Published in final edited form as:

Ann Behav Med. 2003 October ; 26(2): 104–123.

Methodological Challenges in Research on Sexual Risk Behavior:

II. Accuracy of Self-Reports

Kerstin E. E. Schroder, Michael P. Carey, and Peter A. Venable

Center for Health and Behavior Syracuse University

Abstract

Assessing sexual behavior with self-report is essential to research on a variety of health topics, including pregnancy and infertility, sexually transmitted infections, and sexual health and functioning. Recent methodological research has provided new insights regarding the accuracy of self-reports of sexual behavior. In this paper, we review these studies paying particular attention to a promising new development, namely, the use of computer-assisted assessments. The collection of sexual risk behavior data with computers has increased dramatically in recent years, but little is known about the accuracy of such assessments. We summarize the evidence, discuss methodological issues that arise in studies evaluating the accuracy of self-reports, and offer recommendations for future research.

Data regarding sexual behavior are typically obtained by self-report. These data are used to evaluate prevention programs targeting sexually transmitted infection and unintended pregnancy; to assess and treat a variety of sexual health problems; and to inform public policy decisions. Thus, it is critical that such self-reports yield the most precise, detailed, and accurate information. The purpose of this review is to discuss recent evidence regarding the accuracy¹ of self-reported sexual behavior. We identify factors that influence the accuracy of self-report, and integrate recent empirical evidence with earlier reviews by Catania et al. (1, 2) and Weinhardt et al. (3). We pay particular attention to technologically advanced modes of assessment, which have increased in popularity in the past decade. After reviewing the literature, we discuss methodological issues relevant to this research, draw conclusions about the state-of-the-science, and make recommendations for future research.

Factors Influencing the Accuracy of Self-Reported Sexual Risk Behavior

Research has identified a wide array of factors that appear to influence the accuracy of retrospective self-report of sexual behavior. These factors include the demands of the recall task and related memory error, and the social context of assessment, which can affect self-report bias. Our review focuses almost exclusively on retrospective assessment methods. Contemporaneous assessment methods (e.g., by diary) are rarely subject to accuracy evaluations but serve occasionally as a criterion against which retrospective self-reports are compared. Diaries and similar contemporaneous techniques are included in the general methodological discussion following this review. In reviewing the research on retrospective assessment modes, we use the acronyms listed in Table 1.

¹We use accuracy as a summary term that includes both reliability and validity. If self-reports of manifest sexual behaviors are the outcome of interest and no inference to a latent dimension is made, then questions of reliability and validity reduce to the question whether the behavior is measured accurately, that is, without error. If a behavioral measure is accurate, it is a reliable and valid indicator of the enacted behavior. For these reasons, we use the term accuracy whenever questions of reliability and validity are reduced to the precise measurement of manifest behavior. We limit the use of the term "reliability" to a discussion of reliability coefficients such as temporal stability of self-reports, and related research methods (e.g., test-retest reliability studies), and we use the term "validity" only when sexual behavior serves as an indicator for a latent dimension.

Cognitive Demands of the Recall Task and Memory Error

Memory error is affected by at least three factors that determine the demands of recall: the length of the reference interval, the level of measurement, and the frequency of behavior being assessed. Further, accurate responding is affected by individual limitations such as literacy skills.

Length of the Assessment Interval—Common sense suggests that reports for briefer intervals should be more accurate. At the same time, longer reporting intervals are regarded as more representative, especially for the assessment of low-frequency behavior. To achieve a balance between accuracy and representativeness, many investigators have tended to use intermediate reference intervals (e.g., two to three months) to obtain detailed information about sexual risk behavior.

Investigators often justify a 2 or 3 month interval with self-administered questionnaires (SAQs) by referring to Kauth, St. Lawrence, and Kelly (4). Results from this study are often interpreted to suggest that sexual behaviors can be assessed reliably by SAQ for intervals as long as three months but that longer intervals lead to a decline in self-report accuracy. This interpretation needs to be challenged for methodological reasons. In this study, sexual practices of 61 gay, HIV-seropositive men were assessed for three overlapping retrospective reporting periods: past 2 weeks, 3 months, and 12 months. The reports for the three time frames were compared by extrapolating the reports for shorter time intervals to the longer time frames assuming behavioral stability (i.e., one-year reports were compared to the three-month reports multiplied by four and to the two-week reports multiplied by 26; two-week reports were multiplied by 6 for comparison with 3-month reports). For most of the behaviors, the extrapolated frequencies calculated from the shorter report intervals exceeded self-reports for the longer periods. The greatest difference emerged for unprotected anal intercourse: The 2-week retrospective reports indicated an average of 100.5 times per year and the 3-month reports an average of 29.4 times per year; the real average frequency reported for the past year was 9.8 times. However, the results of this study are inconclusive due to the absence of a true test-retest design, or, alternatively, a comparison between contemporary self-reports with retrospective assessments.

To our knowledge, only two recent studies have addressed this issue. In the first study, Carey, Carey, Maisto, Gordon, and Weinhardt (5) examined the test-retest reliability of the TLFB among 66 psychiatric outpatients, using a test-retest interval of one week. A variety of sexual behaviors were assessed, and retest reliability coefficients for the most recent month and for the past three months were computed. Intraclass correlations (ICCs) were consistently high for both time-frames, with values ranging from $r = .71$ to $.97$ for the one-month period and between $.73$ and $.95$ for the three-month retrospective recall. These results indicate that, with TLFB, three-month retrospective reports may be as stable as reports for a one-month reference interval. Other studies with repeated assessments of sexual behaviors have not referred to the same time period, and do not provide information about the accuracy of retrospective self-reports (6).

In the second study, Jaccard et al. (7), compared retrospective SAQ self-reports with weekly self-monitoring, which was done prospectively, and collected over a 1-year period in a sample of 285 single, heterosexual men and women, aged 18 to 35. Weekly self-administered questionnaires were collected by mail, whereas more detailed retrospective questionnaires that assessed sexual behaviors since the beginning of the study were mailed at months 1, 3, 6, and 12. Statistically significant differences between weekly reported unprotected intercourse and long-term retrospective self-reports of unprotected intercourse (tested by regression models assuming an intercept of 0 and a slope of 1) were found only in the 1-month and 6-month periods. No significant discrepancies emerged in the 3-month and 12-month retrospective self-reports of unprotected intercourse, supporting the assumption that self-reports may be reliable as well as reasonably representative at least for a time frame of 3 months. There are at least

two explanations for the finding that retrospective frequency self-reports for a one-month assessment interval was less accurate than those for the three-month interval. First, self-reporting errors receive stronger weights in short time periods because of the lower number of sexual occasions compared to longer time frames; in contrast, longer time frames are likely to be more representative of an individual's behavior. Second, respondents may have chosen a recall strategy that is optimal for longer time frames but error-prone for shorter time frames. According to Jaccard and Wan (8), the use of a semantic memory strategy (e.g., retrieving knowledge about one's habits) may facilitate accurate self-reports for longer time frames, whereas reliance on episodic memory by recalling and counting each single event may be more successful for short time periods. It is possible that the overall task (i.e., recall of behavior for up to one year) caused an inflexible use of semantic memory strategies, and thus, less accuracy for the shorter time frame.

Whether reference intervals of longer than 3 months yield accurate self-reports requires more systematic research. The results provided by the two aforementioned studies may not generalize to other populations, behaviors, and assessment modes. Future research on these questions is needed to examine the trade-off between the information gained with longer time frames and the probable loss in accuracy over time.

Level of Measurement—Catania et al. (1) questioned whether incidence reports should be preferred over frequency reports of sexual behaviors because they found, on average, higher test-retest reliability coefficients for the former. They concluded that use of incidence reports (dichotomies) or categorical measures will increase the reliability of self-reports.

However, a number of recent studies have challenged the notion that relative frequency measures are more reliable. For example, Jaccard et al. (7) reported that only 58% of those who claimed to be perfectly consistent condom users were actually as they described. The accuracy of incidence data is also challenged by recent findings showing that assessment modes elicit different incidence reports of risky and non-normative sexual behaviors (9–11).

In addition, evidence supportive of absolute frequency reports has been reported by several authors. For example, Weinhardt et al. (12) tested the reliability of the sexual behavior TLFB in a sample of college students. Using a three-month time frame and a test-retest interval of one week, they found that the TLFB reports of a variety of risk behaviors were highly stable, as indicated by ICCs between .79 and .99. Carey et al. (5) reported intraclass correlations (ICCs) between $\rho = .71$ and .97 for frequency reports among adult men and women (5). Durant and Carey (13) found consistently high ICCs for both SAQ (.84 to .96) and FTFI (.92 to .95), when these assessment modes were compared to diary records. Carballo-Diequez et al. (14) assessed sexual risk behavior by FTFI in a sample of 75 male couples and reported weight kappas between .83 and 1.00 for a retest interval of one week. McKinnon et al. (15) assessed sexual behaviors by FTFI for a reporting period of six months and reported retest reliability coefficients (ICCs) between .74 and .89.

In sum, recent studies indicate that the stability of sexual behavior frequency reports is at least as high as the stability of incidence reports. Given that count measures yield more detailed quantitative information (16), they seem preferable compared to less informative dichotomous self-reports. Questions about the incidence of an event may make it easier for respondents to deny. Weinhardt et al. (3) even recommended to “place the burden of denial on the participant” by requesting a frequency report instead of an answer to the question whether or not a person engaged in a behavior, thus suggesting to the respondent that there is nothing unusual or abnormal about the behavior in question.

Frequency of Target Behavior(s)—Several studies have documented that accuracy decreases as the frequency of the target behavior increases (9,13,17). Thus, people who have to report only a small number of events tend to provide more accurate reports. For example, Downey et al. (17) assessed sexual behaviors in a sample of 87 men engaging in unprotected same-sex intercourse using a weekly diary over a period of three months. These records were subsequently compared to retrospective reports for the same interval. The authors found clear evidence for over-reporting of low-frequency sexual acts and underreporting of high-frequency sexual behaviors in the retrospective self-reports. Ramjee, Weber, and Morar (18) also found evidence for underreporting of sexual behaviors among highly sexually active, commercial sex workers.

Several kinds of memory error are affected by the frequency of events. Low-frequency events are more salient, and the salience of an event may induce telescoping, leading a person to believe (incorrectly) that an event occurred more recently and fell into the specified reference interval. Thus, the salience of low-frequency events may lead to over-reporting. In contrast, high frequency events are less salient and may lead to forgetting and underreporting. Telescoping and over-reporting of low-frequency events may be reduced by the simple technique of asking a question twice. Loftus, Klinger, Smith, and Fiedler (19) suggest that asking more than one question increases accuracy because it signals to the respondent that greater precision is desired. However, the two-time frame questioning method has not been evaluated in methodological studies of self-reported sexual behavior.

When recalling high-frequency events, respondents use a cognitive heuristic called rounding to provide rough estimates. For example, Tourangeau and Smith (9,10) found an increasing percentage of rounded values corresponding to the number of lifetime partners. Fifty percent of the participants with 10 to 19 partners reported multiples of five, and an equal percentage of participants with 20 to 29 partners reported multiples of ten. For 50 and more partners, nearly 90% of the participants rounded their number to multiples of 10 or 25. The same phenomenon is likely to occur for high-frequency sexual behaviors assessed with single-item questions. Rounding does not occur with event-level self-reports in TLFB, diary, or daily telephone reporting. A disproportionately high number of rounded values signal imprecision. Accuracy-enhancing techniques such as the two-time frame questioning method developed by Loftus et al. (19) or similar multiple item assessments may reduce rounding by communicating to the respondent that greater precision is desired. Further, introducing participants to recall-enhancing heuristics may reduce rough estimates and may improve accuracy of self-reports.

Literacy Skills and Task Demands—Respondents vary in their ability to provide self-report data accurately. It is now well-acknowledged that assessment methods differ in their demands for literacy. Because illiteracy is more likely in high-risk populations, this is a particular concern for HIV research. Assessments that rely on written material (e.g., SAQ) are particularly vulnerable to errors caused by lack of understanding and non-response. SAQs often yield missing values and logically inconsistent answers and subjects may not be able to follow complex instructions and skipping patterns. Assessments with low-literacy persons benefit from audio-presentation of the questions, either by a human interviewer or by audio-assisted methods (e.g., ACASI, T-ACASI). An advantage in interviewer-administered assessments is that uncertainty about the terms can be resolved immediately and errors due to misunderstanding can be eliminated, but such methods are less anonymous. Another method that does not necessarily involve written materials and can be applied in low-literacy populations is the diary. However, diaries are used rarely in sexual risk behavior research because of demands on the participants, the likelihood of high dropout rates, and possible reactivity.

Computerization and audio-assistance promise to reduce response error by eliminating missing values and excluding logically inconsistent response patterns. Audio-assistance combined with a SAQ (20) or a CASI (9,21) reduce cognitive demands, facilitate understanding of the questions, and can substitute for written materials. Several studies have examined the effects of audio-assistance on item non-response. For example, Turner et al. (21) assessed stigmatized sexual behaviors (e.g., same sex contacts) and substance use in two groups of men randomly assigned to ACASI and SAQ. ACASI elicited fewer item refusals and “don’t know” answers compared to SAQ. Boekeloo et al. (20) evaluated the benefits of audio-assistance by comparing SAQ with and without audio-assistance in a sample of 305 STD patients. In the audio-assisted condition, a lower number of missing values emerged (3%) compared to the written SAQ condition (6%).

Social Contexts and Motivational Biases

Motivational biases are affected by social contexts and moderated by person characteristics such as gender. We review recent research on both factors in turn.

The social context of assessment is likely to influence self-report accuracy. We would expect social processes such as social desirability and impression management to lead to over- and under-reporting in order to avoid shame or embarrassment, and to conform to perceived norms. Behaviors perceived as socially desirable are likely to be exaggerated, and behaviors that are perceived as socially disapproved are likely to be underreported. Because of social desirability and impression management, assessment strategies that elicit higher self-reports of sensitive behaviors (e.g., unprotected sex, anal intercourse, sex trading) are often assumed to be more accurate (9,10,20–24). Motivational bias has been examined most often in experimental studies comparing different assessment conditions and assessment modes. Most of these studies rely on the assumption that enhanced perceptions of anonymity, privacy, and credibility of an assessment increase accuracy (13).

Direct Interviewer Contact and Self-report Consistency—Studies reviewed in Weinhardt et al. (3) and Catania et al. (1), as well as more recent studies indicate that the stability of self-reports may be higher for interviewer-administered modes (FTFI, TLFB) than for self-administered assessment modes (SAQ, ACASI). However, an interpretation of this evidence is difficult because of the different time frames, populations, test-retest intervals, and settings. In a randomized controlled trial with drug using men and women, Williams et al. (25) examined the retest reliability of both a FTFI and ACASI in. Among men, the average retest reliability for ACASI was weak for number of partners ($r=.36$), and drug-using sex partners ($r=.54$) but strong for the frequency of vaginal sex ($r=.93$). For the same questions, collected with FTFI, retest correlations ranged from .88 and .90. Among women, both methods elicited stable self-reports. SAQ and ACASI may diminish the need for consistent self-reporting and occasionally lead to weak test-retest correlations, even for shorter reporting intervals (1,25–28). Research is needed to determine whether stronger stability coefficients in interviewer-administered assessments can be interpreted as improved accuracy of the self-reports. It may be that the increased stability in interviewer-administered assessments is a consequence of conscious attempts to report consistently rather than better recall or more truthful responding. Also, interviewer administration may offer more cues for episodic recall of previous answers to the same questions.

Recent Research on Anonymity, Confidentiality, Privacy—Three recent studies have examined the effects of motivational bias by manipulating anonymity, privacy, or the belief that responses can be evaluated by objective data. Tourangeau, Smith, and Rasinski (29) performed a Bogus Pipeline experiment, in which respondents were made to believe that inaccurate self-reports could be discovered by a lie detector. Adult volunteers from the

community ($n=120$) were randomly assigned to the Bogus Pipeline and the control condition. Among five incidence items on sexual behavior, one item (oral sex) was reported by a higher percentage of participants in the Bogus Pipeline condition (52% vs. 32%). Further, Bogus Pipeline participants tended to report a higher number of lifetime partners than those in the control condition ($M=13$ vs. $M=7$).

Gerbert et al. (30) tested the effects of different degrees of confidentiality on the accuracy of self-reports among medical patients ($n=1952$). Patients were randomly assigned to two conditions in which (a) both the researcher and their physician would review their responses or (b) responses would be seen by the researcher only. Finding no differences between conditions, the authors suggested that enhanced confidentiality does not affect responding. However, the manipulation of “confidentiality” was rather weak. Patients' self-reports may not be affected by “additional” disclosure of their answers to their physician, because patients are likely to tell their physicians more, not less, than they tell a researcher, in order to receive appropriate medical care. Also, the analyses were restricted to dichotomized scale scores only. An analysis of quantitative indicators of HIV risk behavior might have provided a more sensitive test.

Durant (31) examined the differential effects of anonymity versus confidentiality on responses in 358 college students. In the confidential condition, participants provided personal information (name, date of birth) but were assured of confidentiality; participants in the anonymous condition did not provide any personal information. Durant (31) also assessed perceived question threat and analyzed the relationship between perceived threat and data quality as reflected by incidence and frequency reports, dropouts, and item non-response. In the confidential condition, students reported lower frequencies for the 12 most threatening questions on health risk behaviors. Further, all 29 dropouts occurred in the confidential condition, and the number of non-responses was greater in this group. This study provides strong evidence for the value of anonymity when collecting sexual behavior self-reports.

Taken together, these studies confirm the role of motivational bias in self-reports of sexual behaviors. They also outline how research on motivational bias could proceed. Methodological research could be refined by directly assessing question threat (31) or perceived norms, in order to determine whether differential responding caused by assessment conditions or assessment modes may be attributed to motivational bias.

Mode of Assessment—Recent research has also focused on the effects of computerized assessment modes on sexual behavior self-reports. In general, methods requiring face-to-face communication with a human interviewer (TLFB, FTFI, CAPI) afford less privacy and anonymity, and can be expected to enhance motivational bias compared to self-administered assessment modes (i.e., SAQ, CASI, ACASI). Further, the technological interface of computerized self-interviewing (CASI, ACASI) may enhance the credibility of a study. In theory, effects of social desirability and impression management should be reduced in nonsocial, credible assessment settings; as a consequence, self-reports should be more accurate.

We identified 14 studies conducted between 1990 and 2000 that tested the effects of computerized assessments on sexual behavior self-reports (see Table 3). All 14 studies are experimental investigations of mode effects in which participants were randomly assigned to different assessment modes. With only a few exceptions (25,32), sexual behavior self-reports were studied as additional outcomes only, and only in the form of incidence reports or dichotomous measures. Twelve studies compared computer-assisted to traditional assessment methods, and two studies (9,23) focused on computerized methods only. We organize these studies into those (a) involving comparisons with FTFI/TI, (b) studies involving comparisons

with SAQ, and (c) studies testing the additional effects of audio-assistance in computerized assessments.

Computer-assisted self-interviewing versus FTFI/TI: Six of the studies listed in Table 3 compared FTFI or TI reports to computer-assisted self-interviewing with or without audio-assistance (CASI/ACASI; (11,22,25,33–36). In five of the six studies, computerized assessments elicited higher reports of sensitive behaviors and lower reports of socially approved behaviors. For example, Des Jarlais et al. (22) found a higher percentage of participants in the CASI condition reported engaging in homosexual relationships and sex trading, and a lower percentage reported consistent condom use with primary and casual partners. Locke et al. (33,34) conducted a randomized crossover trial with 272 blood donors who were assessed with both FTFI and CASI in random order. In this study, 12 blood donors with HIV risk factors were detected with CASI; none was detected by FTFI. With audio-assisted CASI, Metzger et al., (35) found more reports of HIV positive sex partners and of unprotected receptive anal sex in a sample of 1,974 gay men and 903 intravenous drug (ID) users, compared to FTFI. Robinson and West (36) compared CASI reports about symptoms and sexual history to the records from the physician's interview and found significantly more symptom reports with CASI. Turner et al. (11) compared a telephone-ACASI condition to telephone interviewing with a human interviewer. For 7 out of 15 sensitive questions on sexual behaviors, a higher incidence rate emerged in the computer-administered condition. In responding to the computer, more respondents acknowledged engaging in anal sex over their lifetime and during the past 6 months, limited sexual experience, unstable sexual relationships, and never discussing sex with their most recent partner.

The studies reviewed so far confound self-administration and computerization and do not allow clear-cut inferences. Only one study (32) included all four conditions produced by crossing both factors (i.e., traditional versus computer-assisted modes, and interviewer- versus self-administration). With a sample of over 1000 women, only self-administration had an effect on self-reports: Women responding to SAQ and CASI reported more partners for three assessment intervals (past year, past 5 years, lifetime), and (marginally) more STDs, compared to FTFI and CAPI. In a comparison sample of 100 men, self-administration had similar effects (higher averages) on reports of the numbers of partners for the past 1 and 5 years; however, men reported more lifetime partners in the human interviewer conditions (FTFI/CAPI).

One study does not fit this overall pattern of results. Williams et al. (25) examined test-retest results in a four-group crossover trial with 392 drug users. The four conditions were (a) test-retest with FTFI, (b) test-retest with Audio-CASI, (c) ACASI followed by FTFI, and (d) FTFI followed by ACASI. In this study, mode effects were weak and limited mostly to males. Men reported more sexual partners in FTFI than in ACASI, which was true for both within-group comparisons (i.e., ACASI versus FTFI in Conditions C and D) and between group comparisons (i.e., Condition A compared to B). Similarly, women reported more drug-using sex partners in FTFI than ACASI assessments, in both within-and between group tests. The results of this study contrast with findings from other studies and may be specific to the population of drug users who may perceive computerized assessments as more threatening because of an association with criminal files and investigations, thus enhancing self-protective response bias.

In sum, most studies demonstrate that computerized assessments lead to higher (and, purportedly, more accurate) self-reports of non-normative and risky sexual behaviors compared to FTFI. The results of Tourangeau, Rasinski et al. (32) suggest that these effects are due to self-administration rather than to computerization. This conclusion is supported by findings summarized in earlier reviews indicating that SAQ leads to higher self-reports of critical sexual behaviors than FTFI (1,2).

Computer-assisted self-interviewing versus SAQ: Six studies compared computerized assessments to SAQ (21,30,36–39). Paperny et al. (37) compared self-reports of 799 adolescents who were randomly assigned to anonymous CASI, non-anonymous CASI, and SAQ assessments. In both CASI groups, self-reports were higher compared to the SAQ group. Robinson and West (36) assigned 69 patients of a genito-urinary clinic randomly to SAQ and CASI and found higher symptom reports with CASI. Turner et al. (21,39) analyzed differences between SAQ and Audio-CASI, using data from the National Survey of Adolescent Males. Both studies indicate that ACASI enhances self-report of sensitive same-sex sexual behaviors. For heterosexual contacts, few differences between SAQ and ACASI were found. However, a higher percentage of men in the ACASI condition admitted having been drunk or high during heterosexual intercourse and having had sex with an ID user. The studies reported by Turner et al. (21,39) are inconclusive with regard to the reason for higher self-reports in ACASI because they confound computerization with audio-assistance.

Two studies used both CASI and ACASI in addition to SAQ. O'Reilly et al. (38) tested the acceptance of CASI and ACASI in a sample of 40 average to weak readers. Although CASI and ACASI were judged superior to SAQ in likeability, privacy, and honesty of responding, actual response rates did not differ. In a large sample of primary care patients, Gerbert et al. (30) compared five assessments modes, including SAQ, FTFI, and three CASI modes (i.e. CASI alone, with audio-, and video-assistance). Relative to SAQ and FTFI, computerization led to higher incidence rates for risky behaviors. No differences emerged between the three computer conditions (nor between FTFI and SAQ). In sum, studies comparing SAQ with computer-assisted self-interviewing methods indicate that computerization leads to higher self-reports of sexual risk behaviors; evidence for additional effects of audio-assistance was not found.

Computerized assessment with and without human interviewer: Two studies addressed the question of whether the effects of self-administration that are typical for FTFI-SAQ and FTFI-CASI comparisons hold for computerized assessments with and without human interviewer as well. Tourangeau and Smith (9) randomly assigned 365 participants to CASI, ACASI, and CAPI. The strongest effects were found for audio-assistance (ACASI). Compared to CASI and CAPI, ACASI subjects reported a higher number of lifetime sex partners, and a higher percentage of participants reported oral and anal sex. Gribble et al. (23) conducted computerized telephone interviews with a sample of 2881 people, who were randomly assigned to telephone-ACASI versus computer-assisted telephone interview by a human interviewer. In telephone-ACASI, more respondents reported having received or given drugs or money for sex. However, the internal validity of this study is questionable because 38% dropped out of the telephone-ACASI condition.

Summary: Most studies indicate that self-administration leads to higher reports of sensitive behaviors. This is the most consistent pattern in the results and corroborates self-administration effects in SAQ-FTFI comparisons reported in earlier studies and reviews. The effects of computerization and audio-assistance are less consistent, often ambiguous, and more difficult to interpret. Computerization effects may be masked in the presence of self-administration effects, which would explain the weak effects in FTFI versus CASI/ACASI comparisons (32). In SAQ versus CASI comparisons, computerization was clearly associated with higher self-reports of sexual risk behaviors. Evidence for the additional effects of audio-assistance over computerization is ambiguous because most studies confound these factors.

Gender and Motivational Bias—Social norms for sexual behavior differ by gender and may cause gender-specific response biases. Numerous studies have found higher self-reports of sexual behaviors from men and lower reports from women (1,2,29,40). Gender differences are most apparent in men and women's reports of the number of heterosexual partners.

Tourangeau and Smith (9) and Turner et al. (40) have pointed out that, in a closed population, systematic gender differences in the number of heterosexual partners is a mathematical impossibility, meaning that for unbiased self-reports, the number of heterosexual partners reported by men and women should converge. This contradicts typical findings that men report substantially higher number of partners than women (2,9,40). The most plausible explanation would be that the number of heterosexual partners is over-reported by men and under-reported by women (9). In consequence, the systematic gender difference that occurs in assessments of sexual risk behavior should raise doubts on the validity of the self-reports.

Several recent studies have used gender differences in sexual behavior self-reports as indicators of response bias. Tourangeau and Smith (9) analyzed the number of heterosexual partners reported by male and female respondents who were randomly assigned to two self-administered (CASI, ACASI) and one interviewer-supported computerized assessment modes (CAPI). In all three conditions, men reported a higher number of heterosexual partners than women, but this gender-specific difference varied by mode. Compared to interviewer-assistance (CAPI), both self-administered assessment modes (CASI and ACASI) reduced the discrepancy between men and women's self-reports. This may indicate a reduced bias in self-administered assessments. However, these results have to be interpreted cautiously because the interaction between assessment mode and gender was not statistically significant.

Results from the experiment discussed earlier (29) indicate that the Bogus-Pipeline condition affected men and women to the same degree, that is, both genders reported a higher number of partners when they were made to believe that the validity of their answers could be determined by a lie detector. This contradicts the idea of a gender-specific motivational bias. If a gender-specific response bias had been operating, participants in the Bogus-Pipeline condition would have shown a reduction of gender differences instead of a general increase in the average number of partners reported.

The persistent gender difference in the number of partners reported indicates some kind of error. However, this does not necessarily imply gender-specific response bias. Alternatively, gender differences might be a result of a self-selection bias affecting the composition of the samples. Men may be more inclined to participate if they are highly sexually active and decline participation if they "have nothing to report", whereas women may show the opposite pattern, being less likely to participate if they are sexually "over-active" and have sex with many partners. This could explain why neither high-motivation manipulations (29) nor anonymous, private, and self-administered computer-assessments (9,32) eliminate gender differences in the number of heterosexual partners reported. The results presented so far are inconclusive with regard to the question of whether gender differences in self-reports of sexual activity are due to response bias and, thus, question the validity of the self-reports, or whether the self-reports of men and women are valid but differ because of gender-specific self-selection bias. Both biases are likely to be operative.

Durant (31) provides more conclusive evidence with regard to gender-specific response bias. This study included 13 items on sexual behaviors and assessed perceived question threat for each item. Perceived threat for sexual behavior questions was higher for women (ranging from $\underline{M}$ =2.1 to $\underline{M}$ =4.5) than for men ($\underline{M}$ =1.7 to $\underline{M}$ =3.7), indicating a higher question sensitivity among females. The strongest gender difference in threat ratings emerged for frequency items on masturbation, number of sexual partners, oral sex with an opposite sex partner, and unprotected sex. Of further interest is the finding that women perceived greater question threat across a range of health behaviors. This indicates that gender-specific response bias may not be limited to sexual behaviors but may also apply in studies on drug use, alcohol consumption, and other illegal, norm-deviant, or negatively sanctioned behaviors. The findings of Durant

(31) provides strong evidence for the existence for gender-specific response motivations that may contribute to the gender differences observed in many studies.

One recent study, however, did not find the typical gender-specific response bias. Williams et al. (25) analyzed the number of partners and the frequency of vaginal sex among drug-users and found more sexual activity among women. This may be a population-specific finding because female drug-users may be more likely than men to meet costs for living and drugs by sex trading.

When gender differences do emerge, it remains unclear which gender provides the more accurate reports. Research on this question is scarce, mostly restricted to a separate estimation of test-retest reliability coefficients for each gender, and usually without testing for significance (5,12,25). Williams et al. (25) found that women drug users' self-reports of sexual partners may be more stable over time. However, this result may not apply to populations other than drug-using men and women and other sexual behavior self-reports beyond the number of partners.

In sum, most studies are inconclusive with regard to the question of whether gender differences in self-reports are a result of response bias (and thus, inaccuracy in self-reports) or a gender-specific self-selection bias. Further, little is known about the magnitude of response bias among men and women and the question concerning which gender delivers the more accurate reports. More research is needed to further evaluate the degree, causes, mode effects, and other moderators of gender-specific inaccuracies in sexual behavior self-reports.

Conclusions

1. Research evaluating the accuracy of retrospective self-reports remains insufficient.
2. Recent methodological studies provide new insights regarding the determinants of accuracy of sexual behavior self-reports. Promising are recent results showing that frequency data can be reported consistently across different sexual behaviors and for time frames of at least three months.
3. The quality of recall differs by assessment mode. Event-level assessments by TLFB are less prone to over-reporting and rounding but more likely to lead to under-reporting due to forgetting of events compared to single item assessments that request a summary report.
4. Interviewer-administered assessment modes perform better in test-retest comparisons (and thus seem more reliable) but elicit lower self-reports of sensitive behaviors (which seems to indicate inaccurate responding). Further research on this pattern is needed.
5. Computerization and audio-assistance may reduce measurement error. Audio-presentation improves question understanding and reduces response error, item refusal, and "don't know" responses. Computerization may rule out inconsistent and "out-of-range" values and is expected to reduce motivational bias.
6. Recent studies provide direct evidence for the bias-reducing effects of anonymity.
7. Comparisons between computer-assisted and traditional assessment modes indicate that self-administration as well as computerization may reduce the socially desired responding. However, most of these studies have assessed sexual behavior only on incidence level. Further research is needed to determine whether the results of these studies can be generalized to frequency reports.

8. Gender differences are increasingly used to estimate the degree of response bias and to evaluate the bias-reducing effects of anonymity, self-administration, and computerization. However, little attention has been paid to the alternative explanation that gender-specific self-selection bias may be responsible for gender differences in reporting. Research is needed to test this hypothesis. An interesting approach is the simultaneous assessment of perceived question threat as an indicator of socially desirable response motivations.
9. Research on motivational bias has paid little attention to the impact of environmental normative cues on sexual behavior self-reports. Computerized assessment itself may work as a normative cue and may bias responses, as discussed later.

Critique and Discussion of Methodological Research

So far, we have summarized and critiqued recent methodological studies and identified needs for further investigations. In the following section, we discuss the methodological strategies for testing the accuracy of sexual behavior self-reports. Our discussion focuses on the most common and feasible methods for validating self-report measures.³ We first discuss test-retest studies assessing the stability of self-reports. Second, we comment on weaknesses of experimental studies comparing mode effects on responding, which reduce the possibility to draw valid conclusions on the accuracy of self-reports. Third, we discuss the use of diary methods (prospective self-monitoring) to validate retrospective self-reports.

Test-Retest Reliability Studies

Investigators routinely provide test-retest reliability coefficients in studies using self-report measures because a minimum requirement for the accuracy of self-reports is that they are replicable on an individual level. If repeated self-reports of the same behaviors, referring to the same time period, are not stable, then they cannot be regarded as accurate. We address two important issues in the evaluation of test-retest results: (a) The statistical evaluation of self-report stability, and (b) limitations of test-retest reliability as an indicator of accuracy.

Problems in the Statistical Evaluation of Reliability Coefficients—Two common problems emerge in the literature. First, the reliability coefficients reported are sometimes inappropriate. Second, reliability coefficients can be unreliable themselves, which is usually not taken into account in the evaluation of study results.

Reliability coefficients differ in content: In a review of methodological studies on the accuracy of self-reports, we found that four reliability coefficients were most often used in testing the stability of self-reports: Pearson correlations, Spearman correlations, Kappa (and weighted Kappa), and intraclass correlations (ICCs; see Table 2 and Table 4). Investigators must select among these indicators, which are not equivalent. For illustrative purposes, we consider an example: Six sexually active participants report 3, 4, 5, 6, 7, and 8 sexual events; at a second assessment, regarding the same time period and sexual events, the six respondents report the values 5, 6, 7, 8, 9 and 10 in the same rank order. Given these data, both the Pearson and Spearman correlations assume a value of 1, the ICC equals .56, the simple (unweighted) kappa takes a value below 0, weighted kappa equals .22 (with an equidistant weight distribution) or .59 (for weights that decrease proportional with increasing discrepancies). In this example, the reliability coefficients span a range from below 0 to 1. A wide span of values, derived from different reliability coefficients, is not such an unusual occurrence as one might

³We do not address research using biological markers of self-report validity, which are limited by the sample size requirements and sensitivity to self-reporting error (1,7,50). We also do not discuss partner reports as criterion of self-report accuracy because its limited feasibility in high-risk populations with multiple and inconsistent partnerships. This research paradigm and its evidence are discussed in Catania et al (1), Jaccard & Wan (8) and Jaccard et al. (7).

think. For example, Downey et al. (17) examined the retest reliability for an ensemble of ten behavioral items and reported Kappa as .37 and the Spearman correlation as .82.

The discrepancy between the reliability indicators occurs for two major reasons. First, Pearson and Spearman correlations evaluate the relative concordance (linear relationships, rank order) between measures, regardless of the numerical differences in the paired values. In contrast, Kappa and ICCs evaluate the degree of absolute agreement between self-reports, disregarding further distribution characteristics. Second, reliability coefficients define “agreement” differently, which is the reason for discrepancies between unweighted kappa and ICC.

From our point of view, testing the stability of self-reports for the same behavioral events during the same time period requires indicators of absolute agreement, that is, that the same or similar frequencies be reported both times. In our example, we would not consider the self-reports to be accurate and absolutely reliable, as indicated by Spearman and Pearson correlations, because of the discrepancies in the absolute scores. Linear dependence and rank order do not satisfy the need for a measure that captures the true frequency of behavioral events. Thus, for test-retest studies that assess the same behavior at two occasions for the same reference period, ICC and kappa coefficients are the most appropriate indicators of reliability, whereas Pearson or Spearman correlations are imprecise. (The same is true for diary-recall comparisons, which are discussed later.) In contrast, correlation coefficients are more appropriate if self-reports of different time periods are compared and absolute agreement of the scores cannot be expected.

Compared to ICCs and weighted kappa, unweighted kappa seems too conservative as an indicator for behavioral frequency data. Kappa measures the agreement between repeated self-reports, controlled for chance agreement. Because absolute agreement is likely to decrease with an increasing range of values and more frequent target behaviors, the dichotomous criterion of agreement used by kappa is too rigid, in particular in the upper end of the distribution. ICC and weighted kappa are better able to capture the degree of agreement in terms of the numerical discrepancy between first and second self-report. For count measures, ICCs may be most appropriate. However, investigators would do well to report all suitable coefficients for a comprehensive evaluation of an assessment instrument (e.g., Downey et al.; (17)).

Reliability coefficients can be unreliable: There are two main factors that affect the reliability of the reliability coefficients: (a) The effective sample size, and (b) extreme outliers in the bivariate distribution.

First, kappa coefficients and ICCs are unreliable to the degree that the variance in the respective item is determined by a small number of cases. This is true regardless of the total sample size and the magnitude of the resulting coefficient. A simple example may illustrate this problem. In a repeated assessment with 100 people, only two subjects report an occasion of unprotected intercourse with an intravenous drug user during a specified 2-week period. If both subjects report the single event both times, the agreement between the two assessments would be perfect. If Subject A reports this event only at Time 1 and not at Time 2, and Subject B vice versa, the ICC will be near zero. If Subject A reports consistently one event both times but Subject B only at Time 1, the ICC will assume a value of .66. Because, in this example, two people only determine the “effective sample size” (i.e., the number of people with scores > 0), the reliability of the reliability coefficient is questionable, regardless of its magnitude.

As a consequence, the most reasonable approach in evaluating ICC and Kappa coefficients would be to take the effective sample size into account, that is, the number of people reporting frequencies greater than zero in one or both assessments. If the effective sample size is small, reliability coefficients should be interpreted cautiously, regardless of their magnitude. The

wisest approach would be to decide that, due to the small sample of people engaging in this behavior, the reliability of this part of a measure must await a larger sample that can provide more reliable and representative results.

Second, reliability coefficients are affected strongly by outliers in the bivariate distribution of first and second assessment. For example, Carey et al. (5) examined the retest reliability of the TLFB among psychiatric patients. For the frequency of unprotected vaginal intercourse during the past month, an extreme outlier occurred in the bivariate distribution of the repeated self-reports, indicating that a single male participant provided extremely discrepant accounts of his behavior in first and second assessment. If this outlier was removed from the analyses, the ICC for condom-protected vaginal sex improved from .47 to .95 in the male subsample and from .52 to .97 in the total sample. A single participant with extremely discrepant self-reports at first and second assessment may distort the magnitude of test-retest reliability coefficients that are otherwise strong for the rest of the sample.

Recommendations: To address the difficulties in the statistical evaluation of reliability coefficients, we recommend four strategies:

- a. Use indicators of absolute agreement (i.e., ICC, weighted kappa) for the evaluation of test-retest reliability between two measures of sexual behavior frequency.
- b. Report at least two of the most commonly used indicators of reliability, such as the ICC and kappa. When using weighted kappa, provide information about the definition of the weights. If kappa coefficients with self-defined weights are used, present these in addition to more common indices such as kappa with equidistant weights.
- c. Present reliability coefficients for different sexual behaviors together with the "effective sample size" so that readers can determine the reliability of this information.
- d. Evaluate the influence exerted by outliers. In test-retest reliability studies, extreme outliers in the bivariate distribution indicate that a single person was either unable or unwilling to provide a consistent account of his or her behavior. In this case, the unreliability can be attributed to the individual and should not distort the evaluation of the retest-reliability of an instrument. However, the recommendation to remove extreme outliers should not be misunderstood or misused as a general strategy to artificially enhance statistical results. The removal of outliers should be reported and justified, ideally by a figure showing the bivariate distribution (5).

Evaluation of Test-Retest Reliability Does Not Suffice—Test-retest reliability coefficients are indicators of stability. The potential to elicit consistent retrospective self-reports is a necessary precondition, but it does not guarantee the accuracy of the data. This is because an unknown part of the agreement between first and second assessment may be determined by auto-correlated error.

Over-estimation bias in the magnitude of test-retest reliability coefficients has to be taken into account for two reasons. First, because both the first and second assessments have to refer to the same time period, test-retest intervals must be kept brief, and respondents may be tempted to recall their previous report rather than their behavior per se. Such a strategy reduces the demands on memory and also satisfies participants' need for consistency. Second, there is evidence for stable individual bias in sexual behavior self-reports. For example, Coxon (41) compared diary records with retrospective self-reports of sexual behaviors. He found strong individual differences in self-report bias, indicating that people are likely to consistently under- or over-estimate their behavior in retrospect. As a consequence, agreement between first and second assessments in retest reliability studies is only partially determined by the true score,

and the magnitude of the reliability coefficient is, in part, a result of an artifact caused by repeated error.

We conclude that retest reliability studies are an important and valuable source of information about the stability of self-reports but do not provide sufficient evidence for the accuracy of the data. Test-retest coefficients may over-estimate the agreement between the true scores and the observed scores and need to be interpreted with caution. However, lack of agreement between repeated assessments, as expressed in low retest coefficients, is a clear indicator of inaccuracy. Retest reliability studies should be regarded a necessary step in the validation of a self-reports instrument, with the main goal to demonstrate that an important precondition for accurate measurement is met. Whereas research with SAQ or the TLFB often provides adequate information concerning retest-reliability, methodological research on computerized assessments has failed to provide basic evidence for the reliability of the data gathered with these methods (see Table 3)².

Experiments Comparing Retrospective Assessment Modes

In order to draw valid conclusions from experimental research on the accuracy of sexual behavior self-reports, the internal validity of the experimental study must be sound. In this regard, experiments on mode effects have two methodological limitations. First, they rely on assumptions that may not always be valid. Second, self-selection bias and attrition may threaten the internal validity of the results.

Basic Assumptions—Experiments comparing retrospective assessment methods seek to evaluate the convergent validity and relative accuracy of these methods. Explicitly or implicitly, they employ a set of assumptions in order to decide on the superiority of one method over another (1,2,7,9,40): (a) Higher incidence and frequency reports of non-normative and risky sexual behaviors indicate a higher degree of accuracy; (b) privacy, anonymity, and credibility of an assessment reduce bias and lead to more accurate results; (c) gender-specific norms affect perceived social desirability of sexual behaviors and cause gender-specific response bias; and (d) gender differences in self-reports of sexual behaviors are due to response bias and indicate inaccuracy of the data.

One shortcoming of experimental research on assessment mode effects is that these assumptions are not systematically tested. We already challenged the idea that gender differences can be interpreted clearly and exclusively as indicators of inaccurate responding. Further, one might challenge the idea that enhanced credibility leads to more accurate self-reports. Although the credibility of an assessment may have beneficial effects on accuracy, it may, alternatively, bias responses towards over-reporting of non-normative and risky sexual behaviors. The desire to be a "good subject" may be misunderstood as "having something to report." In this case, inferring enhanced accuracy from higher self-reports may not be warranted. This critique applies in particular to the conclusions drawn from higher responding to sensitive questions in computerized assessments. In addition, computerization may alter perceived norms and change response bias instead of merely enhancing credibility and accuracy. Interacting with a computer may be associated with computer games, the liberal norms and attitudes of a younger generation, and the "coolness" of being risky, and may induce less cautious (but not necessarily more accurate) responding.

Sampling Error – Threat to Internal and External Validity—In contrast to retest studies or diary-recall comparisons, experimental evaluations depend on successful randomization and

²The only exception we could identify in the literature is the study of Williams et al. (25) who report test-retest coefficients for computerized assessment modes.

are vulnerable to distortion by self-selection and attrition, which may not occur at random. For example, people high in social anxiety may refuse a FTFI but may be willing to respond to a computer. Although systematic dropout has no direct effect on the accuracy of the data (i.e., does not affect responses on an individual level), it invalidates the preconditions that allow inferences to be drawn from the results.

For example, Gribble et al. (23) tested the hypothesis that T-ACASI would enhance self-reports of drug use and sex trading compared to telephone assessments conducted by a human interviewer. Using a household probability sample of 2,881 telephone interviews, 697 men were randomly assigned to T-ACASI. However, in the T-ACASI condition, 83 men refused or were unable to be interviewed by T-ACASI, 90 participants discontinued in the demographic part, and 47 discontinued during the sensitive sections of the interview. Thus, only 437 men (63%) completed the T-ACASI interview. From the 137 who discontinued the T-ACASI, 96 continued with the human interviewer. The severe dropout in the T-ACASI condition limits any clear conclusions from the results. Although the authors found support for their hypothesis that T-ACASI elicits higher self-reports of sensitive behaviors, this inference is not warranted because dropout in T-ACASI may not have occurred at random.

In sum, experimental mode comparisons testing the convergent validity of different assessment modes provide valuable information about relative tendencies towards over- and under-reporting, but they do not suffice as a means of evaluating the accuracy of the data, and they cannot be used to decide on the superiority of one method over another.

Self-Monitoring Data as Substitute for the Missing Gold Standard

Results from methodological studies on sexual behavior self-reports are often difficult to interpret because there is no "gold standard" to which the retrospective self-reports can be compared (1,3). The closest approximation to a gold standard are diaries (or similar contemporaneous techniques) because they minimize the demands on memory and promise most accurate results (1,13,27,41–43). In the following sections we discuss pros and cons of the diary as a substitute for the missing gold standard and evaluate its contribution to methodological research.

An obvious advantage of the diary is the short time period between behavior and recording. Because memory error is unlikely to occur, one important source of inaccuracy in retrospective self-reports is minimized. Further, the diary is often assumed to be more private and credible, and it can be arranged to optimize confidentiality (if not anonymity) characteristics, which should reduce barriers to honest reporting (13,41). The diary requires only minimal language, reading, or writing skills and can be used with low-literacy populations. Daily recordings require a minimum of time and allow the respondent to decide when and where to complete the measure. Further, the diary can provide event-level data and detailed information about the co-occurrence of events.

If accepted as the standard of comparison, diary records offer the unique opportunity to estimate error variance and true variance in the retrospective self-reports, not (as usual) by correlation and regression, but by direct comparison between "true" score (the diary record) and "observed" score (the retrospective report). This comparison corresponds to the definition of reliability by classical test theory but it is rarely of practical utility. Under these conditions, diary – recall comparisons can provide highly valuable, unique information, as exemplified by the following six methodological benefits.

First, the diary may clarify whether higher self-reports in SAQ relative to FTFI indicate over-reporting with SAQ, under-reporting in FTFI, or both—Table 4 lists 13 studies that used diary measures as an accuracy criterion for retrospective self-reports. Eight

of these 13 studies compared diary records with SAQ, and 2 studies examined differences between FTFI and diary records. For example, Reading (44) compared FTFI against diary records with 68 male volunteers and found that FTFI assessments yield under-estimates of sexual behaviors in absolute terms and not only relative to SAQ.

The studies comparing SAQ self-reports with diary records provide, in sum, mixed results for over- and under-reporting. Three studies found over-reporting of sexual behaviors with SAQ (41,42,45), whereas two studies found under-reporting in SAQ relative to the diary (18,46). Two studies reported over- as well as under-reporting with SAQ (7,17), one study did not find significant differences between SAQ and diary records (13), and one study is inconclusive because diary and SAQ reports did not refer to the same time interval (27). Thus, with regard to SAQ, there appears to be no obvious trend towards either under- or over-reporting. Sample differences as well as item content may affect average tendencies to over- or underreport. The three studies in which over-reporting occurred used students, adolescents, and gay men. Studies in which SAQ reports were lower than the diary records were conducted with female sex workers and adolescent STD patients. Additional diary studies are needed to evaluate the direction and degree of error that can be expected with FTFI and SAQ assessments in different populations.

Second, comparisons to diary records can provide information about the frequency and absolute magnitude of over- and under-reporting—Coxon (41) examined the frequency of insertive and receptive anal and oral sex, condom use, and masturbation among gay men; he found, on average, higher self-reports in SAQs compared to the diary records. Although the majority of error occurred in form of over-reporting (55%), he also found a considerable amount of under-reporting (25%). Unprotected anal intercourse was associated with the highest amount of error, in both directions. Insertive anal sex without condoms was systematically under-reported, and receptive anal sex without condoms was systematically over-reported; thus, it appears that respondents attribute responsibility for the highest-risk behavior to their partner and bias their self-reports accordingly.

Third, diary comparisons allow identification of individual self-report tendencies in retrospective assessments—Coxon (41), for example, found systematic individual differences in the direction of self-report bias. Participants tended to either over-report or under-report their sexual activities. Further research could reveal person characteristics associated with consistent under- and over-reporting. In a recent study, Jaccard et al. (7) found significant relationships between level of education and self-report accuracy, with lower educated individuals showing more self-reporting error than higher educated individuals. However, in this study, age and gender were unrelated to the amount and direction of error. Of particular interest are further studies testing gender effects on heterosexual self-reports in order to determine whether the typical pattern of lower self-reports among females and higher self-reports among males result from gender-specific response bias (9) or are rather a result of gender-specific sampling error.

Fourth, diary comparisons resolve the problem of autocorrelated error and over-estimation bias of reliability coefficients that limit the utility of test-retest studies—This has important implications for the evaluation of test-retest results, as discussed earlier. Diary-recall comparisons can and should be used to estimate the amount of autocorrelated error that affects test-retest reliability coefficients of retrospective assessment instruments.

Fifth, diary comparisons can identify factors associated with the direction of error—For example, the knowledge that high-frequency events are more likely to be under-reported, and that low-frequency events tend to be over-reported, is gained from diary studies

that allow determining the frequency, direction, and absolute amount of error as a function of true behavior frequencies (17,18).

Finally, diary comparisons have the potential to provide information about the true superiority of one retrospective assessment mode over another—For example, Durant and Carey (13) used diary data as the accuracy standard for a comparison between FTFI and SAQ. In this sample of 185 female students, few differences in the self-reports emerged. However, this study may serve as a model for future comparisons between retrospective assessment modes.

These examples demonstrate results that are unique to studies using diaries (or other contemporary behavioral records) as standard. Given these unique contributions, the diary appears underused in studies testing the accuracy of retrospective data. However, it is worth noting that diaries have several disadvantages that may limit their feasibility and value. First, they require consistent self-monitoring over an extended time interval, which may enhance self-selection bias and dropout, and undermine the generalization of the results. Second, participants may not comply with instructions to complete the diary on a daily basis, thus re-inventing memory errors that the diary was supposed to eliminate. Third, diary measures may be reactive and may distort the results by affecting the behavior they are supposed to measure (47).

How do these disadvantages weigh relative to the benefits of diary measures? First, in contrast to experimental mode comparisons and applied research, self-selection bias may be a minor problem in methodological diary studies because of the focus on within-subject accuracy. In addition, there is no reason to expect a systematic dropout by memory ability, meaning that self-selection and dropout may not distort relationships between diary records and retrospective self-reports and, with regard to the generalizability of the sample results, may be limited to a “missing at random” pattern. Second, even if diary reports are provided on a weekly basis only, memory error will be quite limited compared to recall over several months. Third, although reactivity of diary assessments may limit their utility in intervention research, reactivity may be less important in methodological research, because of the focus on the accuracy of recall as expressed in the correspondence between enacted behavior (as indicated by the diary records) and retrospective self-report. In sum, the potential disadvantages of diary assessments seem to be of little consequence in methodological studies evaluating the accuracy of retrospective self-reports. A more serious limitation, however, is the possibility that diary assessments affect subsequent recall. Research on this question is needed to evaluate the generalizability of diary-recall comparisons to mere retrospective assessments.

Overall, however, we strongly recommend making greater use of the diary or similar contemporary assessment methods in future methodological research on the accuracy of retrospective sexual behavior self-reports. In deciding on a contemporary data collection method, the anonymity, privacy, convenience, and lower likelihood of incomplete data with the diary should be carefully weighed against alternatives, such as telephone-based (48) or internet-based assessments.

It is somewhat surprising that virtually no research has examined the relationship between diary data and technologically advanced, computerized assessment modes. In reviewing the literature, we found no overlap between studies testing responses to computerized assessment modes and diary studies. This is especially surprising given the increasing interest in both computer-administered assessment modes and ecological momentary assessment (EMA) in health and behavior research (49). Further research needs to address the degree to which computerized incidence reports and frequency measures of sexual behaviors deliver accurate

data, and whether these methods improve data quality compared to traditional assessment modes.

Summary and Conclusions

We reviewed methodological studies on the accuracy of retrospective sexual behavior self-reports and discussed three major methodological approaches, namely test-retest studies, experimental mode comparisons, and diary-recall comparisons. These study types deliver unique information and complement each other.

Test-retest studies provide indispensable information about the potential of an instrument to elicit stable retrospective self-reports. Because individual response tendencies are likely to lead to auto-correlated errors in repeated assessments, retest coefficients may over-estimate the accuracy of the data and should be interpreted with caution.

In test-retest studies – as well as diary-recall comparisons – the choice of the most appropriate reliability statistic is crucial for the validity of the conclusions they suggest. Repeated self-reports of the same behavioral events, referring to the same time period, require coefficients of absolute concordance (i.e., ICCs or Kappa). Pearson and Spearman correlations are more appropriate if two assessments refer to different time periods or provide data other than absolute frequencies of sexual behaviors. The reliability of reliability coefficients will vary depending on the effective sample size. This needs to be taken into account in interpreting reliability coefficients.

Experimental mode comparisons provide information about the relative performance of assessment modes in terms of incidence and frequency reports. They are vulnerable to systematic self-selection bias and often rely on assumptions that may not always be true. Caution is warranted regarding the crucial assumption that “higher self-reports of non-normative or risky sexual behaviors indicate higher accuracy.” This applies, in particular, to the finding that CASI or ACASI have a consistently higher likelihood to elicit more self-reports compared to other assessment modes, including SAQs. Computerized assessments may elicit over-reporting instead of higher accuracy. This possibility should be taken into account, especially in light of diary-recall comparisons showing that SAQ reports may lead occasionally to over-reporting but are, on average, not biased in any particular direction. Experimental mode comparisons miss independent evidence (i.e., other than the reported frequencies) for the accuracy of the data and the presumed motivational causes of differential responding. Needed are independent indicators of potential bias such as perceived question threat, perceived desirability of a behavior, or results from diary comparisons that inform about over- and under-reporting in retrospective self-reports.

Continued research on the accuracy of sexual behavior self-reports with computerized assessment modes is needed. We call for research that responds to the four most important yet unresolved issues regarding the accuracy of computerized assessments. First, test-retest studies are needed to establish the stability of self-reports obtained by computerized assessments. Second, research comparing computerized assessments with diary records or similar contemporary self-report methods is needed. Diary comparisons would provide the most valuable information about the accuracy of self-reports gathered with computerized surveys. Third, future methodological studies need to integrate a wide variety of sexual risk behaviors. In most studies, except for the number of partners, only dichotomous items on sexual behaviors were used. Fourth, little attention has been paid to the possibility that computerization may activate a different set of norms rather than removing any influence of motivational bias. Future research should address this question.

Diary records and equivalent concurrent assessment methods provide the most important source of information regarding the accuracy of retrospective self-reports. They can clarify the direction and magnitude of measurement error, the meaning of assessment mode differences in terms of absolute error, and individual differences in self-report bias. Further, they can provide information regarding the magnitude of auto-correlated error that may bias test-retest reliability coefficients. We encourage continued research using diaries to identify the conditions that enhance the accuracy of sexual behavior self-reports.

Some limitations need to be mentioned. Our review was restricted to the most common and feasible methodological approaches testing the validity of sexual behavior self-reports. We did not discuss research targeting biological markers of self-report validity or studies using partner reports as validity criterion. To date, both methods are used infrequently in research on computerized assessment modes and are of limited utility in evaluating accuracy of sexual behavior self-reports in most high-risk populations.³ However, both evaluation criteria can add to our knowledge about memory limits, self-reporting bias, and the external validity of retrospective measures when combined with computerized self-reports of sexual risk behavior.

Author Notes

All authors are with the Center for Health and Behavior, Syracuse University. Correspondence and request for reprints to Michael P. Carey, Center for Health and Behavior, 430 Huntington Hall, Syracuse University, Syracuse, NY 13244-2340; email: mpcarey@syr.edu. This work was supported by the National Institute of Mental Health (grants # R01-MH54929 and K02-MH01582).

References

1. Catania JA, Gibson DR, Chitwood DD, Coates TJ. Methodological problems in AIDS behavioral research: influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin* 1990;108:339–362. [PubMed: 2270232]
2. Catania JA, Binson D, Dolcini MM, et al. Risk factors for HIV and other sexually transmitted diseases and prevention practices among US heterosexual adults: changes from 1990 to 1992. *American Journal of Public Health* 1995;85:1492–1499. [PubMed: 7485660]
3. Weinhardt LS, Forsyth AD, Carey MP, Jaworski BC, Durant LE. 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* 1998;27:155–180. [PubMed: 9562899]
4. Kauth MR, St. Lawrence JS, Kelly JA. Reliability of retrospective assessments of sexual HIV risk behavior: A comparison of biweekly, three-month, and twelve-month self-reports. *AIDS Education & Prevention* 1991;3:207–214. [PubMed: 1834142]
5. Carey MP, Carey KB, Maisto SA, Gordon CM, Weinhardt LS. Assessing sexual risk behaviour with the Timeline Followback (TLFB) approach: Continued development and psychometric evaluation with psychiatric outpatients. *International Journal of STD and AIDS* 2001;12:365–375. [PubMed: 11368817]
6. McFarlane M, St Lawrence JS. Adolescents' recall of sexual behavior: consistency of self-report and effect of variations in recall duration. *Journal of Adolescent Health* 1999;25:199–206. [PubMed: 10475496]
7. Jaccard J, McDonald R, Wan CK, Dittus PJ, Quinlan S. The accuracy of self-reports of condom use and sexual behavior. *Journal of Applied Social Psychology*. In press.
8. Jaccard J, Wan CK. A paradigm for studying the accuracy of self-reports of risk behavior relevant to AIDS: Empirical perspectives on stability, recall bias, and transitory influences. *Journal of Applied Social Psychology* 1995;25:1831–1858.
9. Tourangeau R, Smith TW. Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly* 1996;60:275–304.
10. Tourangeau, R.; Smith, TW. Paper presented at the International Conference on Computer-Assisted Survey Information Collection. San Antonio, Texas: 1996 Dec 12. Collecting sensitive information with different modes of data collection.

11. Turner, CF.; Miller, HG.; Smith, TK.; Cooley, PC.; Rogers, SM. Telephone audio computer-assisted self-interviewing (T-ACASI) and survey measurements of sensitive behaviors: Preliminary results. In: R. Banks, JF.; Gerrard, L., editors. *Survey and Statistical Computing 1996*. Chesham Bucks, U.K.: Association for Survey Computing; 1996.
12. Weinhardt LS, Carey MP, Maisto SA, et al. Reliability of the timeline follow-back sexual behavior interview. *Annals of Behavioral Medicine* 1998;20:25–30. [PubMed: 9755348]
13. Durant LE, Carey MP. Self-administered questionnaires versus face-to-face interviews in assessing sexual behavior in young women. *Archives of Sexual Behavior* 2000;29:309–322. [PubMed: 10948721]
14. Carballo-Diequez A, Remien RH, Dolezal C, Wagner G. Reliability of sexual behavior self-reports in male couples of discordant HIV status. *Journal of Sex Research* 1999;36:152–158.
15. McKinnon K, Cournos F, Meyer-Bahlburg HF, et al. Reliability of sexual risk behavior interviews with psychiatric patients. *American Journal of Psychiatry* 1993;150:972–974. [PubMed: 8494081]
16. Schroder KEE, Carey MP, Venable PA. Methodological challenges in research on sexual risk behavior: I. Item content, scaling, and data analytical options. *Annals of Behavioral Medicine*. in press.
17. Downey L, Ryan R, Roffman R, Kulich M. How could I forget? Inaccurate memories of sexually intimate moments. *Journal of Sex Research* 1995;32:177–191.
18. Ramjee G, Weber AE, Morar NS. Recording sexual behavior: comparison of recall questionnaires with a coital diary. *Sexually Transmitted Diseases* 1999;26:374–380. [PubMed: 10458629]
19. Loftus EF, Klinger MR, Smith KD, Fiedler J. A tale of two questions: Benefits of asking more than one question. *Public Opinion Quarterly* 1990;54:330–345.
20. Boekeloo BO, Schiavo L, Rabin DL, et al. Self-reports of HIV risk factors by patients at a sexually transmitted disease clinic: audio vs written questionnaires. *American Journal of Public Health* 1994;84:754–760. [PubMed: 8179044]
21. Turner CF, Ku L, Rogers SM, et al. Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science* 1998;280:867–873. [PubMed: 9572724]
22. Des Jarlais DC, Paone D, Milliken J, et al. Audio-computer interviewing to measure risk behaviour for HIV among injecting drug users: a quasi-randomised trial. *Lancet* 1999;353:1657–1661. [PubMed: 10335785]
23. Gribble JN, Miller HG, Cooley PC, et al. The impact of T-ACASI interviewing on reported drug use among men who have sex with men. *Substance Use and Misuse* 2000;35:869–890. [PubMed: 10847215]
24. Paperny DMN. Computerized health assessment and education for adolescent HIV and STD prevention in health care settings and schools. *Health Education and Prevention* 1997;24:54–70.
25. Williams ML, Freeman RC, Bowen AM, et al. 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* 2000;12:199–213. [PubMed: 10926124]
26. Dowling-Guyer S, Johnson ME, Fisher DG, et al. Reliability of drug users' self-reported HIV risk behavior and validity of self-reported drug use. *Assessment* 1994;1:383–392.
27. McLaws ML, Oldenburg B, Ross MW, Cooper DA. Sexual Behavior in AIDS-related research: Reliability and validity of recall and diary measures. *Journal of Sex Research* 1990;27:265–281.
28. Saltzman SP, Stoddard AM, McCusker J, Moon MW, Mayer KH. Reliability of self-reported sexual behavior risk factors for HIV infection in homosexual men. *Public Health Reports* 1987;102:692–697. [PubMed: 3120234]
29. Tourangeau R, Smith TW, Rasinski KA. Motivation to report sensitive behaviors on surveys: Evidence from a bogus pipeline experiment. *Journal of Applied Social Psychology* 1997;27:209–222.
30. Gerbert B, Bronstone A, Pantilat S, et al. When asked, patients tell: disclosure of sensitive health-risk behaviors. *Medical Care* 1999;37:104–111. [PubMed: 10413398]
31. Durant, LE.; Carey, MP.; Schroder, KEE. The Effects of Anonymity and Respondent Characteristics on Self-Reports of Sensitive Behaviors. Syracuse: Syracuse University: Center for Health and Behavior; 2001. Unpublished manuscript.

32. Tourangeau R, Rasinski K, Jobe JB, Smith TW, Pratt WF. Sources of error in a survey on sexual behavior. *Journal of Official Statistics* 1997;13:341–367.
33. Locke SE, Kowaloff HB, Hoff RG, et al. Computer-based interview for screening blood donors for risk of HIV transmission. *JAMA* 1992;268:1301–1305. [PubMed: 1507376]
34. Locke SE, Kowaloff HB, Hoff RG, et al. Computer interview for screening blood donors for risk of HIV transmission. *MD Computing* 1994;11:26–32. [PubMed: 8145632]
35. Metzger DS, Koblin B, Turner C, et al. Randomized controlled trial of audio computer-assisted self-interviewing: utility and acceptability in longitudinal studies. HIVNET Vaccine Preparedness Study Protocol Team. *American Journal of Epidemiology* 2000;152:99–106. [PubMed: 10909945]
36. Robinson R, West R. A comparison of computer and questionnaire methods of history-taking in a genito-urinary clinic. *Psychology and Health* 1992;6:77–84.
37. Paperny DMN, Aono JY, Lehman RM, Hammar SL, Risser J. Computer-assisted detection and intervention in adolescent high-risk health behaviors. *Journal of Pediatrics* 1990;116:456–462. [PubMed: 2308041]
38. O'Reilly JM, Hubbard ML, Lessler JT, Biemer PP, Turner CF. Audio and video computer-assisted self-interviewing: Preliminary tests of new technologies for data collection. *Journal of Official Statistics* 1994;10:197–214.
39. Turner, CF.; Ku, L.; Sonenstein, FL.; Pleck, JH. Impact of ACASI on reporting of male-male sexual contacts: Preliminary results from the 1995 National Survey of Adolescent Males. In: Warnecke, RB., editor. *Health Survey Research Methods: Conference Proceedings*. DHHS Pub. No. (PHS). Hyattsville, MD: National Center for Health Statistics; 1996. p. 996-1013.
40. Turner, CF.; Miller, HG.; Rogers, SM. Survey measurement of sexual behavior: Problems and progress. In: Bancroft, J., editor. *Researching sexual behavior. Methodological issues*. Bloomington, IN: Indiana University Press; 1997.
41. Coxon AP. Parallel accounts? Discrepancies between self-report (diary) and recall (questionnaire) measures of the same sexual behaviour. *AIDS Care* 1999;11:221–234. [PubMed: 10474624]
42. Leigh BC, Gillmore MR, Morrison DM. Comparison of diary and retrospective measures for recording alcohol consumption and sexual activity. *Journal of Clinical Epidemiology* 1998;51:119–127. [PubMed: 9474072]
43. Morrison DM, Leigh BC, Gillmore MR. Daily data collection: A comparison of three methods. *Journal of Sex Research* 1999;36:76–81.
44. Reading AE. A comparison of the accuracy and reactivity of methods of monitoring male sexual behavior. *Journal of Behavioral Assessment* 1983;5:11–23.
45. Berk, R.; Abramson, PR.; Okami, P. Sexual activities as told in surveys. In: Abramson, PR.; Pinkerton, SD., editors. *Sexual Nature, Sexual Culture*. Chicago: The University of Chicago Press; 1995.
46. Fortenberry, JD.; Cecil, H.; Zimet, GD.; Orr, DP. Concordance between self-report questionnaires and coital diaries for sexual behaviors of adolescent women with sexually transmitted infections. In: Bancroft, J., editor. *Researching sexual behavior. Methodological issues*. Bloomington, IN: Indiana University Press; 1997.
47. Midanik LT. Validity of self-reported alcohol use: a literature review and assessment. *British Journal of Addiction* 1988;83:1019–1030. [PubMed: 3066418]
48. Searles JS, Perrine MW, Mundt JC, Helzer JE. Self-report of drinking using touch-tone telephone: extending the limits of reliable daily contact. *Journal of Studies on Alcohol* 1995;56:375–382. [PubMed: 7674671]
49. Stone AA, Schwartz JE, Neale JM, et al. A comparison of coping assessed by ecological momentary assessment and retrospective recall. *Journal of Personality and Social Psychology* 1998;74:1670–1680. [PubMed: 9654765]
50. Fishbein M, Pequegnat W. Evaluating AIDS prevention interventions using behavioral and biological outcome measures. *Sexually Transmitted Diseases* 2000;27:101–110. [PubMed: 10676977]
51. Tourangeau R, Jobe JB, Pratt WF, Rasinski K. Design and results of the Women's Health Study. *NIDA Research Monograph* 1997;167:344–365. [PubMed: 9243569]
52. Hoppe MJ, Gillmore MR, Valadez DL, et al. The relative costs and benefits of telephone interviews versus self-administered diaries for daily data collection. *Evaluation Review* 2000;24:102–116. [PubMed: 10747768]

53. Minnis AM, Padian NS. Reliability of adolescents' self-reported sexual behavior: A comparison of two diary methodologies. *Journal of Adolescent Health* 2001;28:394–403. [PubMed: 11336869]
54. Jeannin A, Konings E, Dubois-Arber F, Landert C, Van Melle G. Validity and reliability in reporting sexual partners and condom use in a Swiss population survey. *European Journal of Epidemiology* 1998;14:139–146. [PubMed: 9556172]
55. Kalichman SC, Kelly JA. Priming effects of HIV risk assessments on related perceptions and behavior: An experimental field study. *AIDS and Behavior* 1997;1:3–8.
56. Konings E, Bantebya G, Carael M, Bagenda D, Mertens T. Validating population surveys for the measurement of HIV/STD prevention indicators. *AIDS* 1995;9:375–382. [PubMed: 7794542]
57. McElrath K, Chitwood DD, Griffin DK, Comerford M. The consistency of self-reported HIV risk behavior among injection drug users. *American Journal of Public Health* 1994;84:1965–1970. [PubMed: 7998638]
58. Needle R, Fisher DG, Weatherby N, et al. Reliability of self-reported HIV risk behaviors of drug users. *Psychology of Addictive Behaviors* 1995;9:242–250.
59. Sohler N, Colson PW, Meyer-Bahlburg HF, Susser E. Reliability of self-reports about sexual risk behavior for HIV among homeless men with severe mental illness. *Psychiatric Services* 2000;51:814–816. [PubMed: 10828118]
60. Stone VE, Catania JA, Binson D. Measuring change in sexual behavior: Concordance between survey measures. *Journal of Sex Research* 1999;36:102–108.
61. Van Duynhoven YT, Nagelkerke MJ, Van De Laar MJ. Reliability of self-reported sexual histories: test-retest and interpartner comparison in a sexually transmitted diseases clinic. *Sexually Transmitted Diseases* 1999;26:33–42. [PubMed: 9918321]

Table 1**Acronyms**

ACASI	audio-and computer assisted self-interviewing
CAPI	computer-assisted personal interviewing (computerized data collection administered by an interviewer)
CASI	computer-assisted self-interviewing
FTFI	face-to-face interview
SAQ	self-administered questionnaire
T-ACASI	telephone- and computer-assisted self-interviewing (T-ACASI), and
TI	telephone interview with a human interviewer
TLFB	Timeline-Followback assessment (a calendar-aided personal interview method assessing sexual behaviors on event level and providing count measures of target behaviors)

Table 2
Test-retest studies on sexual behavior self-reports and their findings (1990–2000)

Author	Sample	Mode	Variables	Reliability tests	Analyses
Carey et al. (5)	Several samples of mentally ill patients	TLFB	Frequency of ins. & recept. vaginal, anal, & oral sex, condom use, 3 months	Interrater reliability: ICCs between .80 to 1.00, median & mean=.98; Retest-ICCs: .73 to .87; TLFB < frequencies than single item	ICCs, Wilcoxon signed-rank test
Dowling-Guyer et al. (26)	218 adult drug users	SAQ	No of partners & IDU partners, Freq. vag., oral, anal sex w. & wo. condoms, 1 month	No. of days having sex: r=.89; No of partner: r=.92; specific sexual behaviors: range from r=.07 to r=.75; Median TR =.75 11 of 17 retest coefficients > r=.7	Correlations
Jeannin et al. (54)	138 Swiss people age 17–22	Telephone survey	No. lifetime partners; Casual partners past 6 months; New steady partner; Condom use (Yes/no); Condom use past intercourse	Approximately 8–10 week test-retest interval; NOT the same reference periods!! For new steady partner, kappa=.51; For one or more partners lifetime: kappa=.91 Kappa higher for dichot. items (ever sex: .95) counts of partners: kappa=.51, Pearson r=.91	Kappa; Pearson correlation
Kalichman et al. (55)	267 female African American Women	SAQ, Overhead presentation, FTFI	8 of 16 items on freq. of vag., anal, and oral sex w. & wo. condom No. of partners	2-week test-retest reliability (stability coefficients), assessed in 2-week units of the retrospective reports; sex without condoms: .65; vag. sex without condoms: .78	Correlations
Kauth et al. (4)	61 gay men	SAQ	Condom use in oral & anal sex, number of partners	2-week, 3-month- and 12-month retrospective reports; Extrapolation of shorter periods to 1 year; Most striking result: average freq. of anal sex past year based on 2-week reports: 100.5; direct reports past year: 9.8 times!	Pearson correlations, descriptive analyses
Konings et al. (56)	N=75, Ugandan household survey	FTFI	Incidence sex w. casual partner, condom use last intercourse	Condom use: Kappa=.1; No of partners: k=.47	kappa
McElrath et al. (57)	ID users	FTFI	No of sexual partners, condom use	6-, 12-, 18 month retrospective reports kappa: .25 to .65, lower for longer intervals; incidence data more affected than counts	Kappas
McKinnon et al. (15)	27 SMI	FTFI	No of partners, Freq. of sex, proportion of vag. sex & condom use in vaginal sex	Test-retest correlations: .08 (withdrawal in anal sex) and .11 (withdrawal in anal rec.) to .95/.98 for unprotected ins. & rec. anal sex; Freq. condom use: .90/.80 for anal receptive and anal insertive sex;	Pearson correlations, ICCs, Kappas
McLaws et al. (27)	30 sexually active gay men	Diary, SAQ	Test-retest for SAQ: 48 to 72 hours; Recept. & ins. anal sex, w./w.o. condoms, withdrawal, oral sex, masturbation, No. of reg. & casual partners; 1 month	retrospective report period	Spearman Correlations
Needle et al. (58)	209 ID users in five cities	FTFI	No of days having sex; # of FTFI partners; # of IDU partners; freq. vaginal sex; condom use	Kappas ranged from .75 to 1.00	Kappa
Sohler et al. (59)	39 SMI homeless men of a shelter psychiatric program	FTFI	No. of partners, commercial, new partners, Freq. sex. male & female partners, vag. sex, ins./rec.anal sex, condom use	1-2-week test-retest reliability (ICCs) of the measures over a 6-month reporting period varied between .49 (anal sex with women) and .90 (number of episodes with men)	ICCs
Stone et al. (60)	962 heterosexual men and women of the AMEN cohort study	FTFI	freq. sex. risk behavior (vag. & anal sex) in the past year; No. of partners, sex, practices, condom use with each partner;	Difference scores of reported frequencies in Wave 1 and Wave 2 (1 year later) were computed and compared against direct reports of behavior change within the last year	kappa coefficients
Van Duynhoven et al. (61)	288 STD clinic attendees	FTFI, once physician, repeated by a nurse	Number of sexual partners, age at first intercourse	Kappa, exact agreement: .73 to .96, and 54% to 99%; doctor-interview: sign. lower number of partners and higher age at first intercourse Women's coefficients higher than men's	Kappa, exact agreement in %

Author	Sample	Mode	Variables	Reliability tests	Analyses
Weinhardt et al. (12)	110 college students	TLFB & single item method	vaginal, oral, anal protected and unprotected, insertive and receptive sex, Safer sex discussions, number of partners	Freq. & incidences, & retest ICCs were similar in SI and TLFB; ICCs: .86 to .97 for the TLFB; .81 to .95 for SI. No differences between genders	ICCs, bootstrap confidence intervals for the ICCs
Williams et al. (25)	392 subjects in a multi-site study	Audio- CASI, FTFI	Number of different sex partners, Drug-using sex partners, Times had vaginal sex;	Among men, FTFI reliability coefficients higher (.88 to .90) than ACASI coefficients (.36 to .93). For women, rather opposite result: .72-.78 for FTFI, .91 -.99 for ACASI	Correlations

Table 3
Experimental Studies Testing the Effects of Computerized and Audio-Assisted Assessment Methods on Data Quality

Authors	Sample	Outcome variables	Assessment Modes Compared	Refusals / Drop-outs	Missing values	Outcomes	Analyses
Des Jarlais et al., 1999 (22)	1,481 subjects in a syringe-exchange program	Dichotomous outcomes on sexual risk behaviors, risky IDU	ACASI, FTFI RCT	113 (7%) of 1,594 approached	no missings in ACASI; FTFI: <=1%	ACASI: higher incidence rates for sensitive behaviors (sex with same-sex partner, sex trading), and lower incidence rates for normative behaviors (always using condoms)	Logistic regressions
Gerbert et al., 1999 (30)	1952 medical care patients	10 dichotomous items on HIV risk behavior;	SAQ, FTFI Audio-ass. SAQ CASI, ACASI3 RCT	33 % of eligible patients declined	not reported	Enhanced confidentiality had NO effects on the reported risk factors. Mode effects on HIV risk questions: Audio-SAQ, CASI, and ACASI elicited higher percent at risk than written SAQ and FTFI T-ACASI; higher % reported having received or given drugs or money for sex In T-ACASI: Only 62% completed; 19% broke off during demographic part, 19% broke off during sensitive part	Chi ² tests
Gribble et al., 2000 (23)	2881 teleph. interviews; 697 randomly assigned to T-ACASI	Received or gave money or drugs for sex	T-ACASI Human telephone interviewer RCT	22% of 3694 eligible households declined;	not reported	SAQ and FTFI T-ACASI; higher % reported having received or given drugs or money for sex In T-ACASI: Only 62% completed; 19% broke off during demographic part, 19% broke off during sensitive part	Logistic regressions
Locke et al., 1992 (33, 34)	294 male & female blood donors	dichotomous: risky partners (HIV+, IDU), transfusion, male-male sex, HIV-related symptoms	CASI vs. FTFI / SAQ RCT & repeated measures crossover trial	10% of approached subjects declined;	22 (7.5%) of 294 omitted for missing data	CASI identified 12 blood donors with HIV risk factors or HIV related symptoms; None was identified by FTFI or SAQ; all participants were HIV negative; 61 % believed that blood donors would be more honest with computer, 12% vice versa	McNemar tests,
Mezger et al., 2000 (35)	1,974 gay men	dichotomous: HIV+ partners, unprot. anal receptive sex	ACASI FTFI RCT (longitud. study with 1-year FU)	ACASI: 18 of 528 refused or technical problems; 53 of 1446 in FTFI	did not occur for the two outcomes	ACASI elicited more reports of HIV positive partners; Tendencies to more reports of unprotected anal receptive intercourse, but this effect not significant CASI/ACASI were judged superior to SAQ; Among 14 sensitive questions on	Logistic regressions
O'Reilly et al., 1994 (38)	40 average and weak readers	dichotomous: sex risk behavior	SAQ CASI ACASI	Not reported	3 missing values for (A) CASI, 3 for SAQ	MANOVAC hi ² tests	MANOVAC hi ² tests

Authors	Sample	Outcome variables	Assessment Modes Compared	Refusals / Drop-outs	Missing values	Outcomes	Analyses
Papemy et al., 1990 (37)	810 adolescents	dichotomous: Having a lover, ever sexual intercourse	CASI w. or wo. disclosure to physician, vs. SAQ RCT	N=90; 35 in CASI, 6 in the share cond., 49 in SAQ cond.	not reported	sexual behaviors, only 2 led to higher reports in A/CASI Only significant difference: More subjects in the anonymous CASI condition reported having currently a lover compared to SAQ. No differences between other groups. Compared to physician's interview, both CASI and SAQ yielded significantly more symptoms; CASI elicited more symptom reports than SAQ	Chi ² tests
Robinson & West, 1992 (36)	69 patients at a genito-urinary clinic	symptoms, sexual history	SAQ, vs. CASI, compared to subsequent physician FTFI RCT	not reported	not reported		
Tourangeau et al., 1997 (51)	Household probability & clinic sample of 1,059 women, 100 men	Number of heterosexual partners; STDs, abortions	SAQ FTFI CASI CAPI RCT	Response rate: 55%; no differ. between conditions;	SAQ highest number of missing values (5.4%)	Women: higher # of partners, higher # of STDs reported in self-administered conditions (SAQ/CASI); no effects on reports of abortion; no effects of computerization; Men: higher # of partners Men reported more sex partners than women; ACASI: highest # of partners; higher reports of oral & anal sex ;Self-administration (CASI, ACASI vs. CAPI): reduced gender differences in # of partners; No mode effects on condom use.	ANOVAs Chi ² test
Tourangeau & Smith, 1996 (9)	365 subjects in the community	STDs, # of partners over 1 year, 5 years, lifetime; oral and anal sex, condom use	CAPI CASI ACASI RCT	43.2 of 643 subjects declined; no diff. between conditions	Not reported	ACASI elicited more reports of sexual contacts with a prostitute, male-male sex, sex and alcohol or drug use, and IDU partners; Little differences for prevalence of male-female sexual contacts.	ANOVAs Chi ² test
Turner et al., 1998 (21)	1729 men living in the community	Dichot. : Sex with prostitute, male-male sex, sex & alcohol / drug use, IDU partners	ACASI SAQ RCT	25% eligib. declined. 39 drop-outs; no diff. betw. conditions	less item refusals & don't know answers in ACASI		Logistic regressions log-linear modeling

Authors	Sample	Outcome variables	Assessment Modes Compared	Refusals / Drop-outs	Missing values	Outcomes	Analyses
Turner, Ku, et al., 1996 (39)	928 young men	Dichotomous: masturbation, ins.& receptive oral & anal sex	ACASI SAQ RCT	not reported (prelimin. results)	not reported	ACASI elicited higher reports of "ever been masturbated by another male", and "ever had insertive anal sex with another male"; any male-male sex	Logistic regressions
Turner, Miller et al., 1996 (11)	142 men & women age 18–49	Dichotomous: anal& oral sex, limited sexual experience, condom use	T-ACASI human telephone interviewer RCT	not reported (preliminary report)	no missing data in T-ACASI	T-ACASI: more reports of: anal sex over lifetime & 6-month, limited sex. activity past 6 months, instable relationships, & limited discussion of sex life. For 7 of 15 questions less social. desirable responding	logistic regressions
Williams et al., 2000 (25)	392 subjects in a multisite study	Number of partners, IDU-partners, frequency of vaginal sex	A-CASI FTFI, RCT rep. measures w. cross-over design	85 of 477 (18%) could not be interviewed	not reported	ACASI: Low test-retest correlation for # of sexual partners in the male sample; Mode-consistent test-retest correlations higher than cross-over correlations (ACASI first, then FTFI, and vice versa)	ANOVAs, t-tests, correlations

Note: RCT=Randomized Controlled Trial; WS-Design=Within-Subject (Repeated Measures) Design

Table 4
Studies Using the Diary as Standard of Comparison for Retrospective Self-reports

Authors	Sample	Outcome variables	Assessment Modes Compared	Refusals / Drop-outs	Missing values	Outcomes	Analyses
Berk et al., 1995 (45)	217 students	Number of Partners, vaginal & oral sex, fellatio	Diary SAQ WS-Design	Not reported	Not reported	higher reports of vaginal sex in the SAQ; this tendency is significantly higher for men than for women. In both assessments, women report more oral sex than men	t-tests comparing slopes
Coxon, 1999 (41)	86 gay men, volunteers recruited by newspaper ads	Insertive and receptive; anal sex, fellatio, w/w.o. condoms; masturbation.	Diary SAQ WS-Design	52%	Not reported	SAQ estimates were higher than diary; Individual difference analyses showed a high likelihood of being either an under- or overestimator. Discrepancy measure analyzed	Descriptive analyses; Pearson & Spearman correlations
Downey et al., 1995 (17)	87 gay and bisexual men taking part in a counseling program	Insertive and receptive oral and anal sex, w. & w.o. condoms	Diary SAQ WS	51% of eligible men refused	Not reported	Accuracy of SAQ low; Errors increased with frequency; Over-reporting of low-frequency behaviors & vice versa	Wilcoxon, Spearman correlations, Cohen's Kappa
Durant & Carey, 2000 (13)	185 female students	Vaginal, oral, and anal sex w. & w.o. barrier	Diary vs SAQ FTFI RCT & WS	3%	included in drop-outs	Analyses of discrepancy scores between SAQ or FTFI and diary; few differences; Vaginal sex with condoms was under-reported in FTFI; ICCs for FTFI: .92-.95; for SAQ: .84-.96	ICCs, bootstrapping; Wilcoxon
Fortenberry et al., 1997 (46)	48 women attending an STD prevention program	# of partners, each coital event, condom use,	Diary SAQ	not reported	not reported	Lower reports of sexual behaviors in SAQ (only	t-tests?

Authors	Sample	Outcome variables	Assessment Modes Compared	Refusals / Drop-outs	Missing values	Outcomes	Analyses
Hoppe et al. (52)	100 subjects from a study on drinking, drug use, & sex	alcohol / drug use & sex 10 items on sex. behavior s (among other health-related items)	Diary, daily telephone PI RCT & WS	Not reported	# missing days higher in telephone interview	about 62%; number of partners approximately equal No significant differences between conditions on reports of sexual behaviors (Number of days had sex, Proportions of sex occasions used condoms) Comparison between retrospective self-reports and contemporary (or "short-term" retrospective) weekly self-reports of unprotected intercourse were close for 3- and 12-month retrospective reports but differed significantly from weekly self-reports for the 1- and 6-month intervals. No gender differences in self-reporting bias were found. higher reports of sexual intercourse in SAQ in the adolescent sub-sample only; no differences in the proportion of relative condom use reported	t-tests
Jaccard et al., 2002 (7)	285 single, heterosexually active men & women	Frequency of vaginal and unprotected vaginal intercourse; Percent condom use	Postal SAQ (weekly); postal SAQ (at 1,3,6,12 months) WS-Design	30% refusals	3.3%, Treated w. missing value imputation	Regression analyses, robust regression, bootstrap estimates	
Leigh et al., 1998 (42)	79 young adults and adolescents, STD patients or students	Sexual intercourse, condom use, drinking / drug use before sex	Diary SAQ WS-Design	6 (7%) of subjects starting the study	see Morrison et al. (1999)	Correlations weighted kappa, t-tests, ANOVAs	

Authors	Sample	Outcome variables	Assessment Modes Compared	Refusals / Drop-outs	Missing values	Outcomes	Analyses
McLaws et al., 1990 (27)	30 sexually active men, mostly prostitutes	# of reg. and cas. partners, condom use/ withdrawal in anal ins. & receptive sex	PI-assisted SAQ, subsequent diary assessment WS-Design	Not reported	11 of 30 participants (37%) not completed the diary part	Comparisons between the first two weeks and the second two weeks of the diary; Test-retest correlations with SAQ (one-week interval) were between .35 and .92; Comparisons between diary and SAQ do not refer to the same time period! Comparison between the two diary methods, applied over a period of four weeks; Participants in the telephone interview condition reported decreased subjects in the diary condition reported increased condom use over time All 6 drop-outs among those starting the study emerged in the calling-in condition; Number of missing days differed by condition: 10% in calling-out, 10% in calling-in, <1% in diary; Underestimates of clients, condom use, vaginal and anal sex in weekly SAQ;	Spearman correlations
Minnis & Padian (2001) (53)	105 sexually active female adolescents	Use of contraceptive methods, including condom use	Daily diary & daily telephone interview Retrospect. Interviews RCT & WS-Design	11 of 118 eligible subjects refuses	3 drop-outs		GEE modeling binomial or Poisson distribution, Kappa
Morrison et al., 1999 (43)	see Leigh et al. (1998)	Vaginal and anal sex, condom use, drinking and drug use before sex	written diary vs. daily calling-out / daily calling-in RCT	155 of 491 letters undelivered 146 of 336 letters not returned	205 (9.3%) missing days. Item level: # of missings <1 / person		Chi ² tests, ANOVAs
Ramjee et al., 1999 (18)	79 female sex workers	Vaginal, anal, oral sex, type of partner (client/steady), condom use, condom break	Diary vs. weekly SAQ / weekly day-by-day recall WS-design	not reported	not reported		Rank correlations weighted kappa, paired t-tests

Authors	Sample	Outcome variables	Assessment Modes Compared	Refusals / Drop-outs	Missing values	Outcomes	Analyses
Reading, 1983 (44)	68 male volunteers	sexual desire, masturbation, satisfaction, sexual problems etc.	a) Diary & monthly FTFI b) 3 FTFI's, c) FTFI pre-post only	not reported	not reported	Underestimates of days worked, clients, anal sex, and sex with steady partner in weekly day-by-day recall. Correlations between diary cards and monthly FTFI ranged from .01 (morning erections) to .89 (sexual intercourse); Reports on diary cards higher; Under-reporting of sexual difficulties in FTFI	Pearson correlations

Note: RCT=Randomized Controlled Trial; WS=Within-Subject (Repeated Measures) Design