Reliability of Self-Reported Sexual Behavior Risk Factors for HIV Infection in Homosexual Men

SCOTT P. SALTZMAN, MS
ANNE M. STODDARD, ScD
JANE McCUSKER, MD, DrPH
MARTHA W. MOON, RNP
KENNETH H. MAYER, MD

Mr. Saltzman is Project Manager, Ms. Moon is Clinical Research Coordinator, and Dr. Mayer is Research Director at the Fenway Community Health Center, Boston, MA. Dr. Stoddard is Assistant Professor of Public Health and Dr. McCusker is Associate Professor of Public Health at the University of Massachusetts at Amherst. Dr. Mayer also is Assistant Professor of Medicine, Brown University, and Chief, Infectious Disease Division, Memorial Hospital, Pawtucket, RI.

This research was supported by a grant from the Division of Public Health, Commonwealth of Massachusetts.

Tearsheet requests to Dr. Kenneth H. Mayer, The Memorial Hospital, Infectious Disease Division, Prospect St., Pawtucket, RI 02860.

Synopsis

This study was undertaken to determine the reliability of self-reported sexual behavior using the test and retest technique when used with self-reported sexual behavior. The subjects were 116 asymptomatic homosexual men who participated in another study (an examination of behavioral and demographic determinants of HIV antibody status). The subjects were asked to complete two questionnaires. The first contained demographic and sexual behavior questions. The second, administered an average of 6 weeks later, used a subset of the questions in the first questionnaire.

The reliability of the test-retest procedure was measured by the Kappa statistic, which assesses the proportion of agreement between two data items, accounting for the amount of agreement expected by chance. The highest degree of reliability as measured by Kappa was found with demographic information, smoking history, and sexual orientation. Self-reported sexual behaviors for the previous 6 months generally had the next highest degree of reliability as measured by Kappa. Questions examining change over the previous 5 years had the lowest reliability.

Behavior changes during the time between questionnaires, subjectivity of the answer categories, and social desirability of the answers are three factors that may result in a lack of reliability in this self-reported sexual behavior questionnaire. This raises methodological concerns about the measurement of behavioral risk factors for AIDS and the ability to assess meaningfully subjective reports of behavioral change.

HOMOSEXUALLY ACTIVE MEN are the group at highest risk for developing AIDS, presently accounting for more than 70 percent of the cases in the U.S. (*1*). Specific sexual practices, such as receptive anal intercourse with multiple partners, have been identified as risk factors for acquisition of the AIDS-associated virus, human immunodeficiency virus (HIV, formerly known as HTLV-III/LAV) (*2,3*). Other behavior may facilitate transmission of this retrovirus.

Several behavioral research projects are investigating the association of certain sexual behaviors with HIV antibody status. In these studies the adequacy of the behavioral documentation has not been sufficiently explored. This may have implications for models that are developed for transmission and acquisition of HIV, and for clinical outcomes (*4*). The problem of obtaining adequate behavioral data is not unique to research on AIDS, and is an issue for research on other sexually transmitted diseases, as well as contraceptive research.

Estimation of the validity of self-reported sexual behavior is difficult owing to a lack of external sources to corroborate reported activities and the need to maintain confidentiality. Reliability often is an indicator of validity in such circumstances (*5,6*). While high reliability does not assure validity, low reliability indicates low validity. Test-retest reliability (TRTR) is the "degree of stability exhibited when a measurement is repeated under identical conditions" (*7*). TRTR measures the ability of a person to recall a past occurrence consistently.

We investigated the TRTR of sexual behavior data obtained from a seroepidemiologic study of the HIV antibody status of patients at a Boston community health center. We examined the TRTR of questions on demographics, health behaviors, sexual behaviors, differences in TRTR, and factors that may have caused these differences.

Methods and Materials

At least half the patients visiting the Fenway Community Health Center (FCFC) are homosexual or bisexual men living in or near Boston. They come for health maintenance services, sexually transmitted disease screening, and treatment of minor medical problems. Between July 1 and December 31, 1984, 2,000 male patients were given a letter which described the study and invited them to volunteer to participate. To be eligible, they could not have a diagnosis or clinical signs of AIDS or AIDS-related complex and must have been homosexually active during the previous 6-month period.

The study was explained to 201 participants at the health center, and their written, informed consent was obtained. They completed a detailed questionnaire asking for sociodemographic information, their history of health and sexual behavior and use of drugs, and their attitudes and knowledge concerning these practices. The study coordinator reviewed the questionnaire with the participant for missing or inconsistent information. Attitudinal items were checked only for completion. Blood was drawn for a complete count and HIV antibody determination during the clinical examination.

Those who agreed to return for the results of the HIV antibody screening entered the TRTR study. They were asked to complete the second questionnaire, which they were told would take 5 to 10 minutes, prior to receiving the HIV test result. They were not told that the questionnaire was a retest for the purpose of evaluating the reliability of the information they provided on the first questionnaire.

The retest questionnaire contained some of the items from the initial questionnaire. The questions were presented in the same format and order in both questionnaires, although the retest was 12 pages shorter. The two questionnaires were administered under similar conditions in the same room, by the same personnel, and with the same instructions.

The items selected for retest were those on sociodemographic characteristics, general measures of sexuality, specific sexual practices, and other health behaviors, such as smoking behavior and eating a balanced diet. The general sexuality items included sexual orientation, numbers of partners over the subjects' lifetimes and for the past 6 months, and frequency of sexual contact over the last 6 months. Subjects were asked whether they had experienced a period of sexual activity greater than the current period (the "high period") and if so, when the high period occurred.

The retest questionnaire included items on the following specific practices during the previous 6 months, in terms of frequency of practice and numbers of partners: insertive and receptive orogenital activity, insertive and receptive anogenital activity, and deep kissing. To distinguish between unreliability and actual behavior change, participants were asked to note if there had been any changes in their behaviors between taking the two questionnaires, and if these changes were reflected in their answers to the retest. The frequencies of the specific practices were assessed for the high period, if such a period had been reported. Questions were asked related to perceived change in activity during the last 5 years.

The responses to all items on the questionnaire, except for age, were precoded into categories. Items on frequencies of specific sexual practices were precoded into five categories from "never" to "four or more times per week." Items on numbers of partners for the practices were coded in seven categories from "none" to "more than 50." Items on changes in practices during the past 5 years contained five categories from "marked increase" to "marked decrease." Cross tabulations and chi-square statistics were used to compare characteristics of the 116 who completed a retest questionnaire and the 85 who did not.

Agreement in response to a question answered at two different times by the same person was measured by percent agreement and the Kappa statistic (*8*), which gives a measure of agreement relative to that expected by chance, when the frequencies of responses are given. Kappa less than zero indicates less agreement than that which would be expected by chance; Kappa equal to zero indicates the same agreement as would be expected; and positive Kappa indicates greater agreement than that which would be expected by chance. A Kappa equal to one occurs with perfect agreement. Percent agreement and Kappa were computed for each item. Kappa greater than .80 indicates "almost perfect" agreement; Kappa be-

Table 1. Agreement between successive questionnaires, percent agreement and expected agreement, with Kappa statistic and standard error of Kappa (SE)

Item	Percent agreement observed	Percent agreement expected	Kappa	SE
Nonsexual				
Education	91	26	.87	.04
Income	87	26	.86	.04
Current smoking behavior	94	62	.84	.05
Diet habits	79	44	.63	.06
General sexuality				
Sexual orientation	94	67	.84	.06
Ever a high period	91	67	.72	.08
Partners:				
For a lifetime	75	16	.70	.05
For last 6 months	65	19	.56	.05
For high period*	59	16	.52	.06
6-month partners:				
Steady	64	28	.49	.06
Nonsteady	53	16	.44	.05
6-month frequency of contact	70	31	.56	.06
Frequencies of practice in last 6 months:				
Insertive oral	68	34	.52	.06
Receptive oral	60	28	.45	.06
Insertive anal:				
Without condom	66	29	.53	.06
With condom	79	63	.50	.08
Receptive anal	69	29	.56	.06
Kissing	60	25	.47	.06
Numbers of partners for practice in last 6 months:				
Insertive oral	71	29	.60	.06
Receptive oral	66	29	.52	.06
Insertive anal:				
Without condom	61	29	.46	.06
With condom	81	65	.60	.08
Receptive anal	63	29	.50	.06
Kissing	63	30	.53	.06

* Note: Based on 86 subjects who reported a high period at both interviews.

tween .61 and .80 indicates "substantial" agreement; and Kappa between .41 and .60 indicates "moderate" agreement. Kappa values below .40 are associated with "poor" to "fair" agreement (*9*).

After the sample was stratified by characteristics that might affect the reliability of the responses, the agreement achieved within the strata was compared. Four characteristics were hypothesized that might have affected the reliability of the responses: reported behavior change between answering the two questionnaires, the interval between the test and retest questionnaire administrations, educational level of the person, and level of sexual activity. The sample was divided into two strata on each of these dimensions, and agreement in each pair of groups was compared. In order to keep the number of significance tests to a mininum, four questionnaire items were preselected for testing: two general sexuality variables (numbers of partners in the last 6 months, and frequencies of sexual contact in the last 6 months), one very common sexual practice in the last 6 months (numbers of orogenital partners when the subject was the inserter), and one less common sexual practice in the last 6 months (numbers of anogenital partners when the subject was the receptor). Differences in Kappa statistics were compared using the standard normal distribution (*8*).

Results

Of the 201 participants in the study, 140 returned for their test results. Of these, 116 (83 percent) completed the retest questionnaire. The time between the administration of the first questionnaire and the followup visit for the 116 study subjects ranged from 2 to 18 weeks. The median interval was slightly less than 5 weeks. Almost 80 percent of the sample completed the followup questionnaire within 8 weeks of filling out the first.

The subjects ranged in age from 20 to 50 years, 50 percent were between 20 and 30, 96 percent were white, 67 percent had 4 years of postsecondary education or more, and 25 percent were HIV-antibody positive. No statistically significant differences were found between the 116 participants in the reliability study and the 85 nonparticipants with regard to any of the sociodemographic variables, sexual behaviors, or health behaviors studied. Nevertheless, the study sample tended to be older, to have had more total partners, and to have a higher percentage testing positive for the HIV antibody than the nonparticipants.

There was a high degree of agreement between the successive questionnaires with regard to the nonsexual questions as measured by both percent agreement and Kappa (table 1), although the agreement for dietary habits did not approach the level of agreement for other nonsexual variables. Generally sexuality measures had moderate levels of agreement. Sexual orientation had the highest Kappa, 0.84. Total number of partners and whether there ever was a "high period" had high Kappa values (0.70 and 0.72, respectively). Lifetime total partners appeared to be more reliably

Table 2. Agreement between successive questionnaires of selected sexual behaviors in the last 6 months, by interval between questionnaires

Variable	Less than 6 weeks (N = 61)		6 weeks or more (N = 53)		Difference	95 percent confidence interval
	Kappa	SE	Kappa	SE		
Number of partners	.60	.076	.49	.083	.11	−.11, .33
Frequency of contact	.58	.083	.53	.096	.05	−.12, .30
Number of partners:						
Insertive oral	.63	.076	.50	.100	.13	−.12, .38
Receptive anal	.52	.085	.43	.096	.09	−.16, .34

Table 3. Agreement between successive questionnaires of selected sexual behaviors in the last 6 months, by numbers of partners in the last 6 months

Variable	0 to 4 partners (N = 53)		5 or more partners (N = 63)		Difference	95 percent confidence interval
	Kappa	SE	Kappa	SE		
Frequency of contact	.62	.089	.47	.093	.15	−10, .40
Number of partners:						
Insertive oral	.63	.106	.55	.087	.08	−.15, .30
Receptive anal	.62	.094	.38	.085	.24	−.01, .49

reported than either numbers of partners in the last 6 months or numbers of partners during the high period. The other general sexuality measures appeared to be reported with a similar but lower magnitude of agreement.

All of the specific sexuality items for the last 6 months had Kappa values indicating moderate agreement. Overall, there did not appear to be any differences in agreement between the reporting of numbers of partners per activity and frequency of activity during the last 6 months. There was a nonsignificant tendency for the differences in agreement to be in the direction of a decrease in the frequency of activity at the second interview. Agreement in the reporting of frequency of activity during the "high period" (data not shown) was similar to that for the last 6 months, although items for the high period on average had lower Kappa (mean = 0.46, SD = 0.01). Items assessing change in behavior during the last 5 years were much less reliably reported than the rest (mean Kappa = 0.34, SD = 0.03).

There were no statistically significant differences in agreement between subjects who answered the followup questionnaire within 6 weeks of answering the initial questionnaire and those for whom the interval was 6 weeks or more (table 2). Regardless of the number of partners during the last 6 months (table 3), subjects tended to report activities during that period with a similar degree of agreement. Similarly, no significant differences were found between those with less than 4 years of postsecondary education and those with 4 years or more, nor between subjects who reported a change in behavior between the two questionnaire administrations (data not shown). Although not statistically significant, there was a tendency for subjects to report receptive anal activity more reliably if they were better educated, and less reliably if they had had five partners or more in the last 6 months.

Discussion

As anticipated, self-reported sexual behavior information was not as reliable as the demographic information gathered, probably because of the increase within individual variability and the sensitive nature of the information. For example, a subject's sexual activity level might fluctuate widely over a 6-month period, while his educational level would remain nearly constant. There was no evidence that certain sexual behaviors were more reliably reported than others. Nevertheless, items that asked about a perceived change in behavior over time were less reliably reported than specific quantitative measures. This finding may be due to the way the questions were asked or to the fact that the period asked about was remote in time. For items related to the past 6-month period,

'Increasing the reliability of self-reported sexual behavioral data might allow the association of behaviors that transmit the virus less efficiently to become apparent statistically.'

a moderate amount of reliability, over that which could be expected by chance, was attained. Items relating to a subject-defined previous "high period" were only slightly less reliable than similar items for the past 6 months.

Test-retest reliability is sensitive to the interval between questionnaires; too little time and memory of previous answers could bias results, while with too much time behavior change could occur. The interval we used was similar to intervals found in other studies (*5,6*). We found no differences in reliability owing to length of this interval.

Martin and Vance suggested (*4*) that the reliability of sexual behavior information decreases as the frequency of the activities increases. In this study there were no significant differences in reliability as the numbers of partners increased, although receptive anal activity was more reliably reported for subjects with fewer partners. Furthermore, as the numbers of partners increased, the response categories became broader. This may have masked a lower reliability at higher levels of activity.

Educational level had little effect on the reliability of the information collected. An effect of education on agreement might be evident in a sample with a broader range of educational levels.

There were no significant differences in Kappa scores between those persons who reported behavior changes between questionnaires and those who did not. The categories of measurement of sexual activity were so broad, however, that small changes in behavior might not be detected. For example, a person who stated on the initial questionnaire that he practiced insertive orogenital activity once a month during the past 6 months and then increased to three times a month at the time of the followup questionnaire, would be in the same response category each time. Because the subjects were reporting average behavior for 6 months, a change during a 6-week interval might not affect that average. Behavior change itself was measured by asking if there was any change between questionnaires. There was no way to assess the validity or reliability of this behavior change question.

Although we tried to make retest conditions as similar as possible to those of the initial survey, there were differences. For example, the first questionnaire was administered at the initial study visit, while the second occurred just prior to subjects receiving the HIV antibody test results; HIV risk reduction education could have taken place during or as a result of the initial visit, and the second questionnaire was considerably shorter than the first. Only three studies of sexual behavior have described evaluations of the reliability of the data (*10–12*). Major methodologic problems with reliability evaluations include the length of the intervals between questionnaires being too long (*10*) and too short (*11*), and the lack of a systematic structure for the two interviews (*12*).

High reliability, although necessary for validity, does not assure validity. Subjects may be quite consistent in responses, but the responses may not be accurate. There is evidence that people tend to bias responses toward the socially desirable answers (*13*). Under the pressure of the AIDS epidemic, subjects may under-report sexual activity. Some subjects in this study commented that participation in the study had educated them in various aspects of AIDS. As a result of completing the first questionnaire, some subjects may have answered the followup questions in a more socially desirable way. There appeared to be some systematic shifts in the responses from less "safe-sex"-oriented categories to more "safe-sex"-oriented categories. Although one explanation for this is behavior change, this shift is in the direction of greater social desirability. As the shifts were slight, firm conclusions are not possible.

A random lack of reliability in the reporting of sexual behaviors will increase the variance of an estimate. This factor will cause a reduction in the statistical power of a significance test to detect an association between a particular behavior and a measure of disease (or other outcome), other things being equal. This reduction in power may be offset by an increase in sample size. If the lack of reliability is systematic, however, it can mask an association that exists or spuriously show an association, depending on the nature of the bias. For example, if subjects tend to give "socially desirable responses," subjects with low levels of a particular activity may respond accurately, while subjects with a high level will bias their responses downward. An association with a behavior reported with this type of bias is likely to be

masked. If all subjects bias their responses in the same way, the presence of an association would still be detected, but the point estimate of the degree of the association would be biased.

Reliability of information on past behaviors, especially sensitive ones, can be improved through the use of aided-recall techniques (*13,14*) or the random response technique (*13,15*) in interview situations. The use of diaries eliminates the problem of errors due to recall and can supplement questionnaire information (*16*). Since behaviors more distant in time apparently have lower reliability, a prospective study design might improve reliability. As some information is bound to necessitate recall, the length of the recall period should be kept short. This study's shortest recall period was 6 months, to which a moderate amount of reliability was attained. By shortening this time period by 2 or 3 months, one may increase the reliability of the information. This would be likely to reduce the within-individual variability.

The dissemination of HIV, the virus that is associated with AIDS, has been associated with homosexual activity, particularly receptive anal contact with multiple partners (*2,3*). Nevertheless, other types of intimate contact have not been categorized as safe, because so few subjects in these studies practice only one type of sexual behavior. This has posed a dilemma for clinicians and public health educators in advising persons at risk for HIV infection as to what truly constitutes "safe sex." Increasing the reliability of self-reported sexual behavioral data might allow the association of behaviors that transmit the virus less efficiently to become apparent statistically.

References

1. Landesman, S. J., Ginzburg, H. M., and Weiss, S. H.: The AIDS epidemic. N Engl J Med 312:521-25, Feb. 21, 1985.
2. Goedert, J., et. al: Determinants of retrovirus (HTLV-III) antibody immunodeficiency condition in homosexual men. Lancet 2:711-716, Sept. 29, 1984.
3. Mayer, K. H. et al.: Association of human lymphotropic virus type III antibodies with sexual and other behaviors in a cohort of homosexual men from Boston with and without generalized lymphadenopathy. Am J Med 80:357-363 (1986).
4. Martin, J. L. and Vance, C. S.: Behavioral and psychosocial facctors in AIDS: methodological and substantive issues. Am Pscyhol 39:1303-1308 (1984).
5. Blotcky, A. D., Grandmaisson, S. L., and Carscaddon, D. M.: Reliability of retrospective self-reports of illness. Psychol Rep 54:779-782 (1984).
6. O'Mallay, P., Bachman, J., and Johnston, L.: Reliability and consistency in self-reports of drug use. Int J Addict 18:805-824 (1983).
7. Last, J. M.: A dictionary of epidemiology. Oxford University Press, New York, 1983.
8. Bishop, Y. M. M., Fienburg, S. E., and Holland, P. W.: Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, MA, 1975.
9. Landis, R. J., and Koch, G. G.: The measurement of observer agreement for categorical data. Biometrics 33:159-174 (1977).
10. Kinsey, A. C., Pomeroy, W. B., and Martin, C. E.: Sexual behavior in the human male. W. B. Saunders Co., Philadelphia, 1948.
11. Delameter, J., and MacCorquedale, P.: The effects of interview schedule variation of reported sexual behavior. Social Meth Res 4:215-236 (1975).
12. Ehrmann, W. W.: Premarital dating behavior. Henry Holt, New York, 1959.
13. Bradburn, N. M.: Response effects. *In* Handbook of survey research, edited by P. H. Rossi, J. D. Wright, and A. B. Anderson. Academic Press, New York, 1983.
14. Stanley, J. C.: Reliability. *In* Educational measurement, edited by R. L. Thorndike. American Council on Education, Washington, DC, 1971.
15. Spanier, G. B.: Recall data in sexual behavior surveys. Soc Biol 23:244-253 (1976).
16. Conrath, D. W.: Diary versus questionnaire data reliability. Soc Netw 5:315-322 (1983).