Archives of Sexual Behavior, Vol. 35, No. 2, April 2006, pp. 225–234 (© 2006)
DOI: 10.1007/s10508-005-9006-2

Test–Retest Reliability of Self-Reported Sexual Behavior, Sexual Orientation, and Psychosexual Milestones Among Gay, Lesbian, and Bisexual Youths

Eric W. Schrimshaw, M.A.,[1,5] Margaret Rosario, Ph.D.,[1,2] Heino F. L. Meyer-Bahlburg, Dr. rer. nat.,[3] and Alice A. Scharf-Matlick, Ph.D.[4]

Received September 14, 2004; revision received January 31, 2005; accepted January 31, 2005
Published online: 20 April 2006

Despite the importance of reliable self-reported sexual information for research on sexuality and sexual health, research has not examined reliability of information provided by gay, lesbian, and bisexual (GLB) youths. Test–retest reliability of self-reported sexual behaviors, sexual orientation, sexual identity, and psychosexual developmental milestones was examined among an ethnically diverse sample of 64 self-identified GLB youths. Two face-to-face interviews were conducted approximately 2 weeks apart using the Sexual Risk Behavior Assessment Schedule for Homosexual Youths (SERBAS-Y-HM). Overall, the mean of the test–retest reliability coefficients was substantial for 6 of the 7 domains: lifetime sexual behaviors ($M = .89$), sexual behavior in the past 3 months ($M = .96$), unprotected sexual behavior in the past 3 months ($M = .93$), sexual identity ($\kappa = .89$), sexual orientation ($M = .82$), and ages of various psychosexual developmental milestones ($M = .77$). Inconsistent reliability was found for reports of sexual behaviors while using substances. A small number of gender differences emerged, with lower reliability among female youths in the lifetime number of same-sex partners. The overall findings suggest that a wide range of self-reported sexual information can be reliably assessed among GLB youths by means of interviewer-administered questionnaires, such as the SERBAS-Y-HM.

KEY WORDS: reliability; sexual behavior; condom use; sexual identity; psychosexual development; adolescents.

INTRODUCTION

Sex research has typically relied on participants' self-reports of sexual behaviors, sexual orientation, sexual identity, and psychosexual developmental milestones. However, researchers have questioned the ability of individuals to provide reliable self-reports and whether current measurement strategies reliably assess sexual information (Catania, Gibson, Chitwood, & Coates, 1990; Schroder, Carey, & Vanable, 2003; Weinhardt, Forsyth, Carey, Jaworski, & Durant, 1998). Failure to reliably assess self-reported sexual information would have profound consequences for research on sexuality and sexual health. In the absence of reliable self-reports, the ability to predict sexual behavior or evaluate changes in behavior is greatly reduced. Furthermore, if measurement of sexual behavior is unreliable then, by definition, it is also invalid.

Despite the importance of reliable self-reports of sexually relevant information, few researchers have undertaken test–retest studies to evaluate whether individuals can provide reliable self-reports and whether current measures of sexual behavior reliably elicit such reports. Indeed, a recent review of the research conducted since

[1]Doctoral Program in Psychology, The City University of New York–Graduate Center, New York, New York.
[2]Department of Psychology, The City University of New York–The City College, New York, New York.
[3]HIV Center for Clinical and Behavioral Studies, New York State Psychiatric Institute, New York, New York; Department of Psychiatry, Columbia University, New York, New York.
[4]Department of Psychology, Iona College, New Rochelle, New York.
[5]To whom correspondence should be addressed at the Center for the Psychosocial Study of Health & Illness, Mailman School of Public Health, Columbia University, 100 Haven Avenue, Suite 6A, New York, New York 10032; e-mail: es458@columbia.edu.

1990 identified only 15 studies that examined test–retest reliability of self-reported sexual behavior (Schroder et al., 2003), with far fewer studies conducted prior to 1990 (Catania et al., 1990). Taken together, research on the reliability of self-reported sexual behavior has been characterized as "a mixed bag" (Catania et al., 1990). The more recent research continues to have a number of limitations (see Schroder et al., 2003, for a recent critique). To adequately evaluate reliability, questions must assess the same behaviors for the same period in time (e.g., the past month) at both the test and the retest assessments. However, overly long assessment periods (e.g., 6 months) used is much reliability research increases the likelihood that the two assessments will be nonoverlapping in time and therefore not be assessing the same behaviors. Such long test–retest periods are better characterized not as reliability of reports, but rather as consistent patterns of behavior (Nunnally, 1978). Similarly, overly short retest periods (e.g., 48 hr) increase the likelihood of participants recalling their original reports, thus artificially increasing reliability coefficients. The research is also limited in the scope of sexual behaviors examined, with many studies assessing the reliability of only a few global assessments of sexual behavior (e.g., number of partners, number of times had sex) rather than the reliability of specific sexual behaviors (e.g., frequency of oral, anal, or vaginal sex; with or without condoms; sex while using substances).

Catania et al. (1990) also noted the lack of research on the reliability of reported sexual behaviors by specific subpopulations, including different age groups, genders, and ethnic/racial groups. Indeed, the reliability and validity of adolescents and young adults' self-reported sexual behaviors have been questioned. Several studies have suggested that a sizable number of youths admit to lying about their sexual experience (17%, Newcomer & Udry, 1988) or report being dishonest about their sexual behavior (8–24%, Siegel, Aten, & Roghmann, 1998). In particular, male youths are significantly more likely to overreport their sexual behavior (14% report that they reported ''a lot more'' sexual behavior than they really had) than female youths, whereas female youths underreport their behavior (8% report that they reported ''a lot less'' sexual behavior than they really had; Siegel et al., 1998).

The test–retest reliability of youths' sexual behavior has not been extensively examined (for review, see Brener, Billy, & Grady, 2003). Most of the reliability research among adolescents and young adults has found self-reported sexual behavior to be only moderately reliable (mean reliability in each study = .51–.66; Boekeloo, Schamus, Simmens, & Cheng, 1998; Brener, Collins, Kann, Warren, & Williams, 1995; Brener et al., 2002; Hearn, O'Sullivan, & Dudley, 2003), with only a single study finding high levels of reliability among youths (mean reliability = .92, Durant & Carey, 2002). The low reliability found in past research is partly due to methodological limitations such as nonoverlapping assessment periods (e.g., Boekeloo et al., 1998). The extant literature on adolescents is also limited by the inclusion of only a small number of sexual items, most of which are general in scope (e.g., whether youths ever had sex; Brener et al., 2002; Flisher, Evans, Muller, & Lombard, 2004). As such, more research is needed to examine a broad range of specific sexual behaviors (e.g., oral and anal sex, condom use during specific behaviors, and insertive versus receptive behaviors).

Although the reliability of self-reported sexual behavior has been questioned among adolescents in general, at particular risk for low reliability may be gay, lesbian, and bisexual (GLB) youths. Although social desirability and privacy concerns may reduce the reliability of reported sexual behaviors of all populations (for review, see Catania et al., 1990; Schroder et al., 2003), this may be particularly true among GLB individuals, given the stigma attached to same-sex sexuality. However, few studies have examined the reliability of sexual behavior among GLB individuals. Only three test–retest studies have examined the reliability of adult gay men's reported sexual behavior (Coates et al., 1986; McLaws, Oldenburg, Ross, & Cooper, 1990; Saltzman, Stoddard, McCusker, Moon, & Mayer, 1987). These studies found a wide range of reliability coefficients from poor to near perfect: .40–.99 (Coates et al., 1986), .08–.98 (McLaws et al., 1990), .34–.72 (Saltzman et al., 1987). To date, the reliability of reported sexual behaviors has not been examined among GLB youths.

Although the reliability of reported sexual behaviors has been examined (for review, see Schroder et al., 2003), the reliability of other aspects of sexuality remains unexamined, including sexual orientation, sexual identity, and the self-reported ages of various psychosexual developmental milestones. The ability of GLB youths to reliably report such aspects of their sexuality is critical for research on psychosexual or sexual identity development. In the only study identified that has examined the test–retest reliability of sexual orientation, Saltzman et al. (1987) found good reliability (.84) over a 6-week period among adult gay men. The only study to examine the reliability of retrospective reports of the ages of achieving various sexual milestones (e.g., age first kissed, age first sex) was conducted among heterosexual girls and it found high levels of reliability (mean $r = .85$, Hearn et al., 2003). However, no such studies have examined the reliability of sexual identity, orientation, or psychosexual milestones among GLB youths.

In an attempt to address the absence of research on the reliability of self-reported sexual behavior among GLB youths, this test–retest study examined the reliability of a broad range of sexual behaviors (both lifetime and in the past 3 months) over a 2-week period. In addition, the study examined the reliability of youths' sexual identity, sexual orientation, and psychosexual developmental milestones. Furthermore, because past research has found gender differences in the reporting of sexual behavior, the reliabilities of male and female youths were examined separately.

METHOD

Participants

As part of a larger longitudinal study of 156 GLB youths aged 14–21 years, a subsample of 64 youths also participated in a sub-study of the test–retest reliability of their self-reported sexual behaviors, sexual identity, sexual orientation, and psychosexual developmental milestones. Youths were recruited from five GLB-focused organizations in New York City, including three community-based organizations and two student organizations from public colleges. Additional description of the larger study sample, including descriptive data of the youths' sexual behavior, is available in earlier reports (e.g., Rosario et al., 1996; Rosario, Meyer-Bahlburg, Hunter, & Gwadz, 1999).

Initial interviews with youths who would compose the reliability sub-study were initiated 4 months following the start of the larger study. As such, recruitment and interview procedures had become established and interviewers had become experienced with the interview protocol. Youths who had recently participated in the baseline assessment of the longitudinal study were contacted by telephone and asked to participate in a second interview scheduled approximately 2 weeks after their original interview. So as to not bias their responses, youths were not told at baseline that the reliability of their responses would be assessed nor were they told the reason for the retest interview was to assess their reliability. We attempted to reinterview all participants who were accrued into the larger study after initiation of the reliability interviews. For the reliability subsample, specific attention was focused on obtaining approximately equal numbers of male and female youths and youths from all five recruitment sites.

Of the 64 youths who participated in the reliability sub-study, 35 (55%) were males and 29 (45%) were females. Youths were between the ages of 14 and 21 years ($M = 18.1$, $SD = 1.9$ for males; $M = 18.2$, $SD = 1.5$ for females). They self-identified as gay/lesbian (69%), bisexual (28%), or other (3%). The youths were of Latino (42%), Black (28%), White (19%), Asian (5%), and other (6%) ethnic backgrounds. Over one third (38%) of the youths reported that their mother or father received welfare, medicaid, or food stamps (i.e., low socioeconomic status, SES). Most youths (87%) were recruited from community-based organizations and the remainder (13%) from college student organizations. A comparison of youths who participated in the reliability sub-study with youths who did not participate (e.g., those interviewed prior to the initiation of the reliability interviews) found no significant differences on gender, age, ethnicity/race, SES, sexual identity, or recruitment site.

Procedure

As part of the larger study, youths provided voluntary signed informed consent for a longitudinal series of interviewer-administered structured interviews. Parental consent was waived for those youths under age 18 years by the Commissioner of Mental Health for the State of New York. Instead, an adult in each community-based organization served *in loco parentis* to safeguard the rights of the underage participants. The study was approved by the Institutional Review Board of the Psychiatry Department of Columbia University and by the recruitment sites.

All interviews for the test–retest sub-study were conducted between January and June 1994, with follow-up interviews conducted approximately 2 weeks later ($M = 17.1$ days, $SD = 4.36$). A 2-week interval was selected for the test–retest administration because this interval is long enough to minimize recall of responses provided at the original assessment, but sufficiently brief both to reduce the likelihood of new sexual behaviors between the test and the retest assessments and to minimize the nonoverlapping portion of the reporting periods for recent sexual behaviors (e.g., in the past 3 months). Indeed, some researchers have suggested a 2-week interval as ideal for test–retest studies (e.g., Nunnally, 1978; Wiederman, 2002) for these reasons.

Interviews were conducted in a private room at each recruitment site. Each youth received $30 for his or her participation at both the initial and the retest assessments. Interviews were conducted by an ethnically diverse group of college-educated male and female interviewers who were purposefully matched to participants on gender, but not necessarily on race/ethnicity. No attempt was made to have the same interviewer conduct the baseline and retest interviews.

Every interviewer received 20 hr of training on conducting interviews on sexually sensitive topics and interviewing techniques (e.g., probing for accuracy of responses, tracking the logical consistency of responses over the course of the interview, building rapport with the youths; Dugan & Meyer-Bahlburg, 2003). Training was conducted by experts in the area of sexuality assessment. As part of their training, each interviewer conducted four practice interviews. Audiotaped interviews were monitored throughout the study to ensure quality and consistency. Interviewers received feedback from the researchers in both individual and group supervision.

Measures

Sexual behavior (both lifetime and in the past 3 months), sexual identity, sexual orientation, and psychosexual developmental milestones were assessed with the Sexual Risk Behavior Assessment Schedule for Homosexual Youths (SERBAS-Y-HM; Meyer-Bahlburg, Ehrhardt, Exner, & Gruen, 1994). The SERBAS-Y-HM is a semi-structured interview schedule with male (M-1) and female (F-1) versions. The SERBAS-Y-HM consists of approximately 300 items, but because of skip patterns throughout the interview, the number of items administered is dependent on the responses reported by each youth. It requires approximately 45 min to administer. The current version of the SERBAS-Y-HM is based on an earlier version of the SERBAS-Y for gay/bisexual male youths (Meyer-Bahlburg, Ehrhardt, Exner, & Gruen, 1988). Revisions were based on focus groups with GLB youths at community-based agencies serving these youths and discussions with staff serving these youths.

Lifetime Sexual Behaviors

A series of items assessed the lifetime prevalence of various sexual behaviors, including the number of sex partners, number of sexual encounters, sex in exchange for goods, and sexual partners at risk for HIV infection. After defining the various sexual behaviors to be assessed in the survey and the youths' own terminology for each behavior, youths were asked to "count up" all the same-sex partners with whom they had "any kind of sex within their whole lifetime." This was followed by a question about the total number of times they had sex with these partners. The lifetime prevalence of the exchanging sex for goods was assessed by asking youths questions about whether they had ever received money, drugs, or a place to stay from a same-sex partner in exchange for sex. Questions also assessed whether youths had ever *given* money, drugs, or a place to stay in exchange for sex, but no youths reported this behavior. Experiences with potentially risky sexual partners were assessed by asking youths whether they had ever had a same-sex partner who had injected drugs, had a sexually transmitted disease, or had tested positive for HIV/AIDS. Lesbian and bisexual female youths were asked whether they had ever had a sexual partner who was a gay or bisexual male. With the exception of this last question, identical questions were asked regarding the same behaviors with other-sex partners. The other-sex questions for these and all other subsections always followed the same-sex questions.

Recent Sexual Behaviors

A series of items assessed the prevalence of various sexual risk behaviors in the past 3 months. After requesting personally relevant events to clarify the 3-month period of interest, youths were asked whether they had any sex (previously defined for them) with a same-sex partner in the past 3 months. If appropriate, youths were then asked to "count up" the number of same-sex sexual partners they had in the past 3 months. Youths were subsequently asked the number of times they had engaged in various sexual behaviors with each of these same-sex partners (separately for active/insertive and passive/receptive), including vaginal–digital sex (for females only), oral sex, oral–anal sex, and anal sex (for males only). A total number of episodes for each sexual behavior was assessed by adding the number of passive and active encounters. Additional items assessed, for each behavior, the number of sexual encounters in which condoms or other appropriate HIV barrier methods were used and the number of encounters in which the youths used drugs or alcohol right before or during sexual activity. We computed the number of unprotected sexual encounters by subtracting the number of protected encounters from the total number of encounters. Corresponding data on other-sex sexual behaviors were also collected. However, with the exception of the overall prevalence of any sex with the other sex, the frequency of specific sexual behaviors with the other sex was too infrequent for reliability analysis; only 19% of youths reported any recent other-sex sexual behaviors.

Sexual Identity

A single item assessed sexual identity, "When you think about sex, do you think of yourself as lesbian/gay, bisexual, or straight?" Youths who rejected these identities were coded as "other."

Sexual Orientation

Sexual orientation was assessed with three items that asked youths to indicate the degree to which in the past 3 months their recent sexual attractions, thoughts, or fantasies focused on the same sex or the other sex: (1) when in the presence of other individuals in a public setting (i.e., sexual attractions), (2) when masturbating, dreaming, or daydreaming (i.e., sexual fantasies), and (3) when viewing erotic materials in films, magazines, or books (i.e., erotica).[6] A 7-point, Kinsey-type response scale was used ranging from 0 (*always girls/women*) to 6 (*always guys/men*), with a midpoint 3 indicating *equally guys/men and girls/women*. The scale was reversed for female youths. Youths who indicated not experiencing the assessed event were coded as such. The mean of these three items was computed as an assessment of overall cognitive sexual orientation (Cronbach's $\alpha = .92$ in the initial assessment of the reliability subsample).

Psychosexual Developmental Milestones

The youths were asked the ages when they first experienced various milestones in the development of sexual orientation, sexual behavior, and sexual identity. They were asked the ages when they were first (1) erotically attracted to, (2) had thoughts or fantasies about, and (3) were aroused by erotica focused on the same-sex. Similar items assessed ages at which youths first experienced attractions, fantasies, and erotic arousal toward the other sex. Youths were asked the ages when they first engaged in various sexual behaviors with the same sex and the ages when they first engaged in various sexual behaviors with the other sex. On the basis of these responses, the minimum age reported was used as the age when they first had any sex with the same sex and the age when they first had any sex with the other sex. Finally, youths were asked about the ages when they first thought they "might be" bisexual, when they thought they "might be" gay/lesbian, when they thought they "really were" bisexual, and when they thought they "really were" gay/lesbian. Youths who indicated not experiencing the assessed event were coded as such.

Data Analysis

Test–retest reliability was computed using kappa (κ) for categorical variables (Cohen, 1968) and intra-class correlations (ICC) for continuous variables (e.g., Bartko, 1966). The rationale for the use of kappa and ICC over Pearson or Spearman correlations have been argued elsewhere (e.g., Schroder et al., 2003). Briefly, although *inter*class correlations (e.g., Pearson, Spearman) are appropriate for examining the relation between two independent variables, these correlations are inappropriate when the two variables share variance (e.g., two assessments of the same variable). In cases of common variance, *intra*class correlations are used (e.g., McGraw & Wong, 1996). Because correlation coefficients are asymmetrically distributed, correlations were transformed using Fisher's *r*-to-*z* transformation (Hays, 1994), averaged, and then back-translated to correlations, so that mean reliability coefficients could be obtained for each domain.

RESULTS

Lifetime Sexual Behaviors

Test–retest reliability of self-reported lifetime prevalence of sexual behaviors are presented in Table I. Overall, youths were found to reliably report lifetime prevalence of sexual behaviors ($M = .89$, range .69–1.00). The lifetime number of same-sex sexual partners (ICC = .96) and the prevalence of exchanging sex for goods with a same-sex partner ($\kappa = 1.0$) were among the most reliably reported. The one exception to this trend was youths' reports of the lifetime number of same-sex sexual encounters. The moderate reliability found for this variable (ICC = .49) was attributable to a low value among the female youths (ICC = .41)[7] as compared with male youths (ICC = .81). Indeed, examination of this observed difference in the reliability coefficients indicated that although female youths were found to provide somewhat more reliable reports than male youths ($M = .94$ versus .88, respectively), female youths had a wider range of reliability coefficients (.41–1.00) than did male youths (.64–1.00) on reports of lifetime sexual behaviors.

Recent Sexual Behaviors

The reliability coefficients of sexual risk behaviors in the past 3 months are presented in Table II. Overall, youths reliably reported recent sexual risk behaviors ($M = .96$,

[6] A fourth item assessing sexual orientation with respect to past sexual behaviors (consistent with the Kinsey definition of sexual orientation was also assessed. However, because it was interviewer-rated, not self-reported, it was not included in this analysis of the reliability of self-reported sexuality.

[7] Examination of the data did not indicate the existence of any outliers, but rather, a pattern of inconsistency characterizing a large number of female youths. For other low reliability coefficients, the potential impact of outliers also was examined, but not found.

Table I. Test–Retest Reliability of Reports of Lifetime Sexual Behaviors

	Total			Males			Females		
	N	% or *M* (*SD*)	κ or ICC	*N*	% or *M* (*SD*)	κ or ICC	*N*	% or *M* (*SD*)	κ or ICC
Lifetime number of same-sex sexual partners	63	10.6 (22.3)	.96	35	14.7 (28.3)	.95	28	5.6 (10.3)	.98
Lifetime number of same-sex sexual encounters	63	205.6 (602.6)	.49	35	128.2 (236.4)	.81	28	298.9 (855.9)	.41
Lifetime number of other-sex sexual partners	64	4.5 (9.8)	.88	35	3.6 (11.1)	.84	29	5.6 (8.2)	.94
Lifetime number of other-sex sexual encounters	64	64.3 (139.8)	.82	35	30.4 (71.9)	.70	29	105.3 (185.9)	.82
Ever received money, drugs, or lodging for sex with a same-sex partner	57	11%	1.00	33	15%	1.00	24	4%	1.00
Ever received money, drugs, or lodging for sex with an other-sex partner	39	10%	.84	16	13%	.64	23	9%	1.00
Ever had a same-sex partner who injected drugs, tested positive for AIDS, or had an STD	63	16%	.69	35	23%	.75	28	7%	.46
Ever has an other-sex partner who injected drugs, tested positive for AIDS, or had an STD	64	11%	.70	35	6%	.65	29	17%	.71
Ever had an other-sex partner who was gay or bisexual male	29	21%	1.00		NA		29	21%	1.00

Note. %: percent who reported ''yes'' to behavior at the baseline assessment. *M*: mean at the baseline assessment. κ: Cohen's kappa. ICC: intraclass correlation. NA: not asked/not applicable. Percentages are reported for dichotomous variables and means for continuous variables. Cohen's kappa is reported for dichotomous variables and intraclass correlations for continuous variables.

range = .68–1.00), with male and female youths having nearly identical reliability ($M = .96$ and .94, respectively). Indeed, the prevalence of sexual behavior with an other-sex sexual partner ($\kappa = 1.0$), the number of same-sex partners (ICC = .96) and encounters (ICC = .91) were the most reliably reported. Reports of unprotected sex were all quite reliable ($M = .93$, range = .77–.99), with no apparent gender differences ($M = .91$ for males and .94 for females). Two gender-specific exceptions should be noted to this general pattern. First, youths (particularly female youths, $\kappa = .60$) were moderately reliable in their reports of whether they had a same-sex sexual encounter in the past 3 months. Second, whereas reports of vaginal–digital, oral, and analingus sexual behaviors while on alcohol or drugs were generally reliable (range = .69–1.0), reports of anal sex while using drugs or alcohol (which was asked only of male youths) were found to be poor (ICC = −.01–.24).

Sexual Identity, Sexual Orientation, and Developmental Milestones

The reliability coefficients of self-reported sexual identity, sexual orientation, and psychosexual development milestones are presented in Table III. Youths reliably reported ($\kappa = .89$) their sexual identity as gay/lesbian, bisexual, or other. Similarly, youths' sexual orientation was reliable when assessed as attractions to others in public and in their fantasies (ICC range = .85–.89), but both male and female youths were only moderately reliable about erotica (ICC range = .63–.66). Youths reliably reported the ages at which they experienced various psychosexual developmental milestones ($M = .77$, range = .66–.88), with female youths somewhat more reliable than male youths ($M = .85$ and .77, respectively). One gender-specific exception was noted; female youths were found to have only moderate reliability (ICC = .45) in reporting the age when they first were sexually "turned on" by same-sex erotica.

DISCUSSION

Despite the importance of reliable sexual information regarding GLB individuals, the current study, as far as we know, represents the first test–retest study of various aspects of sexuality among GLB youths. Overall, substantial to almost perfect reliability was obtained using the SERBAS-Y-HM among GLB youths on a variety of aspects of their sexuality, including lifetime sexual behavior, recent sexual behavior, unprotected sexual risk behavior, sexual identity, sexual orientation, and ages of psychosexual developmental milestones. The reliability found here is substantially higher than that found among most past research among primarily heterosexual adolescents or GLB adults.

Two potential explanations exist for the strong reliability found in this study. First, the SERBAS-Y-HM includes strategies that have been recommended by

Table II. Test–Retest Reliability of Reports of Sexual Behaviors in the Past 3 Months

	Total			Males			Females		
	N	% or *M* (*SD*)	κ or ICC	*N*	% or *M* (*SD*)	κ or ICC	*N*	% or *M* (*SD*)	κ or ICC
Any same-sex sex in past 3 months	64	69%	.68	35	74%	.77	29	62%	.60
Any other-sex sex in past 3 months	64	19%	1.00	35	9%	1.00	29	31%	1.00
Number of same-sex partners in past 3 months	64	1.4 (2.9)	.96	35	1.8 (3.8)	.97	29	0.9 (1.1)	.85
Number of same-sex sexual encounters in past 3 months	43	22.8 (27.4)	.91	25	17.3 (18.5)	.89	18	30.8 (35.8)	.92
Unprotected encounters with same-sex									
Number of unprotected vaginal–digital encounters	17	34.3 (61.7)	.77		NA		17	34.3 (61.7)	.77
Number of unprotected oral encounters	37	30.2 (43.5)	.90	23	19.8 (20.9)	.82	14	48.6 (64.5)	.90
Number of unprotected oral–anal encounters	19	9.0 (19.7)	.97	13	5.7 (7.6)	.80	6	17.5 (35.7)	.99
Number of unprotected anal encounters	19	8.0 (13.1)	.96	19	8.0 (13.1)	.96		NA	
Number of unprotected receptive anal encounters	16	5.6 (9.4)	.92	16	5.6 (9.4)	.92		NA	
Number of unprotected insertive anal encounters	14	4.4 (7.9)	.96	14	4.4 (7.9)	.96		NA	
Same-sex encounters while using drugs or alcohol									
Number of vaginal–digital encounters while using substances	17	1.1 (2.3)	.91		NA		17	1.1 (2.3)	.91
Number of oral encounters while using substances	37	0.7 (1.7)	.74	23	0.6 (1.4)	.69	14	0.9 (2.1)	.77
Number of oral–anal encounters while using substances	19	0.0 (0.2)	1.00	13	0.0 (0.0)	U	6	0.2 (0.4)	1.00
Number of anal encounters while using substances	19	0.4 (1.2)	.24	19	0.4 (1.2)	.24		NA	
Number of receptive anal encounters while using substances	15	0.1 (0.5)	− .01	15	0.1 (0.5)	− .01		NA	
Number of insertive anal encounters while using substances	14	0.4 (1.1)	.17	14	0.4 (1.1)	.17		NA	

Note. %: percent who reported ''yes'' at the baseline assessment. *M*: mean at the baseline assessment. κ: Cohen's kappa. ICC: intraclass correlation. NA: not asked/not applicable. U: undefined; data were a constant, thus no value can be computed. Percentages are reported for dichotomous variables and means for continuous variables. Cohen's kappa is reported for dichotomous variables and intraclass correlation for continuous variables.

experts in sexual behavior assessment to enhance the reliability and validity of the behaviors assessed, including (1) defining sexual terms (e.g., what do you mean by "sex"; Wiederman, 2002), (2) using nontechnical jargon by exploring and using the youths' own language and terms for sexual behaviors (e.g., "tossing salad"; Catania et al., 1990), (3) focusing on a short, 3-month recall assessment (Schroder et al., 2003), (4) using participant-nominated events in order to personally anchor and clarify the assessment window (Weinhardt et al., 1998), (5) assessing behaviors with respect to each specific partner, and (6) utilizing qualitative research to inform item content and language (Weinhardt et al., 1998). Second, the interviewers were highly trained and experienced with the administration of the SERBAS-Y-HM, comfortable with discussing sexual topics, and comfortable with the GLB population. Unfortunately, it is impossible to determine which aspects of the SERBAS-Y-HM or the interviewer training played critical roles in the reliability of the reports assessed here. Nevertheless, researchers are encouraged to employ measures that, like the SERBAS-Y-HM, incorporate strategies to enhance the reliability and validity of self-reported sexual information.

Despite the generally high reliability found among these youths, some exceptions were noted. Although there were generally few observed differences in the reliability of male and female youths' reports, instances of moderate or low reliability were often gender-specific. For example, female youths were found to have only fair agreement on the number of sexual partners in their lifetime, whereas male youths provided almost perfect reliability on this question. In contrast, male youths were found to provide poor reliability in their reports of anal sex while using alcohol or drugs (female youths were not asked about anal sex). This poor reliability may be due to the rarity of this behavior, which reduced

Table III. Test–Retest Reliability of Reports of Sexual Identity, Sexual Orientation, and Psychosexual Developmental Milestones

	Total			Males			Females		
	N	% or *M* (*SD*)	κ or ICC	*N*	% or *M* (*SD*)	κ or ICC	*N*	% or *M* (*SD*)	κ or ICC
Sexual identity (% gay/lesbian)	64	69%	.89	35	80%	.82	29	55%	.93
Sexual orientation: Attractions	63	5.0 (1.4)	.85	34	5.2 (1.3)	.92	29	4.8 (1.6)	.79
Sexual orientation: Fantasies	63	5.2 (1.3)	.89	34	5.4 (1.1)	.86	29	5.0 (1.5)	.90
Sexual orientation: Erotica	59	5.2 (1.4)	.66	34	5.4 (1.4)	.63	25	4.9 (1.4)	.66
Mean sexual orientation	64	5.1 (1.3)	.88	35	5.3 (1.2)	.90	29	4.9 (1.4)	.87
Psychosexual developmental milestones									
Age first sexually attracted to same sex (in years)	63	10.9 (3.7)	.78	35	10.9 (3.8)	.77	28	10.9 (3.7)	.80
Age first had sexual fantasies about the same sex (in years)	64	11.7 (3.4)	.73	35	11.3 (3.7)	.74	29	12.2 (3.1)	.69
Age first sexually turned on by same-sex erotica (in years)	58	12.2 (3.4)	.66	34	11.8 (3.8)	.70	24	12.8 (2.4)	.45
Age first had sex with the other sex (in years)	41	12.3 (3.8)	.88	18	10.7 (4.3)	.93	23	13.5 (2.9)	.79
Age first thought might be bisexual (in years)	39	13.3 (3.1)	.70	17	13.0 (3.4)	.81	22	13.6 (2.9)	.61
Age first thought might be gay/lesbian (in years)	51	12.5 (3.8)	.74	30	11.5 (3.6)	.62	21	13.8 (3.8)	.93
Age first had sex with the same-sex (in years)	57	13.5 (3.8)	.80	33	13.2 (3.9)	.66	24	14.1 (3.8)	.99
Age first thought really was bisexual (in years)	27	15.0 (2.5)	.84	11	14.5 (3.4)	.89	17	15.3 (1.5)	.62
Age first thought really was gay/lesbian (in years)	48	15.2 (2.5)	.76	28	14.7 (2.5)	.64	20	15.8 (2.5)	.97

Note. %: percent who reported ''yes'' at the baseline assessment. *M*: mean at the baseline assessment. κ: Cohen's kappa. ICC: intraclass correlation. Percentages are reported for dichotomous variables, and means for continuous variables. Cohen's kappa is reported for dichotomous variables and intraclass correlation for continuous variables.

the sample size and potential variability. Nevertheless, it should be noted that the numbers of moderate or low reliabilities observed were less than expected by chance. Future research must determine whether the low reliabilities are chance findings or indicate problematic measurement.

Unreliable findings also have serious methodological implications for sex research in general. For example, youths reported only moderate reliability ($\kappa = .77$ for males and .60 for females) on whether they had any same-sex sexual behavior in the past 3 months. Although this would suggest that youths are only moderately able to recall their recent sexual behaviors, in fact, they provided highly reliable reports of the number of recent partners and the number of recent specific sexual acts (e.g., vaginal, oral, anal; with or without a condom). This inconsistency suggests that perhaps, despite our efforts to clarify what we meant by "sex," some youths were confused by this general term, but not when asked about specific behaviors. Thus, the use of general questions may be unreliable and research should focus on specific sexual behaviors. This would also imply that general questions should not be used to determine whether to skip a section of more detailed sexual inquiry; instead, specific behaviors should be assessed, regardless of any response to more general sex questions.

Given the recent advances in computer-assisted interviewing (e.g., Audio-CASI), some may question whether the use of a face-to-face interview for the assessment of sexual behavior is a reliable and valid method of assessment. Indeed, many have suggested that the greater privacy afforded by Audio-CASI assessments would increase the reliability and validity of self-reported sexual behavior (for review, see Schroder et al., 2003). Although some research has indicated that Audio-CASI results in more reports of potentially stigmatizing sexual behaviors than do face-to-face interviews (Des Jarlais et al., 1999), most of the research has identified only a small number of differences between interviews and Audio-CASI in the reports of sexual behaviors (Ellen et al., 2002; Macalino, Celentano, Latkin, Strathdee, & Vlahov, 2002; Metzger et al., 2000; Williams et al., 2000). Indeed, some of these observed differences are in the opposite direction, with more sexual behaviors disclosed via face-to-face interviews than with Audio-CASI (Ellen et al., 2002; Jennings, Lucenko, Malow, & Devieux, 2002; Williams et al., 2000). Furthermore, at least some past research has suggested that test–retest reliability of sexual behavior is greater in face-to-face interviews than when using Audio-CASI (Williams et al., 2000). Although it is unclear whether Audio-CASI results in more reliable and valid assessments of sexual behavior, face-to-face interviews may have some potential advantages in some populations, such as among those with low educational background or those who are uncomfortable using computers. Face-to-face interviews have the added benefits of allowing for the exploration of the individuals' own terms for various sexual behaviors, perceiving possible confusion and clarification of questions, exploring of potential logical inconsistencies, and building trust and

rapport with the participant—none of which are adequately duplicated with the use of Audio-CASI. Indeed, this report provides evidence that sexual information can be reliably obtained via face-to-face interviews and earlier reports from this study using the SERBAS-Y-HM provide evidence of the construct validity of this interviewer-administered assessment (e.g., Rosario, Hunter, Maguen, Gwadz, & Smith, 2001; Rosario, Mahler, Hunter, & Gwadz, 1999; Rosario, Schrimshaw, & Hunter, 2004).

The present sub-study has limitations. First, the sample size for the test–retest study was limited. Although we had a sufficient sample to examine reliability separately for male and female youths, we had insufficient numbers to examine potential ethnic/racial differences in reliability. A second limitation is that the sample was recruited from GLB-focused organizations in a major urban area. As such, these GLB youths may not be representative of the population of GLB youths. These youths may have been further along in the development of their GLB identity and more comfortable discussing their sexuality than youths who might not be involved in GLB organizations. As such, these youths' reports may have been more reliable than might be found among samples less comfortable with their sexuality. Similarly, the findings from this ethnically diverse and urban sample may not generalize to other GLB populations. A third potential limitation is the use of a 2-week test–retest period. Although the 2-week retest is recommended by psychometric texts to prevent recall (e.g., Nunnally, 1978) and is sufficiently brief to help ensure that new behaviors did not occur between test and retest (thereby biasing the reliability estimates), this brief retest period might increase the possibility of participants recalling their original responses and artificially increasing their reliability coefficients. As such, future reliability research may wish to employ longer test–retest periods to determine whether the reliability in reports observed here are replicated over longer periods (but not so long as to assess behaviors in two nonoverlapping time periods). Finally, this report demonstrated that GLB youths were able to reliably report sexual information, this study does not provide any information about the validity of these reports. Although reliability is necessary for validity, the reverse is not true. Thus, the high reliabilities identified here are not necessarily indicative that youths were accurate in their reports of sexual information. Future research into the validity of sexual reports are needed.

Despite these limitations, the findings provide preliminary but critical information regarding the reliability of self-reported sexual information among GLB youths. However, given the importance of reliable reports of sexual information and the scarcity of empirical reports examining reliability, future research is needed into reliability of self-reported sexual information among all groups including adolescents and GLB individuals.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Mental Health Center Grant P50-MH43520 (Anke A. Ehrhardt, Principal Investigator of the HIV Center for Clinical and Behavioral Studies; Margaret Rosario, Principal Investigator of research project "HIV Risk and Coming-Out Among Gay and Lesbian Adolescents"; and Heino F. L. Meyer-Bahlburg, Principal Investigator of the Psychosexual Core).

REFERENCES

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19,* 3–11.

Boekeloo, B. O., Schamus, L. A., Simmens, S. J., & Cheng, T. L. (1998). Ability to measure sensitive adolescent behaviors via telephone. *American Journal of Preventative Medicine*, *14,* 209–216.

Brener, N. D., Billy, J. O. G., & Grady, W. R. (2003). Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: Evidence from the scientific literature. *Journal of Adolescent Health*, *33,* 436–457.

Brener, N. D., Collins, J. L., Kann, L., Warren, C. W., & Williams, B. I. (1995). Reliability of the youth risk behavior survey questionnaire. *American Journal of Epidemiology*, *141,* 575–580.

Brener, N. D., Kann, L., McManus, T., Kinchen, S. A., Sundberg, E. C., & Ross, J. G. (2002). Reliability of the 1999 youth risk behavior survey questionnaire. *Journal of Adolescent Health*, *31,* 336–342.

Catania, J. A., Gibson, D. R., Chitwood, D. D., & Coates, T. J. (1990). Methodological problems in AIDS behavioral research: Influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin*, *108,* 339–362.

Coates, R., Soskolne, C., Clazavara, L., Read, S., Fanning, N., Schpahard, F., et al. (1986). The reliability of sexual histories in AIDS related research: Evaluation of an interview administered questionnaire. *Canadian Journal of Public Health*, *77,* 343–348.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70,* 213–220.

Des Jarlais, D. C., Paone, D., Milliken, J., Turner, C. F., Miller, H., Gribble, J., et al. (1999). Audio-computer interviewing to measure risk behaviour for HIV among injecting drug users: A quasi-randomized trial. *Lancet*, *353,* 1657–1662.

Dugan, T. M., & Meyer-Bahlburg, H. F. L. (2003). A training program for sex research interviewers. In D. di Mauro, G. Herdt, & R. Parker (Eds.), *Handbook of sexuality research training initiatives* (pp. 80–92). New York: Social Science Research Council.

Durant, L. E., & Carey, M. P. (2002). Reliability of retrospective self-reports of sexual and nonsexual health behaviors among women. *Journal of Sex and Marital Therapy*, *28,* 331–338.

Ellen, J. E., Gurvey, J. E., Pasch, L., Tschann, J., Nanda, J. P., & Catania, J. (2002). A randomized comparisons of A-CASI and

phone interviews to assess STD/HIV-related risk behaviors in teens. *Journal of Adolescent Health*, *31,* 26–30.

Flisher, A. J., Evans, J., Muller, M., & Lombard, C. (2004). Test–retest reliability of self-reported adolescent risk behavior. *Journal of Adolescence*, *27,* 207–212.

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Hearn, K. D., O'Sullivan, L. F., & Dudley, C. D. (2003). Assessing reliability of early adolescent girls' reports of romantic and sexual behavior. *Archives of Sexual Behavior*, *32,* 513–521.

Jennings, T. E., Lucenko, B. A., Malow, R. M., & Devieux, J. G. (2002). Audio-CASI vs. interview method of administration of an HIV/STD risk of exposure screening instrument for teenagers. *International Journal of STD and AIDS*, *13,* 781–784.

Macalino, G. E., Celentano, D. D., Latkin, C., Strathdee, S. A., & Vlahov, D. (2002). Risk behaviors by audio computer-assisted self-interviews among HIV-seropositive and HIV-seronegative injection drug users. *AIDS Education and Prevention*, *14,* 367–378.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1,* 30–46.

McLaws, M. L., Oldenburg, B., Ross, M. W., & Cooper, D. A. (1990). Sexual behavior in AIDS-related research: Reliability and validity of recall and diary measures. *Journal of Sex Research*, *27,* 265–281.

Metzger, D. S., Koblin, B., Turner, C., Navaline, H., Valenti, F., Holte, S., et al. (2000). Randomized controlled trial of audio computer-assisted self-interviewing: Utility and acceptability in longitudinal studies. *American Journal of Epidemiology*, *152,* 99–106.

Meyer-Bahlburg, H. F. L., Ehrhardt, A. A., Exner, T. A., & Gruen, R. S. (1988). *Sexual Risk Behavior Assessment Schedule for Youths* (*SERBAS-Y-SH*). Unpublished measure, HIV Center for Clinical and Behavioral Studies, New York State Psychiatric Institute.

Meyer-Bahlburg, H. F. L., Ehrhardt, A. A., Exner, T. A., & Gruen, R. S. (1994). *Sexual Risk Behavior Assessment Schedule for Homosexual Youths* (*SERBAS-Y-HM*). Unpublished measure. (Available from H. F. L. Meyer-Bahlburg, HIV Center for Clinical and Behavioral Studies, New York State Psychiatric Institute, 1051 Riverside Drive, Unit 15, New York, NY 10032, meyerb@childpsych.columbia.edu.)

Newcomer, S., & Udry, J. R. (1988). Adolescents' honesty in a survey of sexual behavior. *Journal of Adolescent Research*, *3,* 419–423.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Rosario, M., Hunter, J., Maguen, S., Gwadz, M., & Smith, R. (2001). The coming-out process and its adaptational and health-related associations among gay, lesbian, and bisexual youths: Stipulation and exploration of a model. *American Journal of Community Psychology*, *29,* 133–160.

Rosario, M., Mahler, K., Hunter, J., & Gwadz, M. (1999). Understanding the unprotected sexual behaviors of gay, lesbian, and bisexual youths: An empirical test of the cognitive–environmental model. *Health Psychology*, *18,* 272–280.

Rosario, M., Meyer-Bahlburg, H. F. L., Hunter, J., Exner, T. M., Gwadz, M., & Keller, A. M. (1996). The psychosexual development of urban lesbian, gay, and bisexual youths. *Journal of Sex Research*, *33,* 113–126.

Rosario, M., Meyer-Bahlburg, H. F. L., Hunter, J., & Gwadz, M. (1999). Sexual risk behaviors of gay, lesbian, and bisexual youths in New York City: Prevalence and correlates. *AIDS Education and Prevention*, *11,* 476–496.

Rosario, M., Schrimshaw, E. W., & Hunter, J. (2004). Ethnic/racial differences in the coming-out process of lesbian, gay, and bisexual youths: A comparison of sexual identity development over time. *Cultural Diversity and Ethnic Minority Psychology*, *10,* 215–228.

Saltzman, S. P., Stoddard, A. M., McCusker, J., Moon, M. W., & Mayer, K. H. (1987). Reliability of self-reported sexual behavior risk factors for HIV infection in homosexual men. *Public Health Reports*, *102,* 692–697.

Schroder, K. E. E., Carey, M. P., & Vanable, P. A. (2003). Methodological challenges in research on sexual risk behavior: II. Accuracy of self-reports. *Annals of Behavioral Medicine*, *26,* 104–123.

Siegel, D. M., Aten, M. J., & Roghmann, K. J. (1998). Self-reported honesty among middle and high school students responding to a sexual behavior questionnaire. *Journal of Adolescent Health*, *23,* 20–28.

Weinhardt, L. S., Forsyth, A. D., Carey, M. P., Jaworski, B. C., & Durant, L. E. (1998). Reliability and validity of self-report measures of HIV-related sexual behavior: Progress since 1990 and recommendations for research and practice. *Archives of Sexual Behavior*, *27,* 155–180.

Wiederman, M. W. (2002). Reliability and validity of measurement. In M. W. Wiederman & B. E. Whitley (Eds.), *Handbook for conducting research on human sexuality* (pp. 25–50). Mahwah, NJ: Erlbaum.

Williams, M. L., Freeman, R. C., Bowen, A. M., Zhao, Z., Elwood, W. N., Gordon, C., et al. (2000). A comparison of the reliability of self-reported drug use and sexual behaviors using computer-assisted versus face-to-face interviewing. *AIDS Education and Prevention*, *12,* 199–213.