1991;5:181. [PubMed: 2031690]
7. Needle R, Fisher DG, Weatherbee N, et al. Reliability of self-reported HIV risk behaviors of drug users. *Psychol Addictive Behav* 1995;9:242.
8. Kalichman SC, Kelly JA, Stevenson LY. Priming effects of HIV risk assessments on related perceptions and behavior: An experimental field study. *AIDS Behav* 1997;1:3.
9. Downey L, Ryan R, Roffman R, Kulich M. How could I forget? Inaccurate memories of sexually intimate moments. *J Sex Res* 1995;32:177.
10. Schwarz, N.; Hippler, HJ. What response scales may tell your respondents. In: Hippler, HJ.; Schwarz, N.; Sudman, S., editors. *Social information processing and survey methodology*. New York: Springer-Verlag; 1987. p. 163

Table 1

Summary Statistics for Non-Threatening Behaviors, by Mode

Behavior (in the past two months)	SAQ (n = 95)			FTFI (n = 88)		
	Mean	SD	Range	Mean	SD	Range
On average, how many cigarettes per day have you smoked?	1.7	4.1	(0 – 30)	2.1	4.8	(0 – 20)
On average, how many hours per week have you exercised?	4.3	6.4	(0 – 40)	2.9	3.5	(0 – 18)
How many caffeinated sodas and cups of coffee have you had on an average day?	1.6	2.3	(0 – 20)	1.9	1.4	(0 – 10)
How many alcoholic beverages have you consumed on a typical day that you drink?	2.8	2.5	(0 – 10))	3.8	2.6	(0 – 10)
How many times have you participated in non-intercourse activities (e.g., petting w/clothes on or off)?	15.1	32.0	(0 – 240)	12.3	16.4	(0 – 80)

Note. SAQ = data obtained with a self-administered questionnaire; FTFI= data obtained in a face-to-face interview.

Table 2

Summary Statistics for Threatening Behaviors, by Mode

Behavior (in the past two months)	SAQ (n = 95)			FTFI (n = 88)		
	Mean	SD	Range	Mean	SD	Range
How many times have you masturbated?	1.9	5.2	(0 – 40)	1.6	7.9	(0 – 70)
How many partners have you had sex with (i.e., oral, anal, or vaginal sex)?	.70	.65	(0 – 4)	.82	.82	(0 – 3)
How many times have you had vaginal sex with a condom?	2.9	6.5	(0 – 40)	3.8	7.9	(0 – 60)
How many times have you had vaginal sex without a condom?	2.3	5.5	(0 – 25)	4.1	10.1	(0 – 60)
How many times have you had oral sex without a barrier (i.e., condom or dental dam)?	2.9	5.9	(0 – 40)	3.6	6.2	(0 – 35)

Note. SAQ = data obtained with a self-administered questionnaire; FTFI= data obtained in a face-to-face interview.

Table 3
Test-Retest Correlation Coefficients and Reliability Comparison for Non-Threatening behaviors, by Mode

Behavior (in the past two months)	(n = 95)			(n = 88)			(N = 183)	
	SAQ	M Fisher Z	Confidence Interval 95%	FTFI	M Fisher Z	Confidence Interval 95%	Mode Comparison	95% CI
On average, how many cigarettes per day have you smoked?	.98	2.2	(1.6 to 2.9)	.96	2.0	(.94 to 3.1)		(-1.5 to 1.0)
On average, how many hours per week have you exercised?	.74	.95	(.01 to 1.9)	.47	.51	(-.58 to 1.6)		(-1.9 to .99)
How many caffeinated sodas and cups of coffee have you had on an average day?	.48	.53	(-.40 to 1.5)	.63	.75	(.18 to 1.3)		(-.84 to 1.3)
How many alcoholic beverages have you consumed on a typical day that you drink?	.97	2.2	(1.9 to 2.4)	.96	1.9	(1.6 to 2.2)		(-.63 to .15)
How many times have you participated in non-intercourse activities (e.g., petting w/clothes on or off)?	.96	1.9	(1.3 to 2.5)	.89	1.4	(.92 to 1.9)		(-1.3 to .31)

Note. Reliability coefficients are Intraclass correlations; SAQ = data obtained with a self-administered questionnaire; FTFI = data obtained in a face-to-face interview; CI = confidence interval; M Fisher Z = mean Fisher Z transformations; all confidence intervals are Z confidence intervals.

Table 4
Test-Retest Correlation Coefficients and Reliability Comparison for Threatening Behaviors, by Mode

Behavior (in the past two months)	(n = 95)			(n = 88)			(N= 183)	
	SAQ	M Fisher Z	Confidence Interval 95%	FTFI	M Fisher Z	Confidence Interval 95%	Mode Comparison	95% CI
How many times have you masturbated?	.98	2.2	(1.5 to 2.9)	.87	1.4	(.25 to 2.4)	(-2.2 to .46)	
How many partners have you had sex with (i.e., oral, anal, or vaginal sex)?	.91	1.5	(1.2 to 1.9)	.92	1.6	(1.3 to 2.0)	(-.44 to .58)	
How many times have you had vaginal sex with a condom?	.96	1.9	1.3 to 2.6)	.94	1.8	(.75 to 2.8)	(-1.4 to .98)	
How many times have you had vaginal sex without a condom?	.84	1.2	(.70 to 1.8)	.94	1.7	(.93 to 2.5)	(-.47 to 1.5)	
How many times have you had oral sex without a barrier (i.e., condom or dental dam)?	.85	1.3	(.53 to 2.0)	.95	1.9	(1.62 to 2.2)	(-.16 to 1.4)	

Note. Reliability coefficients are Intraclass correlations; SAQ = data obtained with a self-administered questionnaire; FTFI = data obtained in a face-to-face interview; CI = confidence interval; M Fisher Z = mean Fisher z transformations; all confidence intervals are z confidence intervals.