American Journal of Epidemiology
Published by the Johns Hopkins Bloomberg School of Public Health 2009.

Vol. 169, No. 11
DOI: 10.1093/aje/kwp049
Advance Access publication April 24, 2009

Practice of Epidemiology

Truth or Consequences: The Intertemporal Consistency of Adolescent Self-report on the Youth Risk Behavior Survey

Janet E. Rosenbaum

Initially submitted November 25, 2008; accepted for publication February 10, 2009.

Surveys are the primary information source about adolescents' health risk behaviors, but adolescents may not report their behaviors accurately. Survey data are used for formulating adolescent health policy, and inaccurate data can cause mistakes in policy creation and evaluation. The author used test-retest data from the Youth Risk Behavior Survey (United States, 2000) to compare adolescents' responses to 72 questions about their risk behaviors at a 2-week interval. Each question was evaluated for prevalence change and 3 measures of unreliability: inconsistency (retraction and apparent initiation), agreement measured as tetrachoric correlation, and estimated error due to inconsistency assessed with a Bayesian method. Results showed that adolescents report their sex, drug, alcohol, and tobacco histories more consistently than other risk behaviors in a 2-week period, opposite their tendency over longer intervals. Compared with other Youth Risk Behavior Survey topics, most sex, drug, alcohol, and tobacco items had stable prevalence estimates, higher average agreement, and lower estimated measurement error. Adolescents reported their weight control behaviors more unreliably than other behaviors, particularly problematic because of the increased investment in adolescent obesity research and reliance on annual surveys for surveillance and policy evaluation. Most weight control items had unstable prevalence estimates, lower average agreement, and greater estimated measurement error than other topics.

adolescent behavior; health behavior; reliability and validity; respondent error; risk-taking; tetrachoric correlation

Abbreviations: CDC, Centers for Disease Control and Prevention; IQR, interquartile range; TCC, tetrachoric correlation; YRBS, Youth Risk Behavior Survey.

Adolescents engage in risk behaviors such as smoking, illegal drug use, and early or unprotected sex that threaten their future health. Surveys are the primary source of information about many risk behaviors, and the only source for some behaviors (1). Federal, state, and local governments monitor risk behavior prevalence, set policy priorities, and promote legislation by using surveys, including the Youth Risk Behavior Survey (YRBS) (2, 3) and Monitoring the Future (4). The reliability of survey information is important for accurately measuring changes over time, determining geographic areas and demographics with a greater prevalence of risk behavior, and targeting and evaluating public health interventions. Inaccurate data can easily lead to mistakes in policy creation and evaluation.

Adolescents may report their risk behaviors inaccurately in ways that may threaten surveys' validity. When self-report has been compared with the ''gold standard,'' adolescents have been observed to overreport (5) or underreport (6–8) smoking, overreport height and underreport weight (9), both overreport and underreport arrest (10, 11), and misreport circumcision status (12). Adolescents also retract earlier-reported behaviors, initially reporting having engaged in a behavior and subsequently reporting having never engaged in the behavior. Logically, a retracted answer implies that the respondent lied in at least 1 of the 2 surveys, although the data cannot reveal which. Adolescents retract earlier reports of cigarette smoking (13–16), alcohol and illegal drug use (14, 15, 17–21), sexual intercourse (22–24), abortions (25) and pregnancy, virginity

Correspondence to Dr. Janet E. Rosenbaum, Department of Population, Family, and Reproductive Health, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, E-4008, Baltimore, MD 21215 (e-mail: janet@post.harvard.edu).

pledges, having a permanent tattoo, illegal driving, engaging in sex before age 13 years, and having pierced ears for boys (23).

Inconsistent reports may also carry information on adolescents' beliefs about the identity salience of their behaviors, including which behaviors they see as most central to their identities. Respondents are likely to inaccurately report behavior that conflicts with their identities or values (26, 27) and beliefs (26, 28, 29). For example, adults with greater levels of political interest are more likely to overreport voting (30–33), and respondents with more negative views of traffic violations and bankruptcy report fewer of their own traffic violations and bankruptcies (34). Adolescents' retraction of earlier-reported risk behaviors is most common for intimate, deviant, or illegal behaviors (20) and for experimental behaviors initially reported as infrequent (21, 22, 35). Adolescents seem to revise their pasts as their current behavior changes: their retrospective reports of substance use are more highly correlated with self-reported present use than with actual past use (36), adolescents who take a virginity pledge or become born-again Christians are more likely to retract earlier reports of having had sex, and adolescents who have sex or stop being born-again Christians are more likely to retract earlier reports of having taken a virginity pledge (24). Adolescents' self-image may influence them to be less likely to report weight control practices, both healthy and unhealthy practices, in interviews than self-administered surveys including exercise, diet, vomiting, and fasting (37); and to report drug use when they likely do not use drugs, because they report using fictitious drugs and many other drugs (38, 39).

This study compares adolescents' responses to 72 questions about their risk behaviors at a 2-week interval using methods that may overcome potential threats to validity in an earlier analysis of these data (40) and describe more aspects of survey response inconsistency. It assesses prevalence changes, measures unreliability in 3 ways, and identifies question properties that predict more reliable reporting.

MATERIALS AND METHODS

Data

The data were derived from contingency tables from a 2-week test-retest reliability study of the YRBS conducted in 2000 by the Centers for Disease Control and Prevention (CDC) (40). The YRBS was first developed by invited participants in a 1989 CDC workshop, was validated by the Questionnaire Design Research Laboratory at the National Center for Health Statistics with laboratory and field testing with high school students, was revised 3 times before it was first administered in 1991 (41), and was tested twice for reliability (40, 42).

The reliability study used a convenience cluster sample of classes from 61 schools in urban (48%), suburban (39%), and rural (13%) settings in 20 geographically dispersed US states plus the District of Columbia. On survey day, 77% of the students were present in class with a signed parental consent form. Of students completing the first survey, 89% completed the second survey. The CDC excluded questionnaires with fewer than 20 valid responses or with the same response option 15 times in a row, yielding a final sample of 4,619 students that overrepresented females, African Americans, grades 9 and 10, and ages 15–16 years and underrepresented whites, Latinos, grades 11 and 12, and ages 13–14 years (Table 1).

Students answered 97 questions from the YRBS in two 40-minute classroom periods between February and April 2000, at approximately a 2-week interval. To assure students' anonymity, the survey was administered by trained data collectors from Macro International (Washington, DC), and students alone had access to identification numbers used to link responses. The survey used a computer-scannable booklet with questions above answer choices to avoid off-by-1 errors. CDC dichotomized questions with multiple response categories into ''no risk'' and ''at risk.''

The reliability study was conducted for CDC's internal use. CDC policy is not to disseminate data collected for internal use, so the full data set is unavailable (N. D. Brener, CDC, personal communication, 2006). CDC published an analysis of these reliability data that included prevalence data at each survey administration (p1, p2) and Cohen's kappa (κ) for 72 of 97 questions (omitted items include contraception and substance abuse at last sex) rounded to 1 decimal place (40). These published data can be used to recover the contingency tables. The number of respondents who said ''yes'' at both surveys was $a = n/2\ [\kappa(p1 + p2 - 2p1p2) + p1p2]$, which has an error due to rounding in the original paper of no more than 1 respondent.

The survey questions include items on use of tobacco, alcohol, and illegal drugs; sexual intercourse; symptoms of depression and eating disorders; suicide; violence and weapons use; physical activity; and health-preserving behaviors such as wearing seatbelts, helmets, and sunblock and visiting a dentist and doctor. Questions were coded by possible predictors of inconsistent responses. Question topic was the primary predictor of interest, so questions were coded for whether they concerned deviant, illegal, or stigmatized behavior (43), including sexual intercourse; illegal drugs; alcohol and tobacco; perpetrating a violent crime; being a victim of a crime; and history of suicide, depression, and eating disorders. Other potential predictors were question time frame because memories of more recent events are more accurate (43, 44), and true change is more likely for questions about short time frames; readability, including question word count, number of response choices, whether the previous question concerned a different topic (whether the question was preceded by a transition sentence), concerned a different time frame, or had different answer choices; and whether the question was dichotomized from multiple response choices because dichotomization may artificially lower agreement because of loss of information.

The convenience sample was compared with the nationally representative sample in the YRBS by computing z scores of time 1 and 1999 YRBS (45). The 1999 YRBS questionnaire is available online (http://web.archive.org/web/19991128160242/www.cdc.gov/nccdphp/dash/yrbs/survey99.htm).

Table 1. Demographic Characteristics (%)[a] of Test-Retest Survey Respondents (in 2000) vs. a Nationally Representative 1999 YRBS Sample, United States[b]

	Sample (n = 4,619)	YRBS 1999 Sample (n = 15,349)
Gender		
Male	46.6	50.4
Female	53.4	49.6
School grade		
9	30.6	28.9
10	31.8	26.0
11	21.9	23.6
12	15.7	21.4
Race or ethnicity		
White, non-Hispanic	52.2	60.8
Black, non-Hispanic	31.4	14.1
Hispanic, any race	6.1	10.4
Other	10.3	14.7
Age, years		
≤13	0.1	1.6
14	12.4	17.4
15	28.9	24.0
16	28.5	24.5
17	21.2	22.3
≥18	8.9	10.3

Abbreviation: YRBS, Youth Risk Behavior Survey.
[a] Percentages may not add to 100 because of rounding.
[b] This table is based on information from Brener et al. (40) and Kann et al. (45).

Data analysis

Each question was evaluated for prevalence change and 3 measures of unreliability. These measures were inconsistency (retraction and apparent initiation), agreement measured as tetrachoric correlation (TCC), and estimated error due to inconsistency measured as a Bayesian estimate of the standard error due to inconsistent reporting.

First, prevalence change was assessed by the McNemar test. The earlier analysis of these data (40) compared 95% confidence intervals for prevalence constructed with sampling error under the assumption that the 2 observations were independent, which biases results toward finding no difference between groups since independent observations have a higher standard error than nonindependent, repeated observations from the same individuals.

Second, inconsistency was measured as absolute and relative retraction and as absolute and relative initiation. These measures provide easily interpretable means to compare observed inconsistency with inconsistency expected from chance. Absolute retraction is the proportion of the sample contradicting an earlier reported behavior: an affirmative answer followed by a negative. Relative retraction is the proportion of those who initially reported the behavior and subsequently retract their report: absolute retraction divided by wave 1 prevalence. Absolute apparent initiation is the proportion of the sample that appears to initiate the behavior between waves by reporting the behavior at wave 2 but not at wave 1. Finally, relative initiation is the proportion of wave 2 endorsers who did not report the behavior at wave 1: absolute initiation divided by wave 2 prevalence. Retraction and initiation depend on prevalence: absolute retraction and initiation are bounded from above by the prevalence of the risk behaviors; rare behaviors have more variable relative retraction and initiation because the denominator is small.

Third, agreement was measured by using TCC instead of the more common agreement measure kappa, used in the original analysis (40). TCC is constructed to be independent of prevalence, so rare and common behaviors may be compared on the same scale (46–51) and low agreement cannot be attributed to either low prevalence or prevalence change between waves. TCC can be interpreted as conventional correlation, with 0.0 chance agreement and 1.0 perfect agreement. TCC can also adjust for potential differences in response tendency by wave, such as if adolescents redefine risk behaviors on retest (26); TCC is high if the primary response tendency difference is a shift. TCC is computed with standard error by the maximum likelihood method in the R statistical package (52).

Predictors of agreement were found in 2 ways: comparing mean TCC by category and through linear regression. Past results suggest that adolescents are more likely to misreport sensitive or unusual behaviors (20, 24), so behavior category was considered the primary predictor of agreement, especially behaviors with higher levels of inconsistency in past research: sex and tobacco, alcohol, and drug use. The mean TCCs of the categories were compared by using the Tukey test for honest significant difference. The linear regression had outcome variable TCC, and the model was built beginning with question topic as predictors and by also considering time frame and the question characteristics described above. If agreement was due to memory or true change, agreement would be associated with time frame, tested in 2 ways: by including time frame as a predictor in regression on TCC and comparing TCC for the same risk behavior by time frame.

Fourth, error due to inconsistency was estimated as a standard error multiplier derived from a Bayesian simulation model (53–55), which is described in the Web Appendix (this supplementary material is posted on the *Journal*'s website (http://aje.oupjournals.org/)). Estimated error due to inconsistency is another method of quantifying the impact of inconsistency on adolescent risk behavior surveillance. Error is derived from a model that makes different assumptions than TCC, but both are independent of prevalence. Regressions were replicated by using this estimated error multiplier as an outcome variable.

RESULTS

Prevalence change

The prevalence of 41 of the 72 behaviors changed in a 2-week interval (Table 2), some in logically impossible directions. No change was expected between waves 1 and 2 in

respondents' reports of their behavior prior to age 13 years because all respondents were older than that, yet more respondents reported having used cigarettes ($P \leq 0.0001$) and marijuana ($P \leq 0.05$), and fewer reported sexual intercourse ($P \leq 0.0001$). No decrease was expected in reported lifetime prevalence, but fewer respondents reported having ever used cigarettes ($P \leq 0.0001$), alcohol ($P \leq 0.0001$), and marijuana ($P \leq 0.01$) and having had 4 or more lifetime sexual partners ($P \leq 0.01$). Prevalence change was not associated with any question characteristics in 2 logistic regressions and 1 linear regression.

Inconsistency

Even when prevalence does not change, inconsistency may be high. For example, the proportion of respondents reporting pregnancy history—having ever been pregnant or making another person pregnant—was about 8%–9% at both waves. Although prevalence did not change, 45.3% of those initially reporting pregnancy retracted their report 2 weeks later. Furthermore, 42.7% of pregnancies reported at wave 2 seem to have occurred in the 2 weeks between surveys because these pregnancies were not reported at wave 1.

Median relative retraction for the 72 questions was 27% (interquartile range (IQR) = 19.5–38.2); that is, on average 27% of those reporting a behavior at wave 1 denied the behavior at wave 2. Median relative initiation was 28% (IQR = 19.3–44.2); on average 28% of those reporting a behavior at wave 2 had not reported it at wave 1, as if it were initiated in the 2-week interval between surveys.

No retraction and moderate initiation were expected for the 15 items concerning whether respondents engaged in the behaviors in their lifetimes, but median relative retraction was 23.7% (IQR = 11.9–32.7) and median relative apparent initiation was 28.7% (IQR = 15.5–38.8). That is, almost a quarter of those reporting having ever engaged in a behavior at wave 1 denied the behavior at wave 2, and about a quarter of those reporting the behavior at wave 2 had not reported the behavior 2 weeks earlier at wave 1, as if they had initiated the behavior in the interim. No retraction or initiation was expected for items about behavior before age 13 years because all respondents were older than that; however, for the 4 items, 23.3% of respondents at median retracted and 27.7% of respondents at median apparently initiated. Variation regarding rare behaviors may be larger, but when analysis was restricted to the 13 of 19 lifetime and pre-age-13-years behaviors with a prevalence of 10%–90%, 18.4% of respondents retracted (IQR = 6.9–26.4) and 19.4% of respondents apparently initiated (IQR = 19.4–26.3).

Retraction regarding items about the past year only was expected for respondents who performed a behavior 50–52 weeks before wave 1 and not in the 2 weeks between waves. The behavior changes of such respondents would produce a relative retraction of 2/54 (3.7%) and relative initiation of 3.7%, assuming the behavior had a uniform distribution. Relative retraction and initiation for nearly all (17 of 18) questions about the past year were higher than the levels expected if changed reports were due to true change.

Weight control behaviors had the largest retraction rate. More than 20% of those initially reporting that they consider themselves overweight, are trying to lose weight, or exercise and diet to lose weight retracted these reports 2 weeks later; and more than 50% of those initially reporting that they fast, vomit, and take diet pills retracted these reports 2 weeks later. Apparent initiation of these behaviors was similarly high.

Agreement

Agreement, measured by TCC, was high and left skewed (median = 0.87, IQR = 0.80–0.92) (Table 2). The questions with the highest agreement (TCC = 0.99) involved having ever had sex and having used marijuana. Other questions in the top quartile of agreement (TCC > 0.92) included 3 of the 4 items on marijuana; 7 of the 13 items on smoking; having ever used alcohol, cocaine, and methamphetamines; and 2 of the 4 items on suicide. Questions in the bottom quartile of agreement (TCC < 0.80) included having been taught about AIDS or HIV infection in school (TCC = 0.45, an outlier), 6 of the 7 weight control items, and having seen a doctor when not sick. Agreement regarding the remaining weight control question, whether the respondent considers himself or herself to be overweight, was close to the bottom quartile (TCC = 0.82).

Average agreement (TCC) for the topics of tobacco, alcohol, and drugs was significantly higher than for weight control and miscellaneous topics (doctor, dentist, sunscreen, and HIV education) when Tukey's honest significant difference was used. Agreement for the topic of depression was higher than for the miscellaneous topic and was marginally higher than for weight control/physical activity.

Agreement (TCC) for questions on sexual intercourse and on tobacco, alcohol, and illegal drug use was substantially higher than average, and agreement for questions on weight control was substantially lower than average (Table 3). Agreement was not associated with time frame, question length, or other topics and was marginally lower for questions for which the answer choices had been dichotomized from a multi-item scale ($P = 0.07$, not shown). As expected from its derivation, TCC was not associated with prevalence.

For risk behaviors asked about in multiple items, agreement varied by question time frame within the same risk behavior. Agreement was higher for longer time periods: higher for lifetime than for the past 30 days regarding all 6 risk behaviors for which both time frames were asked; higher for lifetime than for before age 13 years for all 4; higher for the past 30 days than for the past 30 days at school for all 3; higher for the past 30 days than for before age 13 years for all 4; and higher for before age 13 years than for the past 30 days at school for 2 of the 3 risk behaviors (Table 4). Additional discussion of the relative levels of agreement (TCC) between questions can be found in the Web Appendix.

The Bayesian simulation model estimated that unreliable data increased standard error at median by a factor of 3 (Table 2); that is, confidence intervals that account for measurement error due to respondents' inconsistent reporting

Table 2. Two-Week Response Consistency in the YRBS Reliability Study, 2000, United States ($n = 4{,}619$)[a]

	Prevalence[b]			Retraction[c]		Initiation[d]		TCC[e]	SE Multiplier[f]
	Wave 1	Wave 2	McNemar *P* Value	Absolute	Relative	Absolute	Relative		
Tobacco								0.96	2.8
Ever try smoking	65.8	63.9	≤0.0001	4.2	6.4	2.3	3.6	0.98	2.1
Smoke a pack per day	17.5	17.1		2.6	14.6	2.2	12.7	0.97	2
Smoked in the past month	27.2	27.5		3.4	12.7	3.8	13.6	0.96	2
Ever smoke regularly	17.7	19	≤0.001	2.4	13.5	3.7	19.4	0.96	2.4
Chewing tobacco	6.6	6.4		1.9	28.2	1.7	26	0.94	2.7
Bought cigarettes	6.4	7.2	≤0.01	1.5	24.1	2.3	32.5	0.93	2.9
Smoke at school	9.7	9.1	≤0.01	2.7	28.1	2.1	23.5	0.93	2.6
Tried to quit smoking	18.4	16.7	≤0.0001	5.2	28	3.4	20.6	0.92	3
Smoked before age 13 years	21.4	23.7	≤0.0001	3.9	18.4	6.2	26.3	0.91	3
Chewed tobacco at school	3.9	3.9		1.5	38.1	1.5	38.1	0.9	3
Ever smoke cigars	12.2	11.8		4.5	36.5	4.1	34.3	0.86	3.1
Bought tobacco and was carded	6.8	8.2	≤0.001	2.6	37.9	4	48.6	0.83	4.1
No usual cigarette brand	1.6	1.5		1	63.5	0.9	60.9	0.78	3.4
Alcohol								0.94	2.9
Ever drink alcohol	76.1	72.5	≤0.0001	5.3	6.9	1.7	2.3	0.97	2.8
Drank alcohol in the past month	41.1	39.9	≤0.01	7.6	18.5	6.4	16.1	0.9	2.5
Binge drink in the past month	23.9	23.7		6	25	5.8	24.4	0.89	2.5
Drank alcohol before age 13 years	28.9	29.9	≤0.01	6.6	22.8	7.6	25.3	0.87	2.7
Drank alcohol at school	3.9	4.1		1.8	47.2	2	49.7	0.83	3.9
Illegal drugs								0.94	3
Ever use marijuana	42.8	41.7	≤0.01	3	7.1	2	4.7	0.99	1.7
Ever use cocaine	5.6	6.2	≤0.01	1.2	20.9	1.8	28.7	0.95	2.5
Marijuana use in the past month	22.6	22.1		4.4	19.5	3.9	17.7	0.94	2.2
Ever use methamphetamines	6.3	6.9	≤0.01	1.5	24.1	2.1	30.5	0.94	2.7
Marijuana before age 13 years	10.5	11.3	≤0.01	2.5	23.7	3.3	29.1	0.93	2.8
Ever use inhalants	11.3	10.6	≤0.01	3.6	31.6	2.9	27	0.91	2.9
Ever use heroin	1.9	3	≤0.0001	0.5	25	1.6	52.2	0.91	2.7
Ever inject an illegal drug	1.4	2	≤0.01	0.5	33.9	1.1	53.3	0.9	2.7
Use marijuana at school	5.5	5.3		2.2	39.8	2	37.6	0.88	3.3
Cocaine use in the past month	2.2	2.7	≤0.01	1	45.1	1.5	55.2	0.84	3.8
Ever use steroids	4	4.1		2.1	51.9	2.2	53.2	0.8	4.1
Inhalants use in the past month	2.9	3.5	≤0.01	1.5	51.5	2.1	59.9	0.79	4.1
Ever offered drugs at school	23	21.9	≤0.01	8.9	38.6	7.8	35.5	0.76	3.5
Sex								0.91	3.2
Ever had sexual intercourse	49.5	50.2	≤0.01	2	4.1	2.7	5.4	0.99	1.6
Sex in the past 3 months	32.9	35	≤0.0001	5.1	15.4	7.2	20.5	0.91	2.7
≥4 sex partners	19.1	17.6	≤0.01	7.1	37	5.6	31.6	0.82	3.3
Ever pregnant	8.6	8.2		3.9	45.3	3.5	42.7	0.81	3.7
Sex before age 13 years	18	14.8	≤0.0001	9.8	54.4	6.6	44.5	0.66	4.7
Traffic								0.93	3.8
Bike helmet, rarely	84.6	83.8	≤0.01	3.6	4.3	2.8	3.4	0.94	2.4
Motorbike helmet, rarely	37.8	46.8	≤0.0001	3.8	10.1	12.8	27.4	0.89	5.6
Seatbelt, rarely	15.7	19.6	≤0.0001	3.6	23.2	7.6	38.5	0.86	4.5
Drove after alcohol drinking	8.5	10.3	≤0.0001	2.8	32.3	4.6	44.1	0.85	3.9
Rode with a drinking driver	30.3	29.6		8.7	28.6	8	27	0.82	2.8

Table continues

Table 2. Continued

	Prevalence[b]			Retraction[c]		Initiation[d]		TCC[e]	SE Multiplier[f]
	Wave 1	Wave 2	McNemar *P* Value	Absolute	Relative	Absolute	Relative		
Suicide and depression								0.9	3
Considered suicide	17	16	≤0.01	4.1	23.8	3	18.9	0.94	2.6
Attempted suicide	8.4	8.5		2.1	24.5	2.2	25.5	0.94	2.5
Planned suicide	13	12.9		3.8	29.3	3.7	28.9	0.9	2.7
Injured in a suicide attempt	2.1	2.7	≤0.01	0.8	39.2	1.4	52.4	0.87	3.4
Felt sad/hopeless	28.2	24.1	≤0.0001	10.5	37.2	6.4	26.5	0.8	4.1
Violence								0.88	3.8
Forced sex	9.1	10.3	≤0.01	2.4	26.4	3.6	34.9	0.91	3.4
Carried weapon	15	13.3	≤0.0001	5	33.5	3.3	24.9	0.89	3.5
Physical fight	34.6	30.3	≤0.0001	9.2	26.7	4.9	16.2	0.89	3.8
Fight at school	13.1	12.4		4.3	32.9	3.6	29.1	0.89	2.9
Weapon at school	5.1	5.7	≤0.01	1.9	36.4	2.5	43.2	0.88	3.5
Carried a gun	4.2	4.4		1.9	45.9	2.1	48.3	0.84	3.6
Injured in a fight	2.9	4.4	≤0.0001	1.1	38.8	2.6	59.6	0.83	4.8
Injured by someone	9.1	9.9	≤0.01	3.6	39.4	4.4	44.3	0.82	3.6
Feel unsafe at school	5.5	5		3.1	57.1	2.6	52.8	0.76	4.3
Threatened by a weapon at school	7.3	5.9	≤0.001	4.4	59.8	3	50.2	0.73	4.2
Weight, physical activity								0.84	3.7
Physical education weekly	62.4	56.8	≤0.0001	6.5	10.4	0.9	1.5	0.98	2.6
Try to lose weight	22.7	26.1	≤0.0001	6	26.2	9.4	35.8	0.82	3.9
Consider self to be overweight	33.8	37.2	≤0.0001	7.9	23.3	11.3	30.3	0.8	3.6
Exercise to lose weight	58.6	53.9	≤0.0001	12.8	21.9	8.1	15	0.79	4
Sports team	54.6	53.3	≤0.01	11.5	21.1	10.2	19.2	0.77	2.5
Diet to lose weight	43.1	40.4	≤0.0001	12.7	29.6	10.1	24.9	0.75	3.5
Vomit to lose weight	4.9	5		2.8	56.2	2.9	57.1	0.74	3.7
Diet pills to lose weight	7.8	7.9		4.1	53.1	4.2	53.7	0.73	4
Sports injury	40.8	35.2	≤0.0001	15.3	37.5	9.7	27.6	0.69	4.7
Watched TV for <2 hours/day	62.4	63.2		12.1	19.3	12.9	20.3	0.68	3.4
Fasted to lose weight	18.4	15.3	≤0.0001	10	54.1	6.9	44.8	0.66	4.9
Physical education: exercise for ≥20 minutes	72.3	69	≤0.0001	13.9	19.2	10.6	15.3	0.63	4.2
Miscellaneous								0.71	3.6
Saw the dentist	66.5	63.4	≤0.0001	9.8	14.8	6.7	10.6	0.85	3.3
Used sunscreen rarely	66.6	66.7		8.6	12.9	8.7	13.1	0.83	2.7
Saw the doctor when not sick	58.9	58.1		12.4	21.1	11.6	20	0.72	3.1
Taught about human immunodeficiency virus	85	86.2	≤0.01	8.8	10.4	10	11.6	0.45	5.1

Abbreviation: YRBS, Youth Risk Behavior Survey.

[a] YRBS categories and items within categories are sorted according to decreasing agreement (tetrachoric correlation (TCC)).

[b] Comparison of prevalence at waves 1 and 2 is from the McNemar test.

[c] Absolute retraction is the proportion of all respondents reporting behavior at wave 1 and denying it at wave 2. Relative retraction is the proportion of respondents reporting the behavior at wave 1 who denied the behavior at wave 2.

[d] Absolute initiation is the proportion of all respondents newly reporting the behavior at wave 2, having not reported the behavior at wave 1. Relative initiation is the proportion of those who reported the behavior at wave 2 who apparently initiated the behavior and had not reported the behavior at wave 1.

[e] TCC measures average agreement between wave 1 and wave 2 responses, with 0.0 representing chance agreement and 1.0 perfect agreement.

[f] The standard error multiplier (SE) is estimated from the Bayesian model as the factor by which the usual standard error calculation underestimates total error, including inconsistency.

Table 3. Average TCC by Question Topic, YRBS, 2000, United States

Question Topic	Average TCC[a]	SE	*P* Value
Sexual intercourse	0.950	0.064	≤0.1
Tobacco	0.913	0.033	≤0.01
Alcohol	0.892	0.044	
Illegal drugs	0.882	0.031	≤0.1
Weight control	0.740	0.038	≤0.1
All questions	0.817	0.016	

Abbreviations: SE, standard error multiplier; TCC, tetrachoric correlation; YRBS, Youth Risk Behavior Survey.

[a] TCC measures agreement between wave 1 and wave 2 responses, with 0.0 representing chance agreement and 1.0 perfect agreement. Averages were found in linear regression. $R^2 = 0.32$.

would be at median 3 times as wide as usual confidence intervals. The questions with the lowest error concerned sexual intercourse (standard error multiplier = 1.6), marijuana use (standard error multiplier = 1.7), and smoking cigarettes (standard error multiplier = 2.0–2.1). The questions with the highest error were those related to fasting to lose weight in the past month (standard error multiplier = 4.9), ever being taught about HIV in school (standard error multiplier = 5.1), and rarely/never wearing a motorcycle helmet when riding a motorcycle in the past month (standard error multiplier = 5.6). As found in regressions using TCC as an outcome, sex, drug, alcohol, and tobacco items had lower error than other topics, and weight had higher error (not shown). As expected, the standard error multiplier was not associated with prevalence but was significantly associated with relative retraction.

DISCUSSION

In a 2-week period, adolescents' reports of their sex, drug, alcohol, and tobacco histories were more reliable than their reports of other behaviors; by contrast, in longer intervals, these behaviors were reported much less reliably than other behaviors. Most sex, drug, alcohol, and tobacco items had stable prevalence estimates, higher average agreement, and lower estimated error than other YRBS topics. In short periods, these behaviors may be reported consistently because the behaviors have high identity salience and few adolescents change identities in a 2-week period. In longer periods, these items may be reported less consistently because adolescents report past sex and substance use in accordance with their current identities (24, 26, 36), so adolescents who change their identities and habits will report inconsistently (20, 24).

The validity of adolescent weight control items is particularly critical because of increased investment in obesity research and reliance on annual surveys for surveillance, but adolescents report their weight control behaviors more unreliably than any other behavior. For most weight control items, compared with any other topic, prevalence estimates were unstable; average agreement was lower, the only category for which agreement was low for all questions; and estimated error was higher.

Adolescents change reporting of their past behaviors as their present behaviors change (24, 26, 36). If adolescents changed their weight control behaviors more frequently than the 1-month time frame of the weight control questions, low agreement on reports of weight control may be due to adolescents' reporting their most recent behavior rather than their past-month behavior. Low agreement may also be due to changed inhibitions about reporting weight control behaviors on retest, which would be consistent with past findings that adolescents underreport both healthy and unhealthy weight control behaviors, including vomiting to lose weight and dieting to lose weight, in interviews compared with self-administered surveys (37). However, no trend was evident regarding how inhibitions to report weight control might change on retest: more adolescents reported that they consider themselves overweight and are trying to lose weight, but fewer adolescents reported exercise, diet, and fasting to lose weight, and the proportion reporting vomiting or using pills did not change. The first explanation seems more likely: adolescents may change their weight control behaviors more frequently than a question about the past month can capture accurately. Questions about weight control practices may yield more accurate responses if phrased in terms of a more recent time period, such as the past week, as dietary intake questions are currently formatted, with repeated measures necessary for longer-term surveillance.

Table 4. Agreement (Tetrachoric Correlation) by Time Frame, YRBS, 2000, United States[a]

Item	Ever	Past 30 Days	Before Age 13 Years	Past 30 Days at School
Marijuana	0.99 (0.002)	0.94 (0.006)	0.93 (0.009)	0.88 (0.02)
Cigarettes	0.98 (0.003)	0.96 (0.004)	0.91 (0.008)	0.93 (0.008)
Alcohol	0.97 (0.003)	0.90 (0.01)	0.87 (0.01)	0.83 (0.02)
Sex[b]	0.99 (0.001)	0.91 (0.007)	0.66 (0.02)	
Inhalants	0.91 (0.01)	0.79 (0.03)		
Cocaine	0.95 (0.008)	0.84 (0.03)		

Abbreviation: YRBS, Youth Risk Behavior Survey.

[a] Standard errors, in parentheses, were computed by using the maximum likelihood estimator in the polychoric correlation library for the R statistical program (52).

[b] Sex is not reported for the past 30 days, but rather for the past 3 months.

Inconsistent responses increase measurement error in proportion to the level of inconsistency. No pattern was evident in the direction or magnitude of prevalence changes from attempted regressions, so prevalence changes may be another manifestation of this measurement error. The error regarding prevalence of an inconsistent behavior such as exercising to lose weight (error multiplier = 4.0) was underestimated by twice as much in magnitude as that for consistent behaviors such as smoking cigarettes (error multiplier = 2.0). This study does not advocate that confidence intervals be constructed to account for all measurement error including inconsistency. Researchers should nonetheless be aware of the limitations of their instruments, as survey experts have advocated (44, 56). For example, borderline-significant findings for items with higher estimated measurement error may be attributable to that error.

Limitations

The hypothesis advanced in this paper about high consistency in short intervals being due to the identity salience of these behaviors to adolescents is a post hoc explanation, but it is plausible because identity is thought to be related to inconsistency during long intervals. The identity salience hypothesis could be studied systematically by using the complete data to find associations between inconsistency and gender, grade, race/ethnicity, and age. This analysis was limited to contingency tables, however, because complete data are not available publicly. Because of a lack of access to full data, this study also could not determine whether inconsistency is a property of the individual, with some individuals more likely to be inconsistent, or the question, with inconsistency correlated among related questions, vital information for improving YRBS validity.

Dichotomization of questions with multilevel categorical responses may have artificially lowered agreement because of loss of information. With full data, agreement on these items could be measured by polychoric correlation, a generalized version of TCC (46, 47, 50, 51). In addition, not all questions were included in the original publication, such as those regarding contraception and substance use during sexual behavior (40).

Another limitation is that the geographically diverse convenience sample is not nationally representative. In addition, this sample is somewhat less likely to engage in risk behaviors compared with the nationally representative YRBS sample.

The Bayesian simulation model for estimating error due to inconsistency was underidentified: there are 3 degrees of freedom in the data to estimate 7 parameters, so many combinations of the 7 parameters could create the observed data, but the use of priors for 4 parameters—sensitivity and specificity for each of the 2 waves—restricted the problem. The estimates of all parameters were stable, so it can be concluded that the priors restricted the problem sufficiently that underidentification was not a major concern.

Comparison with earlier analysis

This study replicates some findings of the original analysis of these data, and it adds others. Brener et al. (40) conducted the original data collection rigorously, analyzed the data thoroughly, and explored some of the same issues as those discussed in this paper but, in a few instances, used ambiguous or inappropriate measures. As in the original analysis (40), this study found substance use and sex to be the most consistent topics, no statistically significant consistency difference across all questions by question time frame, and some instances in which inconsistency could be due to true change.

This study is distinct from the earlier analysis (40). It used a more appropriate test for prevalence change, used a less ambiguous measure of agreement so that low agreement could not be attributed to low prevalence, quantified the impact of inconsistency on measurement error, and found a lack of reliability regarding weight control questions and proposed a potential solution.

Conclusions

Adolescents reported sex and substance use consistently in a 2-week interval, but they reported weight control less consistently than any other risk behaviors. This inconsistency is especially problematic because adolescent obesity is a central public health issue and is potentially more dangerous to adolescents' future health than are other risk behaviors. Revising YRBS weight control questions to encompass a shorter time period may allow more accurate surveillance of adolescents' self-initiated weight control and physical activity. Future survey validity research could examine alternatives to current YRBS weight control questions. In the meantime, researchers should be aware of limitations of the current YRBS data, especially regarding weight control.

Survey report consistency may be connected to adolescents' identities. In short periods, adolescents present their sex and substance use consistently, but, in long periods, adolescents may change their social affiliations and these behaviors and thus report inconsistently.

ACKNOWLEDGMENTS

Author affiliations: Johns Hopkins University Center for Sexually Transmitted Diseases, Baltimore, Maryland (Janet E. Rosenbaum); and Department of Population, Family, and Reproductive Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (Janet E. Rosenbaum).

The author was funded by National Institutes of Health Training Grants T32 AI050056 and T32 HS00060-12 and a dissertation completion grant from the Harvard University Graduate School of Arts and Sciences.

The author thanks Dr. Nancy Brener and Tim McManus (CDC) for helpful information about their data. She also thanks Dr. Joseph Blitzstein, Dr. Joanne Cox, Dr. Michael Cuthbert, Andrew Davis, Dr. Marc Elliott, Lisa Friedland, Arthur Gaer, Dr. Andrew Gelman, Dr. Loren Hoffman, Dr. Mark Irwin, Dr. Richard Light, Dr. Sherry McKee, Dr. Donald Rubin, Dr. Terry Schell, Dr. Julien Teitler, Dr. Kimberly Thompson, Dr. John Uebersax, and Dr. Alan Zaslavsky for helpful conversations.

Presented at the 64th Annual Conference of the American Association for Public Opinion Research, Hollywood, Florida, May 14–17, 2009; Federal Committee on Statistical Methodology, Washington, DC, November 8–10, 2007; 134th Annual Meeting of the American Public Health Association, Boston, Massachusetts, November 4–8, 2006; and the American Statistical Association's Joint Statistical Meeting, Seattle, Washington, August 6–10, 2006.

Conflict of interest: none declared.

REFERENCES

1. Baldwin W. Information no one else knows: the value of self-report. In: Stone AA, Turkkan JS, Bachrach CA, et al, eds. *Science of Self-report: Implications for Research and Practice*. Englewood Cliffs, NJ: Lawrence Erlbaum Associates; 1999: 3–8.
2. Everett SA, Kann L, McReynolds L. The Youth Risk Behavior Surveillance System: policy and program applications. *J Sch Health*. 1997;67(8):333–335.
3. Sussman MP, Jones SE, Wilson TW, et al. The Youth Risk Behavior Surveillance System: updating policy and program applications. *J Sch Health*. 2002;72(1):13–17.
4. Survey Research Center, Institute for Survey Research. *Monitoring the Future*; 2008. (http://www.monitoringthefuture.org/). (Accessed November 6, 2008).
5. Freier M, Bell R, Ellickson P. *Do Teens Tell the Truth? The Validity of Self-reported Tobacco Use by Adolescents*. Santa Monica, CA: RAND Corporation; 1991. (RAND Note N-3291-CHF).
6. Bauman KE, Koch GG. Validity of self-reports and descriptive and analytical conclusions: the case of cigarette smoking by adolescents and their mothers. *Am J Epidemiol*. 1983;118(1): 90–98.
7. Bauman KE, Koch GG, Bryan ES, et al. On the measurement of tobacco use by adolescents: validity of self-reports of smokeless tobacco use and validity of cotinine as an indicator of cigarette smoking. *Am J Epidemiol*. 1989;130(2): 327–337.
8. Patrick DL, Cheadle A, Thompson DC, et al. The validity of self-reported smoking: a review and meta-analysis. *Am J Public Health*. 1994;84(7):1086–1093.
9. Brener ND, McManus T, Galuska DA, et al. Reliability and validity of self-reported height and weight among high school students. *J Adolesc Health*. 2003;32(4):281–287.
10. Hardt R, Petersen-Hardt S. On determining the quality of delinquency self-report method. *J Res Crime Delinq*. 1977;14: 247–261.
11. Kirk D. Examining the divergence across self-report and official data sources on inferences about the adolescent life-course of crime. *J Quant Criminol*. 2006;22:107–129.
12. Risser JM, Risser WL, Eissa MA, et al. Self-assessment of circumcision status by adolescents. *Am J Epidemiol*. 2004; 159(11):1095–1097.
13. Engels RC, Knibbe RA, Drop MJ. Inconsistencies in adolescents self-reports of initiation of alcohol and tobacco use. *Addict Behav*. 1997;22(5):613–623.
14. Pedersen W. Reliability of drug use responses in a longitudinal study. *Scand J Psychol*. 1990;31(1):28–33.
15. Shillington AM, Clapp JD. Self-report stability of adolescent substance use: are there differences for gender, ethnicity and age? *Drug Alcohol Depend*. 2000;60(1):19–27.
16. Stanton WR, McClelland M, Elwood C, et al. Prevalence, reliability and bias of adolescents reports of smoking and quitting. *Addiction*. 1996;91(11):1705–1714.
17. Bailey SL, Flewelling RL, Rachal JV. Characterization of inconsistencies in self-reports of alcohol and marijuana use in a longitudinal study of adolescents. *J Stud Alcohol*. 1992; 53(6):636–647.
18. Fendrich M, Yun Soo Kim J. Multiwave analysis of retest artifact in the National Longitudinal Survey of Youth drug use. *Drug Alcohol Depend*. 2001;62(3):239–253.
19. Fendrich M, Rosenbaum DP. Recanting of substance use reports in a longitudinal prevention study. *Drug Alcohol Depend*. 2003;70(3):241–253.
20. Fendrich M, Vaughn C. Diminished lifetime substance use over time: an inquiry into differential underreporting. *Public Opin Q*. 1994;58:96–123.
21. Mensch B, Kandel D. Underreporting of substance use in a national longitudinal youth cohort: individual and interviewer effects. *Public Opin Q*. 1988;52:100–124.
22. Alexander C, Somerfield M, Ensminger M, et al. Consistency of adolescents self-report of sexual behavior in a longitudinal study. *J Youth Adolesc*. 1993;22:455–471.
23. Rosenbaum JE. Reborn a virgin: adolescents retracting of virginity pledges and sexual histories. *Am J Public Health*. 2006;96(6):1098–1103.
24. Upchurch DM, Lillard LA, Aneshensel CS, et al. Inconsistencies in reporting the occurrence and timing of first intercourse among adolescents. *J Sex Res*. 2002;39(3): 197–206.
25. Fu H, Darroch JE, Henshaw SK, et al. Measuring the extent of abortion underreporting in the 1995 National Survey of Family Growth. *Fam Plann Perspect*. 1998;30(3):128–138.
26. Brener ND, Billy JO, Grady WR. Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: evidence from the scientific literature. *J Adolesc Health*. 2003;33(6):436–457.
27. Midanik LT. Perspectives on the validity of self-reported alcohol use. *Br J Addict*. 1989;84(12):1419–1423.
28. DeMaio T. Social desirability and survey measurement: a review. In: Turner CF, Martin E, eds. *Surveying Subjective Phenomena*. Vol 2. New York, NY: Russell Sage Foundation; 1984:257–282.
29. Pearson R, Ross M, Dawes R. Personal recall and the limits of retrospective questions in surveys. In: Tanur JM, ed. *Questions About Questions*. New York, NY: Russell Sage Foundation; 1992:65–94.
30. Bernstein R, Chadha A, Montjoy R. Overreporting voting: why it happens and why it matters. *Public Opin Q*. 2001; 65(1):22–44.
31. Presser S. Is inaccuracy on factual survey items item-specific or respondent-specific? *Public Opin Q*. 1984;48:344–355.
32. Silver B, Abramson P, Anderson B. The presence of others and overreporting of voting in American national elections. *Public Opin Q*. 1986;50:228–239.
33. Silver B, Anderson B, Abramson P. Who overreports voting? *Am Polit Sci Rev*. 1986;80:613–624.
34. Locander W, Sudman S, Bradburn N. An investigation of interview method, threat, and response distortion. *J Am Stat Assoc*. 1976;71:269–275.
35. Fendrich M, Mackesy-Amiti ME. Decreased drug reporting in a cross-sectional student drug use survey. *J Subst Abuse*. 2000; 11(2):161–172.
36. Collins L, Graham J, Hansen W, et al. Agreement between retrospective accounts of substance use and earlier reported substance use. *Appl Psychol Meas*. 1985;9:301–309.

37. French SA, Peterson CB, Story M, et al. Agreement between survey and interview measures of weight control practices in adolescents. *Int J Eat Disord*. 1998;23(1):45–56.
38. Poulin C, MacNeil P, Mitic W. The validity of a province-wide student drug use survey: lessons in design. *Can J Public Health*. 1993;84(4):259–264.
39. Winters KC, Stinchfield RD, Henly GA, et al. Validity of adolescent self-report of alcohol and other drug involvement. *Int J Addict*. 1990–1991;25(11A):1379–1395.
40. Brener ND, Kann L, McManus T, et al. Reliability of the 1999 Youth Risk Behavior Survey questionnaire. *J Adolesc Health*. 2002;31(4):336–342.
41. Brener ND, Grunbaum JA, Kann L, et al. Assessing health risk behaviors among adolescents: the effect of question wording and appeals for honesty. *J Adolesc Health*. 2004;35(2):91–100.
42. Brener ND, Collins JL, Kann L, et al. Reliability of the youth risk behavior survey questionnaire. *Am J Epidemiol*. 1995; 141(6):575–580.
43. Tourangeau R. Remembering what happened: memory errors and survey reports. In: Stone AA, Turkkan JS, Bachrach CA, et al, eds. *Science of Self-report: Implications for Research and Practice*. Englewood Cliffs, NJ: Lawrence Erlbaum Associates; 1999:29–48.
44. Fowler F. *Survey Research Methods*. Thousand Oaks, CA: Sage Publications; 1992. (Applied Social Research Methods Series; Vol 1, 2nd ed).
45. Kann L, Kinchen SA, Williams BI, et al. Youth risk behavior surveillance—United States, 1999. *MMWR CDC Surveill Summ*. 2000;49(5):1–32.
46. Adejumo A, Heumann C, Toutenburg H. Review of agreement measure as a subset of association measure between raters. May 26, 2004. Ludwig-Maximilians-Universität, München Sonder- forschungsbereich 386, Paper 385. (www.epub.ub.uni-muenchen.de/1755/1/paper_385.pdf). (Accessed January 19, 2009).
47. Banerjee M, Capozzoli M, McSweeney L, et al. Beyond kappa: a review of interrater agreement measures. *Can J Stat*. 1999;27:3–23.
48. Guggenmoos-Holzmann I, Vonk R. Kappa-like indices of observer agreement viewed from a latent class perspective. *Stat Med*. 1998;17(8):797–812.
49. Lee S, Poon W. Maximum likelihood estimation of polyserial correlations. *Psychometrika*. 1986;51:113–121.
50. Pearson K. Mathematical contribution to the theory of evolution. VII. On the correlation of characters not quantitatively measured. *Philos Trans R Soc Lond A*. 1901;(200):1–66.
51. Uebersax JS, Grove WM. Latent trait finite mixture model for the analysis of rating agreement. *Biometrics*. 1993;49(3): 823–835.
52. Fox J. R polycor package for the R statistical package. (www.cran.r-project.org/web/packages/polycor/polycor.pdf). (Accessed January 19, 2009).
53. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Stat Med*. 2002;21(18):2653–2669.
54. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*. 1995;141(3): 263–272.
55. Denkdukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001;57(1):158–167.
56. Fowler F. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage Publications; 1995. (Applied Social Research Methods Series; Vol 38).