Truth or consequences: the inter-temporal consistency of adolescent risk behavior reporting

Janet Rosenbaum

September 12, 2007

Adolescents engage in behaviors which threaten their future health such as smoking, illegal drug use, and early sexual behavior. Surveys are the primary source of information about many risk behaviors, and the only source for some behaviors (Turkkan 2000, Baldwin 2000). Federal, state, and local governments monitor risk behavior prevalence, set policy priorities, and promote legislation using surveys including the Youth Risk Behavior Survey (YRBS) (Sussman, Jones, et al 2002, Everett, Kann et al 1997), Monitoring the Future, and the National Household Survey on Drug Abuse. The reliability of this information is important for accurately measuring changes over time, determining geographic areas with greater risk behavior prevalence, and for targeting and evaluating public health interventions. Inaccurate data can easily lead to mistakes in policy creation and evaluation.

All survey data is compromised somewhat by incorrect responses, but reports of health risk behaviors are particularly vulnerable because respondents answer surveys in accordance with their self-images and beliefs. Adolescence is a time of identity formation [cite] and adolescents experiment with different identities and behaviors [cite], so they may be more likely to give inconsistent survey responses. Adolescents may under-report stigmatized behavior due to self-presentation bias (Brener et al 2003, Robinson et al 2003) or question threat (Fendrich and Makesy-Amiti 2000), fail to recall behavior (Marquis et al 1981), or over-report risk behavior to improve their social status (Brener et al 2003, Midanik 1989). Studies comparing adolescents' self-report with a gold standard have found over-reporting of smoking (Freier, Bell, and Ellickson 1991), under-reporting of smoking (Bauman and Koch 1983, Bauman, Koch, Bryan et al 1989), lack of knowledge of circumcision status (Risser, Risser, Eissa et al 2004), and over-reporting height and under-reporting weight (Brener, McManus et al 2003). Fewer adolescents report five

weight control practices in personal interviews than in self-administered surveys (French et al 1998).

Test-retest studies reveal inconsistencies, logically impossible or unlikely combinations of responses; inconsistency may indicate bad data, but it can also carry information in itself about respondents' attitudes and self-images. Test-retest studies have found that adolescents retract their earlier reports of engaging in sexual intercourse (Alexander et al 1993, Upchurch et al 2001, Rosenbaum 2006), cigarette smoking (Engels et al 1997, Shillington and Clapp 2000, Pedersen 1990, Stanton et al 1996), the use of alcohol and illegal drugs (Shillington and Clapp 2000, Pedersen 1990, Bailey et al 1992, Fendrich and Rosenbaum 2003, Fendrich and Kim 2001, Fendrich and Vaughn 1994, Mensch and Kandel 1988), and pregnancy, virginity pledges, having a permanent tattoo, driving for respondents under age 15, sex prior to age 13, and pierced ears for men (Rosenbaum 2006). Retrospective reports of substance use are more highly correlated with self-reported present use than with self-reported use reported at the actual past time period (Collins et al 1985). Recanting is most common for intimate, deviant, or illegal behaviors (Fendrich and Vaughn 1994), and recanters are no more likely to supply other forms of bad data, such as skipped questions (Alexander et al 1993). Respondents may also recant experimental behaviors which they initially reported as infrequent (Alexander et al 1993, Mensch and Kandel 1988, Fendrich and Mackesy-Amiti 2000). Many factual questions have an attitude component in ambiguous situations in which respondents must use their judgement in deciding whether to reply (Bailar and Rothwell 1984); some adolescent risk behaviors are known to be interpreted differently.

These test-retest results are in addition to measurement error due to ambiguous or varied question wording, but analysis of adults' responses to ambiguously worded questions about crime reveals that the importance of context effects and measurement error decreases as the salience to respondents increases: measurement contexts influence responses much less for assaults with injury than for attempted assaults with a weapon, which in turn is less influenced than attempted assaults without a weapon, and similarly for thefts with large amounts of money vs. smaller amounts (Turner 1984).

Inconsistent reporting also poses statistical challenges for adolescent risk behavior research. When comparing groups with similar reporting behavior such as two years of the YRBS, studies may underestimate true prevalence differences and be underpowered to detect these differences:

the difference in reported rates is smaller than the difference in true rates, and the power to detect differences in true prevalence is smaller than for reported prevalence. Past studies have found that factors associated with greater retraction are low education (Upchurch et al 2002, Fendrich and Kim 2001, Mensch and Kandel 1988), African-American and Latino ethnicity (Upchurch et al 2002, Fendrich and Rosenbaum 2003, Fendrich and Kim 2001, Fendrich and Vaughn 1994, Mensch and Kandel 1988, Fendrich and Mackesey-Amiti 2000), male gender (Fendrich and Mackesey-Amiti 2000, Siddiqui et al 1999), and younger age (Fendrich and Rosenbaum 2003, Fendrich and Mackesey-Amiti 2000). When studying groups with different reporting behaviors such as two demographics within a single wave of the YRBS, apparent differences between groups may be attributable to reporting behavior more than actual behavior. Inconsistency limits the validity of questions as measuring their intended target, and is a lower bound for the measurement error associated with a question (Fowler 1995, p 147). Inconsistency is expected to change the magnitude of regression coefficients, but not their significance, which may cause inaccurate policy evaluation in cases where the magnitude of effects is used for advocacy purposes. Inconsistency is widely acknowledged in the literature, but no tools exist to compensate for it. Researchers often refer to risk behavior survey instruments "as if having been validated was some absolute state, such as beatification" (Fowler 1995), and most surveys are designed to prevent gathering inconsistent data; for instance, the National Longitudinal Study of Adolescent Health (Add Health) asks only 10 questions in which respondents can contradict an earlier report of a risk behavior.

This study compares adolescents' responses to the same questions about their risk behaviors separated by two weeks, and treats inconsistency as both a statistical and a health issue. As a statistical issue which affects the validity of inferences about differences between groups, this paper measures inconsistency directly and evaluates the extent to which alternative explanations such as true change, memory, or large deviations explain inconsistency; calculates a prevalence-independent measure of inconsistency, and identifies question factors associated with greater inconsistency; estimates error due to inconsistency; and estimates retest effect. Substantial evidence indicates that respondents tend to give survey answers which are associated with their current state, and in some cases inconsistency may communicate adolescents' inconsistent performance to health behaviors, both risky and protective. Protective health behaviors such as condoms, sunscreen, seatbelts, and weight control behaviors are most effective when used

consistently; health risk behaviors are most serious when engaged in consistently.

Most risk behaviors are measured cross-sectionally, and since survey responses are highly conditional on respondents' current mood (Tourangeau 2000, others), the intensity of risk behavior reported by a respondent may vary if the respondent were asked again.

Data can be unreliable in two independent ways — the overall prevalence measured in each of the two waves may differ and the specific respondents who choose each risk behavior may differ; we call the latter form of unreliability inconsistency, and it is the focus of this paper because it represents a form of error not currently accounted for. Prevalence may change without many respondents shifting their answer, but the measured prevalence may remain constant while many individual respondents change their answers. This paper will describe prevalence change and inconsistency, identify factors associated with inconsistency, and estimate error due to inconsistency using a Bayesian method.

Materials and methods

Data

This study uses the contingency tables from a test-retest reliability study of the Youth Risk Behavior Survey (YRBS) performed in 2000 (Brener et al 2002). The YRBS was first developed by invited participants in a 1989 CDC workshop, validated by the Questionnaire Design Research Laboratory at National Center for Health Statistics with laboratory and field testing with high school students, and revised three times before its first administration in 1991 (Brener et al 2004).

The reliability study uses a convenience sample of students at 61 schools from a mixture of urban (48 percent), suburban (39 percent), and rural (13 percent) settings in 20 geographically-dispersed states plus the District of Columbia. Classes in each of the schools were selected to participate and given parent consent forms in advance of the survey. 77 percent of the students in the classes were present in class with a parental consent form on the day of the survey. Of these students who completed the first survey, 89 percent completed the second survey. The final sample of 4619 students over-represents females, African-Americans, grades 9 and 10, and ages 15–16, and under-represents whites, Latinos, grades 11 and 12, and ages 13–14 (Table 1).

Trained data collectors from Macro International Inc. administered 97–100 questions from the Youth Risk Behavior Survey (YRBS) to students on two occasions between February and

April 2000, separated by approximately two weeks. Students alone had access to identification numbers used to link responses. For 57 percent of schools, the interval between survey administrations was exactly two weeks, but the interval had mean 15.6 days and range 10–22 days (Brener et al 2002). The survey was administered in a classroom setting on a computer-scannable questionnaire booklet with questions written above answer choices to avoid off-by-one errors, and took students about 40 minutes to complete. The CDC excluded questionnaires with fewer than 20 valid responses or with the same response option 15 times in a row. They dichotomized questions with multiple response categories into "no risk" and "at risk".

The data are contingency tables for 72 questions from this reliability study. Questions about contraception, substance abuse at last sex, foods consumed, and body weight and height were omitted from the CDC's analysis. The CDC does not make the data available for public use (ND Brener, CDC, personal communication, 2006), but they published prevalence at each survey administration (p_1, p_2) and kappa κ from SAS's proc freq command rounded to one decimal place (Brener et al 2002). Cohen's kappa is a measure of agreement which adjusts for chance agreement commonly used in test-retest and inter-rater agreement studies. The number of respondents who say "yes" at both surveys was estimated from the sample size n, kappa κ and prevalences p_1, p_2 as $a = \frac{n}{2}[\kappa(p_1 + p_2 - 2p_1p_2) + p_1p_2]$ and rounded to the nearest integer. The remaining entries in the contingency tables were obtained through subtraction. The error on our estimates of a due to rounding in the original paper is no more than one respondent.

Survey questions

Respondents were administered 97–100 questions, of which the prevalence and kappas are reported for 72 of the questions (Brener 2002). The questions ask about respondents' use of tobacco, alcohol, and specific drugs, sexual intercourse, symptoms of depression and eating disorders, suicide, violence and weapons use, physical activity, and health-preserving behaviors (wearing seatbelts, helmets, sunblock, visiting dentist and doctor).

Questions were coded according to possible predictors of inconsistent responses. Inconsistent answers may be due to true change or to inaccurate answers in one or both waves. Inaccurate answers in one of the waves may be due to lack of understanding of the question, not knowing the answer, inability to recall the answer, or desire not to report the answer (Fowler 1992).

Memories of more recent events are more accurate (Tourangeau 2000, Fowler 1992), and

true change is more likely for questions about short time-frames, so question time frame was coded: respondent's lifetime, past year, past month, past three months, or an unspecified period. Lack of understanding could be associated with readability of the questions, so questions were coded by word count, number of response choices, whether the previous question was about a different topic (gauged by whether the question was preceded by a transition sentence), whether the previous question was about a different time-frame, and whether the previous question had different answer choices. Inability to recall the answer may be associated with question time-frame and salience (Fowler 1992, Tourangeau 2000); . Respondents may be ambivalent about disclosing deviant, illegal, or stigmatized behavior (Tourangeau 2000), so questions were coded for whether they were about sexual intercourse, illegal drugs, alcohol and tobacco, perpetrating a violent crime, being the victim of a crime, and mental disorder symptoms. Dichotomizing multi-item responses at an arbitrary point may artificially lower TCC due to loss of information, so this was coded.

Data analysis

For discussing individual questions, inconsistency is measured as absolute and relative retraction, and absolute and relative initiation. Absolute retraction is the proportion of the sample who give an affirmative answer followed by a negative answer; relative retraction is the proportion of wave 1 endorsers retracting their report at wave 2, absolute retraction divided by wave 1 prevalence. Absolute initiation is the proportion of the sample who report the behavior at wave 2, but not at wave 1. 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, and rare behaviors have more variable relative retraction and initiation due to large deviations.

For comparing questions, inconsistency is measured as tetrachoric correlation (TCC) an agreement measure independent of prevalence which allows rare and common behaviors to be compared on the same scale (Pearson 1900, Lee and Poon 1986, Uebersax and Grove 1993, Adejumo et al 2004, Banerjee et al 1999, Guggenmoos-Holzmann and Vonk 1998.) The most common agreement measure, kappa, is associated with prevalence so low kappa may be attributable to low prevalence (e.g., Brener 2002). TCC assumes that dichotomized questions

come from a normally-distributed latent variable, and that respondents answer in their affirmative if their latent value exceeds a threshold. Adolescents are known to redefine risk behaviors in accordance with their beliefs [cite], and TCC adjusts for potential differences in response tendency by wave, so if respondents are systematically more conservative or liberal in definition on the second wave. TCC is computed with standard error by the maximum likelihood method in R. TCC can be interpreted as conventional correlation, with 0.0 chance agreement and 1.0 perfect agreement. For the purposes of finding which questions within each category have different levels of TCC, the standard error of TCC for each category of questions is calculated under the null hypothesis of no difference within each category, by calculating the TCC with se for the combined contingency table of all behaviors within the category. TCC for the same risk behavior is compared by time-frame using a test for trend. The mean TCC for each category is computed, and compared using the Tukey test.

To find question characteristics associated with inconsistency, we use stepwise-forward linear regression with outcome variable TCC and the twelve question characteristics described above as the predictors.

Data analysis: error due to inconsistency

The error on the prevalence is estimated using a Bayesian method to simulate the distribution of true disease status conditional on response pattern (Joseph, Gyorkos, Coupal 1995, Craig and Black 2002, Dendukuri and Joseph 2001). The other statistical methods in this paper are frequentist, but error estimation is Bayesian because each question has seven parameters and three degrees of freedom, but prior information can be used to constrain the estimates. The procedure and assumptions for this procedure are discussed in Appendix 1. The error of prevalence is estimated from the simulated distribution; the ratio of the error of prevalence to the error from the conventional Bayesian method is called the standard error multipler.

The convenience sample was compared with the nationally-representative sample in the YRBS by computing z scores of time 1 and 1999 YRBS (Kann, Kinchen, et al 2000).

Prevalence change is assessed using the McNemar test for the equality of off-diagonal elements in a two-by-two contingency table. Inconsistency measures the quantity of retraction and initiation, which may be balanced but a large proportion of respondents. Prevalence change measures balance between retraction and initiation.

Results

Inconsistency

Among 72 questions, median relative retraction is 27 percent (IQR (19.5, 38.2)) and median relative initiation is 28 percent (IQR (19.3, 44.2)) (Table 2). That is, 27 percent of those giving an affirmative answer at wave 1 gave a negative answer two weeks later, and 28 percent of those giving an affirmative answer at wave 2 had given a negative answer two weeks earlier. Relative retraction and initiation does not vary significantly among time-frames and is slightly but insignificantly lower for non-rare behavior.

For 15 lifetime risk behavior questions, at median 23.7 percent of respondents reporting a risk behavior at wave 1 retract at wave 2 (interquartile range (IQR): (11.9, 32.7)); at median 28.7 percent of respondents reporting a risk behavior at wave 2 did not report it at wave 1, and apparently initiated the behavior in the two week interval between waves (IQR (15.5, 38.8)). For 4 questions about risk behaviors before age 13, at median 23.3 percent of respondents reporting a risk behavior at wave 1 retract at wave 2; at median 27.7 percent of respondents reporting a risk behavior at wave 2 did not report it at wave 1. No respondents were below age 13.

Thirteen of 19 non-rare lifetime and before age 13 behaviors have prevalence in the range 10–90 percent. Among these non-rare behaviors, at median 18.4 percent of respondents retract (IQR (6.9, 26.4)), and 19.4 percent of respondents apparently initiate (IQR (19.4, 26.3)). Of respondents reporting at wave 1 having ever been pregnant or made another pregnant, 45.3 percent retracted, and 42.7 percent of the wave 2 respondents reporting pregnancy reported at wave 1 that they had never been pregnant.

For questions about the past year, relative retraction and initiation are higher than 4 percent for 17 of 18 questions. Doctor visits had relative retraction of 21.1 percent and relative initiation of 20.0 percent; dentist visits had 14.8 percent retraction and 10.6 percent initiation.

About 8 percent of respondents reported having ever been pregnant or made another pregnant in each wave; of respondents reporting pregnancy at wave 1, 45.3 percent retracted, and 42.7 percent of the wave 2 respondents reporting pregnancy reported at wave 1 that they had never been pregnant.

Agreement

Summary of agreement

Tetrachoric correlation (TCC) is high and left skewed (median 0.87, IQR (0.80, 0.92)) (Table 2). The questions in the top quartile of TCC are in decreasing order: ever have sex, ever use marijuana, attend PE weekly in average week, ever try smoking, smoke a pack per day for past month, ever drink alcohol, ever smoke daily for at least a month, smoked in past 30 days, ever use cocaine, ever use methamphetamines, use marijuana in past month, use chewing tobacco in past month, attempted suicide in the past year, seriously consider suicide in past year, rode a bicycle and rarely/never use a bicycle helmet in the past year, first used marijuana before age 13, smoke at school in past 30 days, bought cigarettes in store or gas station past 30 days (Table 2).

The 20 questions with TCC in the lowest quartile (in increasing order of TCC) are whether the respondent has ever learned about HIV in school (TCC = 0.45, an outlier), exercise at least 20 minutes in average PE class, first sex before age 13, fasted for weight control past month, watch TV less than 2 hours an average day, sports injury past year, saw doctor when not sick in past year, threatened by or injured by weapon at school in past year, diet pills for weight control past month, vomit/laxatives for weight control past month, ate less for weight control past month, stayed home because school unsafe past month, ever been offered or sold drugs at school, play on sports team past year, no usual cigarette brand past 30 days, inhalants past 30 days, exercise for weight control past 30 days, sad/hopeless past year, ever take steroids, and trying to lose weight. Consider self to be overweight is close to the bottom quartile, with TCC of 0.82.

Among the ten questions about violence, in decreasing order of TCC are forced sex, five questions about perpetrating violence, and four questions about being the victim of violence.

The question about visiting the doctor when not sick in the past year has lower TCC than the question about visiting the dentist in the past year (0.72 versus 0.85).

Predictors of agreement

Among risk behaviors where questions ask about more than one time-frame, TCC is higher for lifetime use than for the past 30 days for six of six risk behaviors, higher for lifetime than before age 13 for four of four risk behaviors, and higher for past 30 days than past 30 days at school

for three of three risk behaviors (Table 3).

TCC does not vary on average with the time-frame of the question in linear regression, even after excluding two outliers for lifetime questions. TCC has virtually the same overall distribution for questions about the past month and the past year.

The TCC for the topics of tobacco, alcohol, and drugs are significantly higher than weight control/physical activity and miscellaneous (doctor, dentist, sunscreen, and HIV education) using Tukey's honest significant difference; TCC for depression is higher than miscellaneous and marginally higher than weight control/physical activity.

In linear regression, TCC is higher for illegal drugs, alcohol, smoking, and sexual intercourse, and lower for questions about weight control(Table 4). TCC is not associated with the other factors evaluated. As expected from its properties, TCC is not associated with prevalence.

Within the category of violence, the most consistent question was about forced sex; the five questions about perpetrating violence were the next most consistent questions; and four questions about being victim of violence were the least consistent.

Within the category of depressive symptoms in the past year, the highest TCC was considered suicide (0.94), attempted suicide (0.94), planned suicide (0.90), injured in suicide attempt to need medical attention (0.87), and felt sad or hopeless for at least two weeks (0.80).

Within the category of sex, the highest TCC question was whether the respondent had ever had sexual intercourse (TCC=0.99), next was sex in the past 3 months (0.91), 4 or more lifetime sex partners (0.82), ever pregnant or made another pregnant (0.81), and had sex before age 13 (0.66).

Error due to inconsistency

Unreliable data increases standard error at median by a factor of 3.The questions with the lowest error are sexual intercourse (1.6), marijuana (1.7), and smoking cigarettes (2.0-2.1), and the questions with the highest error are fasting to lose weight in the past month (4.9), ever being taught about HIV in school (5.1) and rarely/never wearing a motorcycle helmet when riding a motorcycle in the past month (5.6). There is no pattern in estimated error due to inconsistency for all questions, but error is inversely proportional to prevalence among the legal and illegal substance use questions.

Prevalence change/ retest effect

The prevalence of 41 of 72 behaviors changed under retest in a two week interval, at the 0.05 level (Table 2). More respondents reported using alcohol, cigarettes, and marijuana before age 13 in the second survey than in the first, and fewer reported sexual intercourse. The prevalence of 11 of the 15 lifetime risk behavior questions changed. Five behaviors decreased in lifetime incidence: inhalants, having had four lifetime sexual partners, alcohol, smoking, and marijuana.

Prevalence change was not associated with any predictors in two logistic regressions and 1 linear regression.

Discussion

Inconsistency

Measurement error for adolescents' self-reported risk behavior appears substantial. About a quarter of respondents reporting risk behaviors give inconsistent answers, and this proportion is not significantly smaller where inconsistency is logically impossible or the behavior is not rare. The inconsistency in a test-retest survey is a conservative estimate of measurement error since respondents tend to repeat answers given in an earlier survey (Fowler 1995, 147).

Inconsistency appears to be driven by topic rather than qualities of the survey.

Inconsistency cannot be explained by true change. If true change explained inconsistency, questions where respondents could not logically change answers would have negligible inconsistency and questions about longer time-frames would have less inconsistency than shorter time-frames, but questions of all time-frames have similar levels of inconsistency. Respondents can logically initiate in questions about the past year, but in order to retract they would have needed to engage in the behavior 11.5–12 months prior to the first survey. If behavior is uniformly distributed throughout the year, fewer than 4 percent of affirmative respondents would change their answers to questions about the past year, but nearly all questions have retraction and initiation above 4 percent.

Inconsistency cannot be attributed to large deviations for rare behaviors. Even for non-rare behaviors (defined conservatively as behaviors with prevalence above 10 percent), median relative retraction and initiation are about one fifth of affirmative responses.

Some inconsistencies evoke absurd conclusions such as that about half of lifetime pregnancies occurred in the past 2 weeks, and about half of people reporting pregnancy forget about their pregnancy in a two week interval.

Agreement

Virgin status has the highest TCC. Substance use questions have substantially higher TCC than other questions, and comprise 13 of 18 questions with TCC in the top quartile.Weight control questions have the lowest agreement of all question types: 6 of 7 weight control questions are in the lowest quartile of TCC, and self-image as overweight is close to the lowest quartile. Patterns in agreement are not attributable to risk behavior prevalence because TCC is independent of prevalence, and this independence was verified empirically. Many adolescents may have well-

defined preferences how to portray their substance use history and virgin status because these are salient to adolescents (Brener 2002, Tourangeau 2000), but fewer may have well-defined preferences how to portray their body image and past year's weight control history. When respondents do not have clear preferences how to portray one aspect of their history, they may answer randomly, treating these health behaviors as non-attitudes (Smith 1984). Respondents are more likely to remember (Fowler 1992) and give consistent responses about (Smith 1984) events with significant impact on their lives and in the recent past than events with less impact and in the distant past , so the difference in consistency may indicate that adolescents believe lifetime substance use and virgin status are relatively important in their lives, and weight control is relatively unimportant.

Social desirability appeared not to be a major factor in inconsistency in the expected direction: adolescents were more consistent about illegal drug use than other risk behaviors and they were more consistent about some suicide behaviors than other risk behaviors; they were no less consistent about perpetrating or being a victim of a crime than about other risk behaviors. The low consistency of weight control questions could indicate low social desirability or inconsistent efforts. Social desirability may be less of a factor due to the survey being self-administered and in a format which attempts to reassure respondents of the anonymity of their answers due to respondents' sole access to the code which links their responses (Fowler 1992).

Patterns in agreement are not attributable to risk behavior prevalence because TCC is independent of prevalence, and this independence was verified empirically.

Respondents may change their responses to weight control questions due to changing their intent to lose weight. Weight control questions are about the past month, so respondents could logically change their answers if the behavior occurred exactly 2–4 weeks prior to the survey. Weight control questions have substantially lower TCC than other questions about the past month, and a larger portion of the inconsistency may be due to true behavior change: many adolescents may initiate and terminate weight control efforts on time scales as short as two weeks. Past studies which show that adolescents' self-initiated weight control efforts are associated with later weight gain after controlling for initial BMI speculate that the reason for the association is that adolescents alternate restrictive and overeating (e.g., Field et al 2007), but studies of nationally representative samples look at intervals of a year. These data for a large diverse sample show inconsistent weight control efforts over a two week period.

Adolescent intent to lose weight seems to vary with time and may even affect adolescents' adherence to physician-supervised weight control interventions.

Virgin status has the highest TCC, but sex before age 13 has among the lowest TCC. This discrepancy could be due to poorly defined preferences for presentation of age of first sex. The discrepancy could be partially a statistical artifact of the dichotomization of a multi-item scale, but such questions had only marginally lower TCC than others.

Adolescents admit perpetrating violence more consistently than being the victim of violence, except forced sex; forced sex has substantially higher TCC than other victim questions. Victims of forced sex may report more consistently due to saliency compared with other acts of violence. Victims may be reluctant to consistently admit the crime, or may regard the crime as less significant than the perpetrators. The victim questions may also be more ambiguous or subjective, such as the low TCC question asking whether respondents have stayed home from school in the past year because it felt too unsafe, or whether they were threatened with a weapon at school (Turner 1984).

Adolescents are more consistent in reporting suicidal ideation and attempts than depressive symptoms. Within the category of depressive symptoms in the past year, the highest TCC was considered suicide (0.94), attempted suicide (0.94), planned suicide (0.90), injured in suicide attempt to need medical attention (0.87), and felt sad or hopeless for at least two weeks (0.80).

Within the category of sex, the highest TCC question was whether the respondent had ever had sexual intercourse (TCC=0.99), next was sex in the past 3 months (0.91), 4 or more lifetime sex partners (0.82), ever pregnant or made another pregnant (0.81), and had sex before age 13 (0.66).

The questions about doctor and dentist appointments may have substantially different TCCs because the doctor question has an extra qualification, "visited the doctor in the past 12 months when you were not sick" as opposed to "visited the dentist in the past 12 months" which may decrease the salience of the answer since it requires evaluation and memory whether had been sick when visited the doctor the last time; respondents may have misunderstood the intent of the question to measure wellness visits as opposed to visits provoked by sickness.

Learn about HIV at school may have lowest TCC because virtually all adolescents have learned about HIV, and they may not reliably recall whether school was one of the places that they learned about HIV.

Looking at the TCCs of individual questions, it seems that many questions which are more inconsistent share these traits: ambiguity and greater number of qualifications, subjectivity in interpretation, and low importance to adolescents.

Past studies have noted that verbal (rather than quantitative) frequency descriptors exhibit substantial individual variation in how people interpret them based on their tastes and prior probabilities (Moxey and Sanford 1992), but questions with verbal quantifiers were not less accurate than those with numerical quantifiers: bike helmets were among the most consistent while motorcycle helmets were less consistent.

TCC may be used as a measure of the relative stigma or social desirability of admitting certain behaviors among respondents who will admit the behaviors at least once, or as the relative importance of the event to adolescents. The social desirability of these questions could be measured using one of several scales developed for the purpose, but these scales have been critiqued as measuring a poorly formulated concept since social desirability can be viewed alternatively as a quality of a survey item wording, a subject, or an approval-seeking personality trait (DeMaio 1984).

Inconsistent responses may be due to the time between surveys. The second answers may not necessary be more accurate, but they are based in greater time for reflection and recall. Respondents may change their answers due to recalling their behavior differently or due to discussion of the survey with friends, and may decide to conceal previously revealed behavior or reveal previously concealed behavior, or to falsely claim to engage in a risk behavior they had previously (and correctly) not reported. Adolescents' communication behavior is sometimes information seeking: they say something to see what the reaction is, or to try out an identity, and some adolescents may be inconsistent because their answer was experimental.

When reports change in impossible ways, at least one of the waves must be incorrect, but it's not possible to know which report is accurate.

The YRBS is administered only once, so change due to retest does not compromise the test, but if inconsistency indicates that a large proportion of wave 1 answers are inaccurate, that is problematic. If peer discussion affects consistency, the full data could reveal a general trend towards increased or decreased prevalence within each school. A future reliability test of the YRBS could have two arms: a test-retest arm and a preview arm in which respondents are given the questions in advance of the test-retest. If delay explains inconsistency, the preview

arm would have higher TCC.

Predictors of agreement

TCC is consistently highest for lifetime incidence, possibly due to the higher salience of lifetime behavior. Adolescents may consider lifetime use part of their identity, but require recall to answer more specific questions. If true change accounted for inconsistency across all questions, we would expect greater overall TCC for questions about the past year than the past month irrespective of question topic, but TCC for questions about the past year does not differ from the past month. This pattern is expected from the fact that the retrieval of autobiographical memories and assessment whether they took place within a specific context is more cognitively complex than verifying autobiographical facts (Conway 1996); the more specific question which adds a place to the time-frame requires more complex evaluation than the question which asks only about the time-frame.

Questions later in the survey had lower TCC than earlier questions, which could be due to respondent fatigue or to the lower salience of the questions later in the survey. The high salience and high TCC topics such as drug use and sex appear early in the survey, and it is likely that question number is simply associated with saliency rather than being an independent factor: question topic and question wording are thought to have larger effects on responses (Bradburn 1992).

Errror

Error due to inconsistency was estimated by assuming that there is one true answer to the question for both time periods which does not change, particularly reasonable for questions with at most negligible changes in a two week period, such as before age 13, lifetime, and past year.

Prevalence change / Retest effect

A majority of YRBS questions are unreliable at a level which is statistically significant, but unlikely to be practically significant. Inconsistency in which individuals report behavior differently does have practical implications. The original CDC report compared 95 percent confidence intervals constructed with sampling error under the assumption of independence (Brener 2002), which biases results towards finding no difference between groups since independent observa-

tions have higher standard error than non-independent repeated observations from the same individuals.

Changes in prevalence may be attributable to the retest effect, the tendency for respondents to give different answers when answering a test for the second time than when answering the test the first time. No pattern in prevalence changes are evident other than a weak tendency for reported drug use to increase. Retest effects do not impact the validity of the YRBS because the YRBS is given only once.

As Brener (2002) found, we find that substance use and sex are the most consistent topics, but looking within categories we find that virgin status is the only sexual question which has high consistency: the other sexual questions have low consistency. As did Brener, we find there is no statistically significant difference in the mean consistency of questions about lifetime, past year, and past month, but we find a significant difference within each risk behavior by time-frame. We find that weight control has substantially lower TCC than other behaviors. The use of TCC instead of kappa means that low agreement cannot be attributed to low prevalence. While inconsistency could be due to true change (Brener 2002), in some cases, true change cannot fully account for the differences between waves.

Limitations

Published papers about the YRBS break down data by characteristics such as gender, ethnicity, school, age for grade, and age, but this study cannot find demographic correlates of inconsistency due to lack of access to the complete data. Demographic data would also allow the evaluation of whether inconsistency follows patterns of memory accuracy, such as females more consistent than males (Bradburn 2000). This study cannot say whether the inconsistency is perpetrated by the same individuals, or whether a large proportion of respondents are inconsistent on some questions, and whether individuals are inconsistent on related questions. The test-retest study examined the consistency of reports about contraception and other sexual behavior, but the prevalences and kappas for these data did not appear in the original report. Response behavior likely varies with demographics, and so sensitivity and specificity are not uniform across all groups; the thresholds estimated for TCC also vary by demographic group, and could be computed separately. Inconsistency on weight control questions may be associated with inconsistent intent to lose weight or inconsistent body image or overweight status. These

questions and others can be answered using analysis of the existing data, and these limitations can be overcome by the release of the full data, as has been advocated in other social sciences (King 2006).

Social desirability is likely to be associated with retraction. Respondents may retract undesirable answers, but desirability is relative to demographic group such as age. We cannot model the likely direction of the bias with aggregated data.

The geographically diverse convenience sample is not nationally representative; Latinos and whites are under-represented and blacks are over-represented, and 15 and 16 year olds are over-represented relative to other adolescent age groups (Table 1). The sample is also less likely to engage in risk behaviors than the nationally-representative YRBS sample.

The Bayesian model is under-identified, but the priors for four parameters – sensitivity and specificity — restrict problem, and estimates are stable.

Conclusions

Adolescents report most accurately behaviors about which they have well-defined preferences and which are salient for their identities, such as lifetime substance use and virgin status (see Brener 2002). Adolescents are most inconsistent in reporting weight control behaviors, possibly due to poorly-defined preferences about weight control and inconsistent body image. Inconsistent weight control behaviors can be explored to greater depth using the existing full test-retest data. The rise of adolescent overweight makes understanding adolescents' current weight control behaviors particularly important for understanding the dynamics of adolescent overweight and predictors of erratic dieting efforts in a large diverse population, and also because adolescents' inconsistent body image may affect their adherence to weight control interventions. The association between question reliability and topic saliency does not rule out the possibility that some YRBS questions could benefit from rewriting and further testing. The recent importance of adolescent obesity and weight control makes it particularly important that measurements be accurate.

Both consistency in a short period and inconsistency in a long time period seem linked to adolescents' identities. In the long-term, recanting is more common for intimate, deviant, or illegal behaviors (Fendrich and Vaughn 1994, Fendrich and Mackesy-Amiti 2000), but sex, drugs, smoking, and suicide have the lowest short-term inconsistency. This apparent contradiction can

be resolved if we view short-term inconsistency as due to behaviors' salience to identity, and long-term inconsistency associated with changed identity (e.g., Rosenbaum 2006).

Differences in reported risk behaviors found between different regions or demographic groups in the YRBS and similar surveys may be attributable to different reporting behaviors. Survey data is used to make policy for rare adolescent risk behaviors such as pregnancy, and cocaine, inhalant, and methamphetamine use, and such rare behaviors have particularly inflated errors. Without a model of demographic variation in reporting behavior, prevalence estimates cannot account for such differences.

Adolescents' risk behavior and self-report are self-consciously chosen often in connection with perceived social desirability of the behavior and the claim. Adolescent risk behavior research needs to be conscious of the implications that adolescents may view even answering anonymous surveys as a social statement.

		Stool test T_1		
		+	-	Total
Serology T_2	+	38	87	125
	-	2	35	37
	Total	40	122	162

Table 1: (a) Results of two tests for *Strongyloides* infection from a study to estimate the prevalence of *Strongyloides* in Cambodian refugees to Canada in 1982–83 for which only two imperfect tests are available: a stool test and a serology test. The two tests estimated dramatically different prevalences — 25% and 77%, respectively — due to the properties of the tests. As pointed out by original analysis of data (Joseph et al 1995/2000) these different prevalences do not even take into account the possibility of sampling error, false positives and false negatives.

Appendices

Data analysis: error due to inconsistency

Estimating risk behavior prevalence using imperfect survey questions of unknown accuracy is analogous to the epidemiological problem of estimating disease prevalence using imperfect diagnostic instruments of unknown accuracy. We use a Bayesian method designed for this epidemiological problem.

The data from two medical binary tests for a disease have three degrees of freedom, but are described by seven parameters: population prevalence, sensitivity and specificity for each test, the probability of true positives on both tests, and the probability of true negatives on both tests. The problem is under-determined, that is, any combination of sensitivity, specificity, and prevalence may be consistent with the observed data. For example, in the below data for *Strongyloides* infection it would be equally consistent with the data for 100% of the population to have the disease, but one test only finds the disease 77% of the time and the other test only finds it 25%, and for 0% of the population to have the disease but the test have false positive rates of 77% and 25%. Prior knowledge leads us to believe that neither is the case, but the data itself does not exclude these possibilities. We use priors in a Bayesian model to exclude such improbable answers.

Frequentist methods for two tests in one population assume that tests are independent conditional on disease status and that two of the test parameters are known with precision, and compute the remaining three parameters and estimate variance (Hui and Zhou 1998). Fixing parameters underestimates uncertainty, and the values which are fixed are often chosen

arbitrarily.

Bayesian methods use a prior distribution on the seven parameters to compensate for lack of identifiability and view all seven parameters as uncertain. Frequentist methods which assign an exact value to some parameters yield answers equivalent to Bayesian method which uses a prior which is a point mass at the assigned value (Joseph et al 1995). The Bayesian methods are more realistic because it's unlikely to know any parameters so precisely, and they also avoid parametric assumptions in creating credible intervals around estimated parameters (Joseph et al 1995). Bayesian methods to estimate prevalence of a disease with two non-gold standard tests are described in (Joseph, Gyorkos, Coupal 1995, Craig and Black 2002, Dendukuri and Joseph 2001), and in greater detail below. The prior compensates for lack of identifiability, so the prior influences posterior results, even with more data (Dendukuri and Joseph 2001).

Parameters and notation

Two tests with binary results yield a two-by-two contingency table with cells n_{ij} where n_{ij} is the number of respondents giving answer i at wave 1 and answer j at time 2. Divide up each cell n_{ij} into the latent data: truly positive (denoted a_{ij}) and the truly negative (denoted b_{ij}).

Let C_1 and C_2 be specificity of tests 1 and 2, respectively, S_1 and S_2 sensitivity, and π be the prevalence of the behavior in the population. Let $p_{ij|k}$ be the probability that someone with disease state k will give responses (i, j). The tests are not independent. Due to inter-test agreement, $p_{ii|k} > p_{i.|k}p_{.i|k}$ for test result i and true disease state k. We include two parameters to describe correlation between answers $p_{00|0}, p_{11|1}$ which must satisfy $S_1S_2 < p_{11|1} < \min(S_1, S_2)$ and $C_1C_2 < p_{00|0} < \min(C_1, C_2)$ (Craig and Black 2002, Dendakuri and Joseph 2001).

Priors: original approach

Rare behaviors, with 1999 YRBS prevalence less than 15 percent, are given a uniform distribution on the interval (0,0.5). All other risk behaviors are given a uniform distribution on the unit interval (0.0,1.0). The 29 rare behaviors are: lifetime forced sex, cocaine, inhalant, heroin, methamphetamine, steroid, injected illegal drugs, and pregnancy; past year injured in fight, threatened with weapon at school, physically hurt by boyfriend/girlfriend, attempted suicide, injured in suicide attempt; past month drove after drinking, carried gun, carried weapon at school, felt too unsafe for school, bought cigarettes, bought cigarettes and carded, smoke at

school, smokeless tobacco, smokeless tobacco at school, no usual cigarette brand, alcohol at school, marijuana at school, cocaine, inhalants, diet pills for weight control, vomit for weight control.

We tried several priors for sensitivity and specificity: uniform, Beta(8,1), Beta(16,2), Beta(32,4), and choose the most diffuse prior which minimizes the number of prevalence estimates which differ substantially from past survey data. In the end, we use Beta(16,2) which means that we assume that sensitivity and specificity are probably at least 50 percent, with a near-zero probability that these parameters are less than 50 percent.

Parameter	Prior	
Prevalence of rare risk behavior	$\pi \sim \mathrm{U}(0, 0.5)$	
Prevalence of non-rare risk behavior	$\pi \sim \mathrm{U}(0, 1)$	
Sensitivity	$S_i \sim \mathrm{Beta}(\alpha_\mathrm{i}, \beta_\mathrm{i})$	$(\alpha_i, \beta_i) \in \{(1,1), (8,1), 16, 2)\}$
Specificity	$C_i \sim \mathrm{Beta}(\alpha_\mathrm{i}, \beta_\mathrm{i})$	C_i such that $C_i + S_i > 1$
Probability true positive agreement	$p_{11\|1} \sim \mathrm{U}(S_1 S_2, \min(S_1, S_2))$	
Probability true negative agreement	$p_{00\|0} \sim \mathrm{U}(C_1 C_2, \min(C_1, C_2))$	

Priors

Sensitivity and specificity are given beta priors.We allow for potential retest effects by not requiring sensitivity or specificity to be identical in the first and second surveys since a respondent may view a question differently the second time than the first time.

The survey questions are assumed to convey some information, so $S+C > 1$ for each wave of the survey. Sensitivity and specificity are constant over the population due to lack of covariates.

The correlation between answers $p_{00|0}, p_{11|1}$ have uniform distribution on the intervals to which they are constrained due to inter-test agreement. $p_{11|1} \sim U(S_1 S_2, \min(S_1, S_2))$, $p_{00|0} \sim U(C_1 C_2, \min(C_1, C_2))$.

Gibbs sampler

After drawing the seven parameters from the above priors, the probabilities of response patterns conditional on true behavior status $p_{ij|k}$ are computed from the values of the six test properties. Bayes's rule uses these probabilities and the prevalence estimate to compute $p_{1|ij}$ the probability that the respondent truly engaged in the behavior conditional on response pattern. The number of people who truly engaged in the behavior and gave survey response pattern ij is drawn from a binomial distribution $a_{ij} \sim \mathrm{Bin}(\mathrm{n_{ij}}, \mathrm{p_{1|ij}})$. The posterior is a Beta distribution with

parameters updated by the simulated latent data a_{ij}. Due to the large amount of data, the estimates converge quickly, but we use 5000 iterations and discard the first 1000 as burn-in. The programming was checked using the original *Strongyloides* data to ensure the same answers were reached.

The Gibbs sampler gives accurate results, but if the model is insufficiently specific the estimates will diverge. Due to lack of identifiability, some estimates of rare events differ substantially from past survey results. For each variable, we do 100 Gibbs samplers to be able to detect divergent prevalence estimates, and use the most vague prior that also avoids divergent prevalence estimate, that is, estimates which differ more than 50 percentage-points from the prevalence found in past surveys.

Gibbs sampler procedure

1. Draw initial values for the seven parameters from the prior distributions.

Parameter	Prior	
Prevalence	$\pi \sim \text{Beta}(1,1)$	truncated at 0.5 if rare
Sensitivity	$S_i \sim \text{Beta}(\alpha_\text{i}, \beta_\text{i})$	
Specificity	$C_i \sim \text{Beta}(\alpha_\text{i}, \beta_\text{i})$	require $C_i + S_i > 1$
Prob both true positive	$p_{11\|1} \sim \text{U}(S_1 S_2, \min(S_1, S_2))$	
Prob both true negative	$p_{00\|0} \sim \text{U}(C_1 C_2, \min(C_1, C_2))$	

   α, β are chosen to avoid divergent estimates.

   If tests are independent $p_{11|1} = P[T_1^+ T_2^+ | D^+] = P(T_1^+ | D^+) P(T_2^+ | D^+) = S_1 S_2)$, and if the tests are totally correlated/dependent, then $p_{11|1} = P[T_1^+ T_2^+ | D^+] = P(T_1^+ | D^+) P(T_2^+ | T_1^+) = S_1 \times 1 = S_1$ or vice versa, so we constrain $p_{11|1}$ to the interval between these two estimates, and similarly for $p_{00|0}$.

2. Compute the probabilities of each response pattern conditional on true behavior status and the seven parameter values.

Parameter	Function
$p_{10\|1}$	$S_1 - p_{11\|1}$
$p_{01\|1}$	$S_2 - p_{11\|1}$
$p_{00\|1}$	$1 - S_1 - S_2 + p_{11\|1}$
$p_{10\|0}$	$C_2 - p_{00\|0}$
$p_{01\|0}$	$C_1 - p_{00\|0}$
$p_{11\|0}$	$C_1 - C_2 + p_{00\|0}$

3. Compute the probability of truly having engaged in the behavior, based on response pattern.

   $p_{1|ij} = \frac{\pi p_{ij|1}}{\pi p_{ij|1} + (1-\pi) p_{ij|0}}$

4. Draw latent data: the number of respondents with each response pattern who are truly positive:

   $a_{ij} \sim \mathrm{Bin}(\mathrm{n_{ij}}, \mathrm{p(1|ij)})$

   $b_{ij} \sim n_{ij} - a_{ij}$

5. Calculate the likelihood given the latent data.

   $p_{11|1}^{a_{11}} p_{10|1}^{a_{10}} p_{01|1}^{a_{01}} p_{00|1}^{a_{00}} p_{11|0}^{n_{11}-a_{11}} p_{10|0}^{n_{10}-a_{10}} p_{01|0}^{n_{01}-a_{01}} p_{00|0}^{n_{00}-a_{00}}$

6. The posterior probabilities are

Parameter	Prior	
Prevalence	$\pi \sim \mathrm{Beta}(1 + \sum \mathrm{a_{ij}}, 1 + \sum \mathrm{b_{ij}})$	
Sensitivity 1	$S_1 \sim \mathrm{Beta}(\alpha + \mathrm{a}_{11} + \mathrm{a}_{10}, \beta + \mathrm{a}_{01} + \mathrm{a}_{00})$	
Specificity 1	$C_1 \sim \mathrm{Beta}(\alpha + \mathrm{b}_{01} + \mathrm{b}_{00}, \beta + \mathrm{b}_{11} + \mathrm{b}_{10})$	require $C_1 + S_1 > 1$
Sensitivity 2	$S_2 \sim \mathrm{Beta}(\alpha + \mathrm{a}_{11} + \mathrm{a}_{01}, \beta + \mathrm{a}_{10} + \mathrm{a}_{00})$	
Specificity 2	$C_2 \sim \mathrm{Beta}(\alpha + \mathrm{b}_{10} + \mathrm{b}_{00}, \beta + \mathrm{b}_{11} + \mathrm{b}_{01})$	require $C_2 + S_2 > 1$
Prob both true positive	$p_{11\|1} \sim \mathrm{U}(S_1 S_2, \min(S_1, S_2))$	
Prob both true negative	$p_{00\|0} \sim \mathrm{U}(C_1 C_2, \min(C_1, C_2))$	

Questionnaire

The YRBS questionnaire from 1999 is available at

http://web.archive.org/web/19991128160242/www.cdc.gov/nccdphp/dash/yrbs/survey99.htm

References

Aaron DJ, Kriska AM, Dearwater SR, Cauley JA, Metz KF, LaPorte RE. Reproducibility and validity of an epidemiologic questionnaire to assess past year physical activity in adolescents. Am J Epidemiol 1995;142:191–201.

Adejumo AO, Heumann C, Toutenburg H. Review of agreement measure as a subset of association measure between raters. Manuscript. May 26, 2004.

Alexander MG, Fisher TD. Truth and consequences: using the bogus pipeline to examine sex differences in self-reported sexuality. J Sex Res 2003;40:27-35.

Alexander CS,Somerfield MR, Ensminger ME, Johnson KE, Kim YJ. Consistency of adolescents' self-report of sexual behavior in a longitudinal study. Journal of Youth and Adolescence 1993;22:455–71.

Bailar BA, Rothwell ND. Measuring employment and unemployment. In: Turner CF, Martin E, ed. Surveying subjective phenomena, vol 2. New York : Russell Sage Foundation; 1984: 129–142.

Bailey SL, Flewelling RL, Rachal JV. Characterization of inconsistencies in self-reports of alcohol and marijuana use in a longitudinal study of adolescents. Journal of Studies on Alcohol 1992;53:636–47.

Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: A review of interrater agreement measures. Canadian J Statistics. 1999;27:3–23.

Banken JA. Drug Abuse Trends among youth in the United States. Annals of the New York Academy of Sciences. 2004; 1025 (1): 465–71.

Bauman KE, Koch GG, Bryan ES, Haley NJ, Downton MI, Orlani MA. On the measurement of tobacco use by adolescents: validity of self-reports of smokeless tobacco use and validity of continine as an indicqator of cigarette smoking. Am J Epidemiol 1989; 130:327–37.

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): 90–8.

Beebe TJ, Harrison PA, Mcrae JA, Anderson RE, Fulkerson JA. An evaluation of computer-assisted self-interviews in a school setting. Public opinion quarterly 1998; 62(4):623–32.

Boekeloo BO, Schamus LA, Simmens SJ, Cheng TL. Ability to measure sensitive adolescent behaviors via telephone. American J preventative medicine 1998;14(3):209–16.

Berkvens D, Speybroeck N, Praet N, Adel A, Lesaffre E. Estimating disease prevalence in a Bayesian framework using probabilistic constraints. Epidemiology 2006; 17(2):145–53.

Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. Statistics in Medicine 2002; 21:2653–69.

Bradburn NM. What have we learned? In: Schwarz N, Sudman S, ed. Context effects in social and psychological research. New York, NY: Springer-Verlag; 1992: 315–323.

Brener ND, Kann L, Kinchen SA, Grunbaum JA, Whalen L, Eaton D, Hawkins J, Ross JG. Methodology of the Youth Risk Behavior Surveillance System. MMWR 2004;53(No. RR-12)

Brener ND, Billy JOG, Grady WR. Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: evidence from the scientific literature. Journal of Adolescent Health. 2003;33:436–457.

Brener ND, Grunbaum JA, Kann L, McManus T, Ross J. Assessing health risk behaviors among adolescents: the effect of question wording and appeals for honesty. Journal of adolescent health. 2004;35:91–100.

Brener ND, Kann L, McManus T, Kinchen SA, Sundberg EC, Ross JG. Reliability of the 1999 Youth Risk Behavior Survey questionnaire. Journal of Adolescent Health. 2002;31:336–342.

Brener ND, McManus T, Galuska DA, Lowry R, Wechsler H. Reliability and validity of self-reported height and weight among high school students. Journal of Adolescent Health. 2003;32:281–287.

Brener ND. Personal communication to Janet Rosenbaum. October 18, 2006.

Brener ND, Collins JL, Kann L. Reliability of the Youth Risk Behavior Survey questionnaire. Am J Epidemiology. 1995; 142:191–201.

Conway MA. Failures of autobiographical remembering. In: Herrmann D, McEvoy C, Hertzog C, Hertel P, Johnson MK, ed. Basic and applied memory research: Theory in context, vol 1. Mahway, NJ: Lawrence Erlbaum Associates; 1996: 295–315.

Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied statistics. 1979;28:20–28.

DeMaio TJ. Social desirability and survey measurement: A review. In: Turner CF, Martin E, ed. Surveying subjective phenomena, vol 2. New York : Russell Sage Foundation; 1984: 257-282.

Denkdukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001; 57:158–167.

Eisenmann JC. Secular trends in variables associated with the metabolic syndrome of North American children and adolescents: a review and synthesis. Am J Hum Biol. 2003 Nov-Dec;15(6):786-94.

Engels RCME, Knibbe RA, Drop MJ. Inconsistencies in adoelscents' self-reports of initiation of alcohol and tobacco use. Addictive Behaviors. 1997;22:613–623.

Erkanli A, Soyer R, Costello EJ. Bayesian inference for prevalence in longitudinal two-phase

studies. Biometrics. 1999;55:1145–1150.

Everett SA, Kann L, McReynolds L. The Youth Risk Behavior Surveillance System: policy and program applications. J Sch Health. 1997 Oct;67(8):333-5.

Fendrich M, Rosenbaum DP. Recanting of substance use reports in a longitudinal prevention study. Drug and Alcohol Dependence. 2003;70:241–253.

Fendrich M, Kim JYS. Multiwave analysis of retest artifact in the National Longitudinal Survey of Youth drug use. Drug and Alcohol Dependence. 2001;62:239–253.

Fendrich ME, Makesy-Amiti ME. Decreased drug reporting in a cross-sectional student drug use survey. Journal of substance abuse. 2000;11:161–172.

Fendrich M, Vaughn CM. Diminished lifetime substance use over time: an inquiry into differential underreporting. Public Opinion Quarterly. 1994;58:96–123.

Fowler FJ. Improving survey questions: Design and evaluation. Applied Social Research Methods Series, vol 38. Thousand Oaks, CA: SAGE publications. 1995.

Fowler FJ. Survey Research Methods, second edition. Applied Social Research Methods Series, vol 1. Thousand Oaks, CA: SAGE publications. 1992.

Freier MC, Bell RM, Ellickson PL. Do teens tell the truth? The validity of self-reported tobacco use by adolescents. RAND Note N-3291-CHF. 1991.

French SA, Peterson CB, Story M, Anderson N, Mussell MP, Mitchell JE. Agreement between survey and interview measures of weight control practices in adolescents. Int J Eat Disord 1998; 23:45–56.

Gfroerer J, Wright D, Kopstein A. Prevalence of youth substance use: the impact of methodological differences between two national surveys. Drug and alcohol dependence. 1997; 47:19–30.

Guggenmoos-Holzmann I, Vonk R. Kappa-like indices of observer agreement viewed from a latent class perspective. Statistics in medicine. 1998; 17:797–812.

Gustafson P. The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. Statistics in Medicine. 2005;24:1203–1217.

Gustafson P, Le ND, Saskin R. Case-control analysis with partial knowledge of exposure misclassification probabilities. Biometrics. 2001; 57:598–609.

Hallfors D, Khatapouse S, Kadushin C, Watson K, Saxe L. A comparison of paper vs computer-assisted self interview for school alcohol, tobacco, and other drug surveys. Evaluation

and program planning. 2000; 23:149–155.

Hearn KD, O'Sullivan LF, Dudley CD. Assessing reliability of early adolescent girls' reports of romantic and sexual behavior. Archives of sexual behavior. 2003;32(6):513–521.

Hendrickson S, Mattes R. Financial Incentive for Diet Recall Accuracy Does Not Affect Reported Energy Intake or Number of Underreporters in a Sample of Overweight Females. J Am Diet Assoc. 2007 Jan;107(1):118-121.

Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Statistical methods in medical research. 1998;7:354–370.

Johnson WO, Gastwirth JL, Pearson LM. Screening without a gold-standard: the Hui-Walter paradigm revisited. American Journal of Epidemiology. 2001; 153:921–924.

Johnston LD, O'Malley PM. The Recanting of Earlier Reported Drug Use by Young Adults. NIDA Res Monogr. 1997;167:320-43.

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;

Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology. 1995;141(3):263–272.

Kann L, Kinchen SA, Williams BI, et al. Youth Risk Behavior Surveillance - United States, 1999. MMWR. 2000;49(SS-05):1-96.
(http://www.cdc.gov/mmwr/preview/mmwrhtml/ss4905a1.htm)
Accessed February 2, 2007.

Kann L, Brener ND, Warren CW, Collins JL, Giovino GA. An assessment of the effect of data collection setting on the prevalence of health risk behaviors among adolescents. Journal of adolescent health. 2002;31:327–335.

King, G. Publication, Publication. PS: Political Science and Politics. 2006; 34(1):119–125. at http://gking.harvard.edu/files/abs/paperspub-abs.shtml. Accessed June 12, 2007.

King G, Replication, Replication. PS: Political Science and Politics. 1995; 28(3): 443–499.

Lafay L, Mennen L, Basdevant A, Charles MA, Borys JM, Eschwege E, Romon M. Does energy intake underreporting involve all kinds of food or only specific food items? Results from the Fleurbaix Laventie Ville Sante (FLVS) study. Int J Obes Relat Metab Disord. 2000 Nov;24(11):1500-6.

Lee, S.Y., and W. Y. Poon. 1986. Maximum Likelihood Estimation of Polyserial Correlations. Psychometrika 51: 113-121.

Marquis KH, Duan N, Marquis MS, Polich JM, Meshkoff JE, Schwarzbach DS, Stasz C. Response errors in sensitive topic surveys: Estimates, effects, and correction options. R-2710/2-HHS, RAND, April 1981.

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.

Mendez MA, Wynter S, Wilks R, Forrester T. Under- and overreporting of energy is related to obesity, lifestyle factors and food group intakes in Jamaican adults. Public Health Nutr. 2004 Feb;7(1):9-19.

Mensch BS, Kandel DB. Underreporting of substance use in a national longitudinal youth cohort: individual and interviewer effects. Public Opinion Quarterly. 1988; 52:100–124.

Moxey LM, Sanford AJ. Context effects and the communicative functions of quantifiers: Implications for their use in attitude research. In: Schwarz N, Sudman S, ed. Context effects in social and psychological research. New York, NY: Springer-Verlag; 1992: 279–296.

Muhlheim LS, Allison DB, Heshka S, Heymsfield SB. Do unsuccessful dieters intentionally underreport food intake? Int J Eat Disord. 1998 Nov;24(3):259-66.

Murphy DA, Durako S, Muenz LR, Wilson CM. Marijuana use among HIV-positive and high-risk adolescents: a comparison of self-report through audio computer-assisted self-administered interviewing and urinalysis. Am J Epidemiol . 2000;152(9):805–813.

Nagelkerke NJD, Fidler V, Buwalda M. Instrumental variables in the evaluation of diagnostic test procedures when the true disease state is unknown. Statistics in Medicine. 1988; 7:739–744.

Novotny JA, Rumpler WV, Riddick H, Hebert JR, Rhodes D, Judd JT, Baer DJ, McDowell M, Briefel R. Personality characteristics as predictors of underreporting of energy intake on 24-hour dietary recall interviews. J Am Diet Assoc. 2003 Sep;103(9):1146-51.

O'Malley PM, Johnston LD, Bachman JG. Alcohol use among adolescents. Alcohol Health Res World. 1998;22(2):85-93.

Pearson, K. (1900) Mathematical contribution to the theory of evolution. VII. On the correlation of characters not quantitatively measured. Philosophical Transactions of the Royal Society of

Pedersen W. Reliability of Drug use responses in a longitudinal study. Scandinavian Journal

of Psychology. 1990;31:28–33.

Poulin C. Validity of a province-wide student drug use survey: lessons in design. Canadian Journal of Public Health. 1993;84:259–264.

Reinisch EJ, Bell RM, Ellickson PL. How accurate are adolescent reports of drug use? RAND Note N-3189 CHF. 1991.

Risser JMH, Risser WL, Eissa MA, Cromwell PF, Barratt MS, Bortot A. Self-assessment of circumcision status by adolescents. Am J Epidemiol . 2004;159(11):1095–1011.

Robinson LA, Vander Weg MW, Riedel BW, Klesges RC, McLain-Allen B. 'Start to stop': results of a randomised controlled trial of a smoking cessation programme for teens. Tob Control. 2003 Dec;12 Suppl 4:IV26-33.

Rootman I, Smart RG. A comparison of alcohol, tobacco, and drug use as determined from household and school surveys. Drug and alcohol dependence. 1985;16:89–94.

Rosenbaum JE. Reborn a virgin: adolescents' retracting of virginity pledges and sexual histories. American Journal of Public Health. 2006; 96(6):1098-1103.

Scagliusi FB, Polacow VO, Artioli GG, Benatti FB, Lancha AH Jr. Selective underreporting of energy intake in women: magnitude, determinants, and effect of training. J Am Diet Assoc. 2003 Oct;103(10):1306-13.

Shillington AM, Clapp JD. Self-report stability of adolescent substance use: are there differences for gender, ethnicity and age?. Drug and Alcohol Dependence. 2000;60:19–27.

Siddiqui O, Mott JA, Anderson TL, Flay BR. Characteristics of inconsistent respondents who have "ever used" drugs in a school-based sample. Substance use and misuse. 1999;34:269–295.

Siegel DM, Aten MJ, Roghmann KJ. Self-reported honesty among middle and high school students responding to a sexual behavior questionnaire. Journal of Adolescent Health. 1998;23:20–28.

Sieving R, Hellerstedt W, McNeely C, Fee R, Snyder J, Resnick M. Reliability of self-reported contraceptive use and sexual behaviors among adolescent girls. Journal of sex research. 2005;42(2):159–166.

Smith TW. Nonattitudes: A review and evaluation. In: Turner CF, Martin E, ed. Surveying subjective phenomena, vol 2. New York : Russell Sage Foundation; 1984: 215–255.

Stanton WR, McClelland M, Elwood C, Ferry D, Silva PA. Prevalence, reliability and bias

of adolescents' reports of smoking and quitting. Addiction. 1996;91:1705–1714.

Sussman MP, Jones SE, Wilson TW, Kann L. The Youth Risk Behavior Surveillance System: updating policy and program applications. J Sch Health. 2002 Jan;72(1):13–7.

Tooze JA, Subar AF, Thompson FE, Troiano R, Schatzkin A, Kipnis V. Psychosocial predictors of energy underreporting in a large doubly labeled water study. Am J Clin Nutr. 2004 May;79(5):795-804.

Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL. Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. Science. 1998;280:867–873.

Turner CF. Why do surveys disagree? Some preliminary hypotheses and some disagreeable examples. In: Turner CF, Martin E, ed. Surveying subjective phenomena, vol 2. New York : Russell Sage Foundation; 1984: 159–214.

Uebersax JS, Grove WM. Latent trait finite mixture model for the analysis of rating agreement. Biometrics. 1993;49:823–835.

Upchurch DM, Lillard LA, Aneshensel CS, Li NF. Inconsistencies in reporting the occurrence and timing of first intercourse among adolescents. Journal of Sex Research. 2002;39:197–206.

U.S. Bureau of the Census. School Enrollment–Social and Economic Characteristics of Students (Update): October 1998 Detailed Tables for Current Population Report, P20-521, Tables 1 and 5, Internet Release Date: October 6, 1999.
http://www.census.gov/population/www/socdemo/school/p20-521dt.html
Accessed January 11, 2007.

Walter SD, Irwig LM. Estimation of error rates, disease prevalence, and relative risk misclassified data: A review. Journal of clinical epidemiology 1988; 41:923–37.

Wentland EJ, Smith KW. Survey responses: An evaluation of their validity. San Diego, CA: Academic Press, Inc. 1993.

Winters KC, Stinchfield RD, Henly GA, Schwartz RH. Validity of adolescent self-report of alcohol and other drug involvement. International Journal of the Addictions 1991;25:1379–95.

Tables for Rosenbaum, "Truth or Consequences", September 12, 2007.
DRAFT: DO NOT CITE.

TABLE 1. Demographic characteristics of respondents, compared with nationally-representative sample (Youth Risk Behavior Survey 1999). May not add to 100% due to rounding. Source: Brener 2002 and Kann, Kinchen, et al 2000.

		Sample (n=4619)	YRBS 1999 (n=15,349)
Gender			
	Male	46.6	50.4
	Female	53.4	49.6
Race/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
Grade			
	9	30.6	28.9
	10	31.8	26.0
	11	21.9	23.6
	12	15.7	21.4
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

TABLE 2. Test-retest effect and agreement.
Reliability study of the Youth Risk Behavior Survey administered to 4619 high school students in 61 high schools in 20 geographically diverse states at an interval of two weeks in Feb-Apr 2000. Survey sections appear in the order of the YRBS and are sorted by tetrachoric correlation (TCC), a measure of agreement uncorrelated with prevalence. Survey sections are summarized in the row prior to the questions. p-values are from the McNemar test, indicating presence of a significant retest effect (difference in prevalence at time 2 vs. time 1). The average retest effect appears in each section's summary. Prevalence is the proportion of respondents endorsing at times 1 or 2, expressed as a percent. Absolute retraction is the proportion of the sample giving an affirmative response followed by a negative response; relative retraction is the proportion of the wave 1 affirmative respondents who give a negative response at wave 2. Absolute initiation is the proportion of the sample giving a negative response followed by an affirmative response; relative initiation is the proportion of the wave 2 affirmative respondents who answered in the negative at wave 1. Tetrachoric correlation (TCC) measures average agreement between wave 1 and wave 2 responses: 0.0 is chance agreement and 1.0 perfect agreement. Standard error of TCC is computed for aggregated data from all questions in a category under assumption of uniform TCC within category. Standard error (se) multiplier is the estimated increase in standard error due to inconsistency estimated with a Bayesian method.
p ≤ 0.1, ** p ≤ 0.01, *** p ≤ 0.001, **** p ≤ 0.0001

		Prevalence			Retraction		Initiation			se
Survey section		t1	t2	p	abs	rel	abs	rel	TCC	mult
									0.93	3.8
Traffic				2.7					(0.003)	
	Bike helmet, rarely/never	84.6	83.8	*	3.6	4.3	2.8	3.4	0.94	2.4
	Motorcycle helmet, rarely/never	37.8	46.8	****	3.8	10.1	12.8	27.4	0.89	5.6
	Seatbelt, rarely/never	15.7	19.6	****	3.6	23.2	7.6	38.5	0.86	4.5
	Drove after drank	8.5	10.3	****	2.8	32.3	4.6	44.1	0.85	3.9
	Rode w drinking driver	30.3	29.6		8.7	28.6	8.0	27.0	0.82	2.8
									0.88	
Violence				-0.4					(0.004)	3.8
	Forced sex	9.1	10.3	**	2.4	26.4	3.6	34.9	0.91	3.4
	Carried weapon	15.0	13.3	****	5.0	33.5	3.3	24.9	0.89	3.5
	Physical fight	34.6	30.3	****	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	*	1.9	36.4	2.5	43.2	0.88	3.5
	Carried gun	4.2	4.4		1.9	45.9	2.1	48.3	0.84	3.6
	Injured in fight	2.9	4.4	****	1.1	38.8	2.6	59.6	0.83	4.8
	Injured by s.o.	9.1	9.9	*	3.6	39.4	4.4	44.3	0.82	3.6
	Feel unsafe at school	5.5	5.0		3.1	57.1	2.6	52.8	0.76	4.3
	Threatened w weapon at school	7.3	5.9	***	4.4	59.8	3.0	50.2	0.73	4.2

Survey section	Prevalence t1	Prevalence t2	p	Retraction abs	Retraction rel	Initiation abs	Initiation rel	TCC	se mult
Depression			-0.9					0.90 (0.004)	3.0
Considered suicide	17.0	16.0	*	4.1	23.8	3.0	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	13.0	12.9		3.8	29.3	3.7	28.9	0.90	2.7
Injured in suicide attempt	2.1	2.7	*	0.8	39.2	1.4	52.4	0.87	3.4
Felt sad/hopeless	28.2	24.1	****	10.5	37.2	6.4	26.5	0.80	4.1
Tobacco			0.1					0.96 (0.001)	2.8
Ever try smoking	65.8	63.9	****	4.2	6.4	2.3	3.6	0.98	2.1
Smoke pack per day	17.5	17.1		2.6	14.6	2.2	12.7	0.97	2.0
Smoked past month	27.2	27.5		3.4	12.7	3.8	13.6	0.96	2.0
Ever smoke regularly	17.7	19.0	***	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	0.94	2.7
Bought cigarettes	6.4	7.2	**	1.5	24.1	2.3	32.5	0.93	2.9
Smoke at school	9.7	9.1	*	2.7	28.1	2.1	23.5	0.93	2.6
Tried to quit smoking	18.4	16.7	****	5.2	28.0	3.4	20.6	0.92	3.0
Smoked before age 13	21.4	23.7	****	3.9	18.4	6.2	26.3	0.91	3.0
Chewed tobacco at school	3.9	3.9		1.5	38.1	1.5	38.1	0.90	3.0
Ever smoke cigars	12.2	11.8		4.5	36.5	4.1	34.3	0.86	3.1
Bought cig and carded	6.8	8.2	***	2.6	37.9	4.0	48.6	0.83	4.1
No usual cigarette brand	1.6	1.5		1.0	63.5	0.9	60.9	0.78	3.4
Alcohol			-0.8					0.94 (0.002)	2.9
Ever drink alcohol	76.1	72.5	****	5.3	6.9	1.7	2.3	0.97	2.8
Alcohol past month	41.1	39.9	*	7.6	18.5	6.4	16.1	0.90	2.5
Binge drink past month	23.9	23.7		6.0	25.0	5.8	24.4	0.89	2.5
Alcohol before age 13	28.9	29.9	*	6.6	22.8	7.6	25.3	0.87	2.7
Alcohol at school	3.9	4.1		1.8	47.2	2.0	49.7	0.83	3.9
Illegal drugs			0.1					0.94 (0.002)	3.0
Ever marijuana	42.8	41.7	**	3.0	7.1	2.0	4.7	0.99	1.7
Ever cocaine	5.6	6.2	*	1.2	20.9	1.8	28.7	0.95	2.5
Marijuana past month	22.6	22.1		4.4	19.5	3.9	17.7	0.94	2.2
Ever methamphetamines	6.3	6.9	*	1.5	24.1	2.1	30.5	0.94	2.7
Marijuana bef age 13	10.5	11.3	*	2.5	23.7	3.3	29.1	0.93	2.8

Survey section		Prevalence t1	t2	p	Retraction abs	rel	Initiation abs	rel	TCC	se mult
	Ever inhalants	11.3	10.6	*	3.6	31.6	2.9	27.0	0.91	2.9
	Ever heroin	1.9	3.0	****	0.5	25.0	1.6	52.2	0.91	2.7
	Ever inject illegal drug	1.4	2.0	**	0.5	33.9	1.1	53.3	0.90	2.7
	Marijuana at school	5.5	5.3		2.2	39.8	2.0	37.6	0.88	3.3
	Cocaine past month	2.2	2.7	*	1.0	45.1	1.5	55.2	0.84	3.8
	Ever steroids	4.0	4.1		2.1	51.9	2.2	53.2	0.80	4.1
	Inhalants past month	2.9	3.5	*	1.5	51.5	2.1	59.9	0.79	4.1
	Ever been sold drugs at school	23.0	21.9	*	8.9	38.6	7.8	35.5	0.76	3.5
Sex				-0.5					0.91 (0.003)	3.2
	Ever sexual intercourse	49.5	50.2	*	2.0	4.1	2.7	5.4	0.99	1.6
	Sex past 3 months	32.9	35.0	****	5.1	15.4	7.2	20.5	0.91	2.7
	4 or more sex partners	19.1	17.6	**	7.1	37.0	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	18.0	14.8	****	9.8	54.4	6.6	44.5	0.66	4.7
Weight, physical activity										
		PE class weekly		-1.5					0.84 (0.002)	3.7
	Consider self overweight	62.4	56.8	****	6.5	10.4	0.9	1.5	0.98	2.6
	Try lose weight	22.7	26.1	****	6.0	26.2	9.4	35.8	0.82	3.9
	Exercise to lose weight	33.8	37.2	****	7.9	23.3	11.3	30.3	0.80	3.6
	Sports team	58.6	53.9	****	12.8	21.9	8.1	15.0	0.79	4.0
	Diet to lose wt	54.6	53.3	*	11.5	21.1	10.2	19.2	0.77	2.5
	Vomit to lose wt	43.1	40.4	****	12.7	29.6	10.1	24.9	0.75	3.5
	Diet pills to lose wt	4.9	5.0		2.8	56.2	2.9	57.1	0.74	3.7
	Sports injury	7.8	7.9		4.1	53.1	4.2	53.7	0.73	4.0
	TV less than 2 hrs/day	40.8	35.2	****	15.3	37.5	9.7	27.6	0.69	4.7
	Fasted to lose wt	62.4	63.2		12.1	19.3	12.9	20.3	0.68	3.4
	PE: exercise > 20 minutes	18.4	15.3	****	10.0	54.1	6.9	44.8	0.66	4.9
Misc		72.3	69.0	****	13.9	19.2	10.6	15.3	0.63	4.2
		Saw dentist		-0.7					0.71	3.6
	Sunscreen rarely/never	66.5	63.4	****	9.8	14.8	6.7	10.6	0.85	3.3
	Saw doctor/nurse when not sick	66.6	66.7		8.6	12.9	8.7	13.1	0.83	2.7
	Taught about HIV	58.9	58.1		12.4	21.1	11.6	20.0	0.72	3.1

Survey section	Prevalence t1	Prevalence t2	p	Retraction abs	Retraction rel	Initiation abs	Initiation rel	TCC	se mult
	85.0	86.2	*	8.8	10.4	10.0	11.6	0.45	5.1

TABLE 3: Agreement (TCC) by time-frame: TCC (standard error). Standard error computed using the 2-step estimator when it can be computed.

	Ever	Past 30 days. (sex: 3 mos)	Before age 13	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	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)	-	-

TABLE 4. Question characteristics associated with Tetrachoric Correlation (TCC) in linear regression. TCC is measured on a scale from 0.0 chance agreement to 1.0 perfect agreement. R^2=0.32.
. p ≤ 0.1, * p ≤ .05, ** p ≤ 0.01, *** p ≤ 0.001, **** p ≤ 0.0001

Factor	coefficient	se	p
Weight control	-0.077	0.034	*
Sexual intercourse	0.133	0.062	*
Tobacco	0.096	0.029	**
Illegal drugs	0.066	0.026	*
Alcohol	0.075	0.041	.