Effects of Validity Screening Items on Adolescent Survey Data

Dewey Cornell

Jennifer Klein

Tim Konold

Francis Huang

Curry School of Education, University of Virginia

In press, *Psychological Assessment*

Author Note

June 17, 2011. Dewey Cornell is Professor of Education and Director of the Virginia Youth Violence Project in the Curry School of Education, University of Virginia. Jennifer Klein is a doctoral student in the Programs in Clinical and School Psychology, Curry School of Education. Tim Konold is a professor in and Program Coordinator for Research, Statistics, and Evaluation in the Curry School of Education. Francis Huang is a senior scientist at the Phonological Awareness Literacy Screening office at the Curry School of Education.

We thank June Jenkins and Christopher Gilman for their work through the Safe Schools/Healthy Students Albermarle/Charlottesville Project. We also thank Donna Bowman Michaelis of the Virginia Department of Criminal Justice Services and Arlene Cundiff of the Virginia Department of Education, and their colleagues, for their support of the Virginia High School Safety Study. We thank Michael Hull for statistical consultation on this study. We also thank the research assistants for the Virginia Youth Violence Project at the time of this project, including Michael Baly, Victoria Phillips, Anna Lacey, and Erin Nekvasil. This project was supported in part by grants from the Federal Safe Schools/Healthy Students Initiative (sponsored by the U.S. Departments of Education, Health and Human Services, and Justice) and the Office of Juvenile Justice and Delinquency Prevention of the U.S. Department of Justice, but the views in this article do not necessarily reflect policies or recommendations of the funding agencies.

Correspondence concerning this article should be addressed to Dewey Cornell, Curry School of Education, University of Virginia, P.O. Box 400270, Charlottesville, VA 22904-4270. Dcornell@virginia.edu

Abstract

Two studies examined the use of validity screening items in adolescent survey data. In each study, adolescent respondents were asked whether they were telling the truth and paying attention in answering survey questions. In Study 1 (N = 7,801), the prevalence rates of student risk behaviors were significantly lower after inappropriate ("invalid") responders were screened out of the sample. In addition, confirmatory and multi-group factor analyses demonstrated significant differences between the factor structures of school climate scales using valid versus invalid responders. In Study 2, student perceptions of school climate were correlated with teacher perceptions in 291 schools. A bootstrap resampling procedure compared the correlations obtained using valid versus invalid responding students in each school and found that valid responders had more positive views of school conditions and produced higher correlations with teacher perceptions. These findings support the value of validity screening items to improve the quality of adolescent survey data.

Keywords: adolescent risk behavior, school surveys, self-report, validity screening

Effects of Validity Screening Items on Adolescent Survey Data

Researchers rely heavily on adolescent self-report surveys for a variety of psychological assessment purposes. The prevalence rates for self-reported adolescent drug use are widely reported as indicators of national trends (Johnston, O'Malley, Bachman, & Schulenberg, 2010) and student self-reports of fighting, weapon-carrying, and other concerns are used to guide national school safety and discipline practices (Robers, Zhang, & Truman, 2010). School-based prevention programs make extensive use of student surveys to evaluate the effectiveness of their prevention efforts (Sharkey, Furlong, & Yetter, 2006).

Although there have been studies concerned with the potential for under-reporting of drug use and delinquent behavior (e.g., Hindelang, Hirschi, & Weiss, 1979), there is surprisingly little research on the problem of over-reporting (Fan et al., 2006; Sharkey et. al, 2006). Adolescents, because of immaturity and rebelliousness, may be tempted to offer inflated reports of their engagement in socially proscribed or illicit behaviors. Or, they may not take the survey seriously and mark it haphazardly, producing an elevation in otherwise low base rate behaviors. The purpose of this study was to bring attention to the relatively neglected problem of adolescent over-reporting, to demonstrate its impact on survey results, and to present evidence in support of validity screening items as a means of improving survey validity.

National Adolescent Student Health Survey. Uncritical acceptance of data from student self-report surveys can have widespread impact on the general public as well as professional and policy-making organizations. A most egregious example was the 1987

National Adolescent Student Health Survey (American School Health Association, 1989), which asked 11,419 8th and 10th grade students, "Think back over the last 12 months. While at school, how often did you carry a handgun?" Students could respond "*Never*," "*Less than once a month*," "*A few times a month*," "*A few times a week*," or "*Nearly every day*." Approximately 2.6% of boys reported that they brought a gun to school, including .8 percent who claimed to have brought a handgun to school "nearly every day." How many adolescent boys might have been tempted to answer such a question in a provocative manner and how many others might have simply marked the wrong answer by accident? Despite these potential problems, the survey results were extrapolated to the sensational conclusion that 135,000 guns are brought to school every day in the United States, which was reported first in *U.S. News & World Report* (Witkin, 1991) and subsequently many other news reports (Cornell, 2006).

Over the next decade, the claim that "135,000 guns are brought to school every day" was cited by numerous professional organizations and advocacy groups, including the National School Boards Association (1993), the Family Research Council (Maginnis, 1995), the American Sociological Association (Levine & Rosich, 1996), and the National Crime Prevention Council (1995). In 2011, a Google search of "135,000 guns" generated about 176,000 results, including archived reports by the National Education Association (2005) and the U.S. Office of Juvenile Justice and Delinquency Prevention (OJJDP; n.d.), as well as articles published by PBS (Public Broadcasting Service; Singer, 2005) and *The New York Times* (Celis, 1994).

Youth Risk Behavior Surveillance System. Perhaps the most widely used student survey in the United States is the CDC's Youth Risk Behavior Survey (YRBS), which is administered in thousands of schools each year. This survey asks students questions about substance use, weapon carrying, fighting, and suicidal thoughts, among other topics (U.S. Department of Health and Human Services, 2009). A summary report of the methodology of this survey acknowledged that self reports could be affected by "cognitive and situational factors" and that "each type of behavior differs in the extent to which it can be validated by an objective measure," but did not identify any studies examining the accuracy of its risk behavior questions (Brener, McManus, Galuska, Lowry, & Wechsler, 2004, p 5). According to this report, the only direct attempt to confirm student self-reports on the YRBS involved student height and weight. This study found that students, on average, over-reported their height by 2.7 inches and under-reported their weight by 3.5 pounds (Brener et al, 2003).

Even without direct studies of student accuracy, there is reason to be suspicious of over-reporting on the YRBS. Furlong and colleagues (2004) identified a group of 414 respondents in one wave of the YRBS who claimed to have carried a weapon to school 6 or more times in the past month (the most extreme response). Many of these weapon-carrying students claimed to make frequent suicide attempts, use heroin, sniff glue, and take steroids, but also incongruously claimed to exercise every day, eat plenty of carrots, and drink lots of milk. The researchers concluded that these students gave extreme responses to survey questions regardless of item content and that the "presence of this

response bias may inflate estimates of the prevalence of school violence and related concerns (p 110)."

The YRBS makes a limited effort to identify implausible response patterns, primarily involving the detection of inconsistent responses to pairs of similar items. Responses that conflict in logical terms are both set to missing. For example, if a student responds to one question that he or she has never smoked but then responds to a subsequent question that he or she has smoked two cigarettes in the previous 30 days, the processing system sets both responses to missing, but retains the rest of the survey. Only questionnaires with fewer than 20 valid responses remaining after editing or with the same answer to 15 or more questions in a row are deleted from the dataset. In 2009, a total of only 50 questionnaires from a national survey of 16,460 students failed these quality-control checks and were excluded from analysis (CDC, 2010).

Add Health Survey. Although most adolescent surveys are anonymous, the well-known National Study of Adolescent Health (Add Health) survey was administered on a confidential basis that made it possible to check the accuracy of school survey responses against other sources of information. The Add Health survey is considered the largest and most comprehensive longitudinal survey of adolescents ever undertaken and has generated more than 3,800 articles and reports (Carolina Population Center, n.d.).

Some researchers have found substantial discrepancies in Add Health survey results. A study of adoptees in the Add Health sample by Miller et al. (2000) found that adopted adolescents reported consistently higher rates of smoking, drinking, skipping school, fighting, lying to parents, and other problematic behavior in comparison to non-

adoptees. However, when the researchers subsequently examined in-home interviews for these same students, they found that about 19% of the adolescents who claimed to be adopted were not adopted (Fan et al., 2002). When the data were reanalyzed, the group differences diminished or disappeared. It is noteworthy that even a low rate of over-reporting could produce statistically significant group differences and false findings.

Fan and colleagues (2006) later identified other evidence of adolescent over-reporting on the ADD Health survey again by comparing student reports on the school survey with information obtained during subsequent face-to-face in-home interviews and interviews with parents. Among their findings were that 20 percent (176 of 863) of adolescents who claimed on the school survey to have been born outside the United States later acknowledged that they were born in the United States in a face-to-face home interview (and a parent confirmed that they were born in the United States). Another group of 253 adolescents claimed on the survey to have used an artificial limb for the past year or more, indicating a permanent physical disability. When interviewed at home, only two of these 253 adolescents could be confirmed as using an artificial limb, with 248 retracting their claim and three not answering the question.

These researchers then examined the other survey responses of the adolescents who either misrepresented their adoption status, their nation of birth, or their disability status. In every comparison between inaccurate responders and truthful responders on behavioral items, the inaccurate responders reported higher rates of problem behaviors (such as drinking, skipping school, and fighting) as well as lower scores on positive variables (such as self-esteem).

Fan and colleagues (2006) noted that it is difficult to distinguish between respondents who are intentionally misrepresenting themselves or simply answering in a careless or haphazard manner, but that both groups produce results with similar problems: over-reporting of otherwise low base-rate behaviors and attitudes. They concluded that even if the proportion of over-reporting students is small relative to the total sample, these respondents could have a substantial impact on subgroup analyses. They noted that conventional validity scales for social desirability were not adequate for this problem and recommended that future surveys include a validity scale or some other screening procedure.

Use of Validity Screening Items

One simple strategy for detecting over-reporting is to ask students directly whether they are answering carefully and truthfully. Such an approach would not detect students who chose not to reveal that they were answering inappropriately, but nevertheless could detect a sizable number of students who are either answering in a careless manner (thus marking an inappropriate answer by accident) or who are willing to admit that they are not answering honestly.

In a survey of 10,909 middle and high school students, Cornell and Loper (1998) employed two validity questions: “I am reading this survey carefully” and “I am telling the truth on this survey.” Approximately 7.8% (933) students marked one or both of these items inappropriately (by responding “no”). These students also tended to endorse fighting, carrying a gun, drinking, and using illegal drugs at rates that were three to five times higher than students who answered the validity items appropriately.

In a sample of 6,189 middle and high school students, Rosenblatt and Furlong (1997) assessed response consistency by comparing student responses to two similar questions about feelings of personal safety at school and screened for validity in the form of an improbable question ("I took ten field trips in the previous month."). Students who failed either the consistency or validity checks were compared with a randomly selected matched group of students who answered consistently and appropriately. Students who failed the checks reported dramatically higher rates of violence in their schools (15%) than students who passed (2%). It was also noted that many screened responders showed systematic bias toward portraying their school in a negative light.

The California Healthy Kids Survey (2010) is administered throughout California schools, with results from approximately 745,000 secondary school students for the 2007-09 reporting period. The survey includes one honesty question ("How many questions in this survey did you answer honestly?") that is used in combination with other survey checks (e.g., a pattern of extreme or inconsistent responding) to identify invalid surveys. However, there is no published or unpublished report on the use of this question and how it affects survey results (Jerry Bailey, personal communication, January 6, 2011). There is a need to assess the impact of validity screening items on survey results and determine whether it might be a useful, cost-effective way to improve the accuracy of student self-report surveys.

Present study

There is a body of evidence that even a small proportion of students who intentionally misrepresent themselves can have a distorting effect on survey results

(Cornell, 2006; Cornell & Loper, 1998; Fan et al., 2006; Furlong, Sharkey, Bates, & Smith, 2004). The purpose of this study is to demonstrate the value of validity screening items to help improve the accuracy of student self report surveys. We contend that use of validity screening items can lower inflated estimates of the prevalence of high-risk behavior and provide data that are more reliable and valid for clinical, research, and program evaluation purposes. In support of this contention, Study One demonstrates that reductions in high-risk behavior rates are observed when surveys that contain inappropriate responses to validity items are removed from the sample. Study One also shows differences in the factor structure of school climate scales based on valid versus invalid respondents. Study Two finds differences in the criterion-related validity of student reports of school conditions based on a comparison of valid respondents versus invalid respondents.

Study One

Methods

Participants

The School Climate Bullying Survey (SCBS; Cornell, 2011) was administered in fall 2010 to 7,801 students in six middle schools and five high schools. The schools were located in two adjacent public school systems serving a small city and surrounding suburban and rural county in central Virginia. The percentage of students eligible for a free or reduced price meal in each school ranged from 10 to 52 with a mean of 28 (SD = 14). The sample included 3,854 (49%) males and 3,947 (51%) females, including 1,188 (15.2%) sixth graders, 1,122 (14.4%) seventh graders, 1,091 (14.0%) eighth graders,

1,167 (15.0%) ninth graders, 1,101 (14.1%) tenth graders, 1,096 (14.0%) eleventh graders, and 1,036 (13.3%) twelfth graders. Participants ranged from 10 to 19 years of age with a mean age of 14 years (*SD* = 2.12).

There were 4,705 (60.3%) students who identified themselves as White, 1,206 (15.5%) as Black, 417 (5.4%) as Multiracial, 311 (5.1%), as Asian, 340 (4.4%) as American Indian/Alaskan, and 1,007 (12.9%) as Other. In a separate demographic item, 765 (9.8%) students specifically identified themselves as Hispanic. Information on the socioeconomic status of individual students was not available.

A total of 7,801 students (91% of total enrolled) completed surveys on the survey administration days at the participating middle and high schools. Although the survey was administered on an anonymous basis, a letter was sent home to parents that allowed them to refuse participation by their child. Approximately 40 parents opted their child out of the survey and approximately 91 students were unavailable to take the survey due to illness or long-term suspension.

Measures

Students completed the School Climate Bullying Survey (Cornell, 2011), a 45-item self-report instrument that collects demographic information, reports of involvement in bullying as an aggressor or victim, and three school climate scales. The SCBS was selected for use in this project because it contains measures of bullying and school climate that have generated scores with favorable psychometric properties (Bandyopadhyay, Cornell, & Konold, 2009; Branson & Cornell, 2009).

To assess bullying, students were presented with a standard definition of bullying, "Bullying is defined as the use of one's strength or status to injure, threaten, or humiliate another person. Bullying can be physical, verbal, or social. It is *not* bullying when two students of about the same strength argue or fight." Bully victimization was assessed with the item, "I have been bullied at school in the past month" and bullying others was assessed with the item, "I have bullied others at school in the past month". There were four response options for both questions (*never, once or twice, about once a week,* or *several times per week*). Consistent with recommendations by Solberg and Olweus (2003) to use a threshold of approximately once per week, frequencies of "*about once a week*" or "*several times per week*" were classified as involvement in bullying. These two items have been found to correspond with independent measures obtained from peer nominations and teacher nominations (Branson & Cornell, 2009; Cornell & Brockenbrough, 2004). Self-reports of victimization were also correlated with depression, negative perceptions of school, and lower academic performance, whereas self-reports of bullying others were correlated with aggressive attitudes, discipline referrals, and suspensions from school (Branson & Cornell, 2009). The SCBS produced estimates of the prevalence of bullying victimization and bullying others that are similar to the Olweus Bullying Victimization Questionnaire (Cornell, 2011).

In order to meet federal grant reporting requirements, the SCBS was augmented with ten risk behavior items derived from the Youth Risk Behavior Surveillance System (CDC, 2010). These items are used nationwide to assess the prevalence of student risk behavior. The items (see Table 1) had answer choices ranging from either "*0 days*" to

"*20-30 days*" or "*0 times*" to "*6 or more times*". For purposes of this study, each item was dichotomized to distinguish no use from any use.

The 20-item school climate portion of the SCBS consisted of a seven-item Aggressive Attitudes scale, a four-item Prevalence of Bullying and Teasing scale, and a nine-item Willingness to Seek Help scale (see Table 2). Previous exploratory and confirmatory factor analyses in a middle school sample (n = 2,111) supported the factor structure of these three school climate scales (Bandyopadhyay et al., 2009). Further analyses with a sample of 7,318 ninth grade students found that these scales were predictive of teacher reports of bullying and teasing, teacher reports of student help-seeking behaviors, teacher reports of gang-related violence, and school records of suspensions and expulsions (Bandyopadhyay et al., 2009).

The nine items on the Willingness to Seek Help scale were reverse coded to align with the direction of the other SCBS items. Measures of internal consistency (Cronbach's alphas) for scores from the three scales in the current sample were: Prevalence of Bullying and Teasing scale ($\alpha = .75$), Aggressive Attitudes scale ($\alpha = .88$), and Willingness to Seek Help scale ($\alpha = .87$). The school climate survey included three screening validity items: (1) "I am telling the truth on this survey," (2) "I am not paying attention to how I answer this survey," and (3) "The answers I have given on this survey are true." The first two items both had four Likert-type answer choices ranging from "*Strong disagree*" to "*Strongly agree*." The items were then dichotomized into those students who either disagree or agree. The third item allowed students to answer either "*No*" or "*Yes*." The intercorrelations among these three items were r = -.33 between items

(1) and (2); .31 between items (1) and (3), and -.20 between (2) and (3); all $p < .01$. Although the three items were not combined into a scale, their internal consistency (alpha) was .65. Students who endorsed either not telling the truth or not paying attention were compared with those students appropriately endorsed all three validity screening items.

Procedure

Students completed the school climate survey online in classrooms under teacher or staff supervision. Students listened to a standard series of directions and then answered questions anonymously. A Spanish translation of the survey was also made available online. Data were provided to the researchers in archival form.

Results

The first phase of data analysis consisted of developing a procedure for validity screening to establish how many students responded inappropriately to the three validity-screening items. Four contrasting variables were created that identified students as invalid responders if they endorsed either (1) not telling the truth on the survey, (2) not giving true answers, (3) not paying attention in answering questions, or (4) any of these three items.

The effect of using each of these validity items singly, or in combination, to identify invalid survey respondents was contrasted with risk groups (endorsed/did not endorse the risk behavior) using chi-square tests of association. Table 3 presents the association between validity items and each risk behavior. Separate analyses were conducted for each of the three validity items and the ten risk items, generating a total of

30 comparisons, of which 29 were statistically significant at $p < .05$. Phi statistic effect sizes for the statistically significant comparisons ranged from .01 to .16. A fourth set of analyses was conducted using the variable that combined responses to the three validity items. All ten of these comparisons were statistically significant, with effect sizes ranging from .03 to .20. Invalid responders endorsed significantly higher rates of risk behaviors than valid responders across all analyses. For example, 28.0% of invalid responders endorsed marijuana use compared to 10.8% of valid responders.

Table 1 compares the prevalence rates of the ten high-risk behaviors for the total sample (N = 7,801) with the sample after validity screening (N = 6,883). Results reveal that students consistently endorsed higher percentages of risk behaviors before validity screening, leading to inflated prevalence rates. Several items had a particularly high rate of inflation, such as weapon carrying (38.6%), bullying (29.7%), feeling unsafe (28.9%), and smoking cigarettes (23.1%) or marijuana (18.5%). The estimated prevalence of students feeling sad or hopeless had the lowest rate of inflation (2.0%).

Second phase. The purpose of the second phase of data analysis was to evaluate whether or not the factor structures of the three SCBS scales were similar across valid and invalid responders. We hypothesized that they would differ across groups because the exaggeration tendency among invalid responders would alter the relationships among items and factors. Confirmatory factor analysis (CFA) was employed to evaluate the school climate portion of the SCBS measurement structure in the screened sample of valid responders. We then examined variations in the measurement properties of the

SCBS between groups of valid and invalid responders through multi-group confirmatory factor analysis (MGCFA).

A graphic representation of the 20-item three factor model hypothesized to fit within this sample that is based on previous theoretical and empirical work (Bandyopadhyay et al., 2009; Cornell, 2011) is presented in Figure 1. Directly measured and observed SCBS items are enclosed in boxes to differentiate them from the estimated latent factors and uniqueness terms. Each observed variable was modeled to be directly influenced by its respective factor as illustrated by single-headed arrows. Factor correlations were freely estimated as depicted by curved, double-headed arrows. Parameterization of the model included scaling the factors to one of the observed variables by fixing a single factor loading to unity for each factor. Curved double-headed arrows connecting a limited number of uniqueness terms indicating shared variance between items not accounted for by the hypothesized factors were not specified in advance of testing this model. Rather, they emerged throughout the model testing process as important model relaxations that materially contributed to improved model fit. Theoretical justification for their estimation is provided below.

Numerous measures of model fit exist for evaluating the quality of measurement models within a structural equation modeling framework, with many focusing on different components of fit (Browne & Cudeck, 1993; Hu & Bentler, 1995). As a result, it is generally recommended that multiple indices be evaluated to highlight different aspects of model fit (Tanaka, 1993). As a stand-alone measure of fit, chi-square (χ^2) is known to reject reasonably specified models when estimated on large samples (Hu & Bentler,

1995; Kaplan, 1990; Kline, 2005). Consequently, use of this statistics was limited to evaluating competing models between valid and invalid groups through chi-square difference tests (χ^2_D) in the context of the MGCFA analyses.

Several additional measures of fit were considered in evaluating stand-alone model quality. These included the Bentler-Bonett normed fit index (NFI), Tucker-Lewis index (TLI), comparative fit index (CFI), and root mean square error of approximation (RMSEA). The first three measures generally range between 0 and 1.0. Traditionally, values of .90 or greater were taken as evidence of good fitting models (Bentler & Bonett, 1980), with more recent research suggests that better fitting models produce values around .95 (Hu & Bentler, 1999). Others consider these thresholds overly restrictive and may result in a decision to reject otherwise good fitting models, particularly when item-level data are considered (Marsh, Hau, & Wen, 2004). In contrast, smaller RMSEA values are reflective of better fitting models, with values of .08 or less generally typically indicating reasonable fit (Browne & Cudeck, 1993).

All models were estimated with the Analysis of Moment Structures (AMOS; Arbuckle, 2007) program, and full information maximum likelihood estimation was employed to accommodate missing data. Multivariate outliers were identified through measures of Mahalanobis distance ($p < .001$) within both the valid and invalid samples. Approximately 1% of the most offending cases in each of the two groups were removed, resulting in $N = 6{,}814$ members of the valid group and $N = 909$ members of the invalid group available for analysis. In the aggregate, a total of 78 cases were removed from the total sample. The preliminary CFA results based on members of the valid group for the

model depicted in Figure 1, without correlated error terms, failed to reveal evidence of good fit (NFI = .87, TLI = .84, CFI = .87, and RMSEA = .08).

Inspection of modification indices revealed four theoretically justifiable correlated error term constraints that could be relaxed in order to improve upon this condition. Two pairs of items appeared to be uniquely associated beyond the first factor due to similar item content. The content of items 2 and 8 referred to efforts to stop bullying, and items 3 and 5 asked students whether they would report a classmate who threatened homicidal violence. Two other pairs of items appeared to be uniquely associated beyond the second factor due to similar item content. Items 11 and 13 both referred to the perceived social status achieved by fighting and items 15 and 16 offered excuses for bullying. Estimation of these associations resulted in a much improved and reasonable level of model fit (NFI = .93, TLI = .91, CFI = .93, and RMSEA = .06). Moreover, model parameter estimates were well within expectation with moderate to large factor loadings, and moderate factor correlations ranging from .35 - .59.

Configural invariance was investigated to determine whether the model characteristics illustrated in Figure 1 held across valid and invalid groups. Results of a multi-group confirmatory factor analyses with no cross group equality constraints on parameter estimates revealed good fit (NFI = .92, TLI = .90, CFI = .92, and RMSEA = .04), indicating that item factor alignment was reasonable for these two groups. A comparison of this model (χ^2 (326) = 5175.45) with between group equality constraints imposed on the factor loadings (χ^2 (343) = 5288.05) resulted in a statistically significant decline in fit, χ^2_D (17) = 112.6, $p < .05$, indicating that at least some of the items were

differentially related to the factors for these two groups. It is also worth noting that a similar multi-group model comparison on the unmodified form of the measurement model (i.e., without correlated errors) similarly revealed a statistically significant decline in fit between the multi-group general from model (χ^2 (334) = 8490.90) and a model in which the factor weights were constrained to be equal across groups (χ^2 (351) = 8618.50); χ^2_D (17) = 127.6, $p < .05$.

An investigation of partial metric invariance of the factor loading was conducted to determine the number of between-group differences that existed within this sample. Here, factor loadings were constrained to be equal between groups in turn. If a factor loading was identified as being invariant between groups, the constraint was left in place as the remaining items were examined. In instances where a factor loading was not found to be invariant between groups, the parameter estimate was allowed to be freely estimated between groups throughout examination of the remaining items. This iterative process resulted in the identification of ten items that failed to measure their respective factor with the same degree of accuracy across groups. Table 2 shows standardized factor loadings that were statistically indistinguishable between groups (denoted by ‘ = ‘) along with the separate estimates that were obtained when equality constraints were found to result in a statistically poorer fitting model (denoted by an absence of ‘ = ‘). The resulting model with these partial constraints (χ^2 (333) = 5187.58) was statistically indistinguishable from the general form configural model in which no parameter constraints on the factor loadings were imposed, χ^2_D (7) = 12.13, $p > .05$.

Between group invariance across factor correlations was also examined. A comparison between the partially constrained factor loading model and a multi-group model with further constraints across the three factor correlations (χ^2 (336) = 5809.65) revealed a statistically significant decline in fit, χ^2_D (3) = 622.07, $p < .05$. Evaluation of sequential constraints across factor correlations, similar to the approach used to evaluate factor loadings, failed to reveal any of the three correlations to be statistically equal between groups. Table 2 shows the resulting factor correlations for the two groups.

Study Two

Participants

Student and teacher data were obtained from the Virginia High School Safety Study (VHSSS; Cornell & Gregory, 2008), a statewide assessment of school climate and safety conditions in Virginia public high schools. All Virginia public high schools were eligible for this study (except for several schools that served only grades 10-12 or provided supplementary services to other high schools and did not award a diploma). Of the 314 eligible high schools, 289 (92%) submitted student surveys and 284 (90%) submitted teacher surveys, resulting in a final sample of 284 schools with both teacher and student data. The percentage of students eligible for a free or reduced price meal in each school ranged from 1 to 83 with a mean of 30 (*SD* = 16).

Student Sample. Ninth grade was selected for study primarily because it is the first year of high school (permitting future longitudinal study as the cohort proceeds through high school) and is a pivotal year for student adjustment and achievement (Donegan, 2008). Ninth grade students were not included in the study if (1) they did not

read English well enough to complete the survey; (2) they had cognitive or physical limitations that prevented them from completing the survey.

Principals were asked to identify enough ninth grade students in order to gain a sample of 25 student surveys from each school. A target sample of 25 represented the size of a typical classroom of students that could be tested on one occasion and is consistent with several national studies of student performance, including the National Educational Longitudinal Study, NELS:88 (Ingels, 1992) and the National Assessment of Educational Progress program (Chromy, 1998).

In order to obtain a reasonably representative sample, the principals were instructed to choose students from an alphabetized student roster using random numbers. In the event that students were absent or for some other reason unavailable to complete the survey, principals selected an alternate using the next random number available. In order to standardize recruitment, principals were sent written instructions, form letters to use in inviting survey participants, and a random number list calibrated to the number of 9^{th} grade students in their school.

After the surveys were completed, participation questionnaires were obtained from 291 school principals. Principals reported that approximately 73% of the students initially identified by the sampling procedure participated in the study. The reasons for nonparticipation among the other 27% (1,983 students) included: student absence due to illness (32 percent of those who did not participate); student declined to participate (16 percent), student moved or transferred (7 percent), parent declined (6 percent), student

suspended from school (5 percent), student language barrier (3 percent), or some other reason (this ranged from a severe disability to attending a field trip; 30 percent).

The final student sample consisted of 7,246 9th grade students with an average of 25 students from each school. Approximately 49% of the students were girls and the mean age was 14.8 (*SD* = 0.70), with a range of 13 to 17 years. The self-reported racial/ethnic distribution of the sample was 63% White/Caucasian, 22% Black/African American, 5% Latino/Hispanic, 3% Asian American, 1% American Indian, and 5% Other. Information on the socioeconomic status of individual students was not available.

Teacher Sample. The teacher sample was selected using a similar random number list procedure. Principals were asked to identify enough ninth grade teachers in order to gain a sample of 10 teachers per school. Schools with fewer than ten ninth-grade teachers were encouraged to have all available ninth-grade teachers complete the survey. The estimated completion rate among teachers was 83%. According to principal reports, there were 163 teachers who declined to participate, 140 who were absent the week of administration and another 162 who, for unknown reasons, did not complete the survey. The final sample consisted of 2,353 teachers (62% female) with a self-reported racial/ethnic distribution of 83% White/Caucasian, 12% African-American, 2% Latino, 1% Asian-American, 1% American Indian, and 1% Other.

Measures

The eight measures (six scales and two single items) used in this study were embedded in a 137-item school climate survey (Cornell & Gregory, 2008). Each measure was administered to both students and teachers. To present alpha coefficients for scores

from each of the scales, the scores for each student in a school were averaged into a single schoolwide student score and similarly the scores for each teacher in a school were averaged into a single schoolwide teacher score. It should be noted that subsequent data analyses are based on individual students and teachers randomly selected from each school.

Academic Press. The Academic Press scale (Midgley et al., 2000) contained 6 items measuring how much teachers press the student to study hard and do challenging work. Respondents were asked whether items such as, "When I've figured out how to do a problem, my teachers give me more challenging problems to think about," and "My teachers accept nothing less than my full effort," with four response options (*Not very True, Not at all True, Somewhat True, True,* or *Very True*). Internal consistency (Cronbach's alpha) was .88 for students and .92 for teachers.

Daily Structure. This index (Cornell, 2006) consisted of 6 items devised to measure student perceptions of how strictly rules were enforced for common problems such as cutting class, coming late to class, smoking, fighting, and speaking sarcastically to a teacher. Respondents were asked about likelihood (*Not at all Likely, Not Likely, Likely* or *Very Likely*) of statements such as, "If a student cut class, how likely would the student be caught?" and "If a student was five minutes late for class, how likely would teachers overlook it?" Internal consistency was .53 for students and .81 for teachers.

Experience of School Rules. This is a 7-item scale that measured perceptions of the school rules as fair and strictly enforced. This scale was used in the School Crime Supplement to the National Crime Victimization Survey (NCES, 2005). Respondents

were asked to *Strongly Disagree, Disagree, Agree,* or *Strongly Agree* with statements such as, "Everyone knows the school rules for student conduct," and "If a school rule is broken, students know what kind of punishment will follow." Internal consistency was .74 for students and .91 for teachers.

Perceptions of Bullying and Teasing at School. This scale, the same Prevalence of Bullying and Teasing measure used in Study One, consists of 4 items describing the extent of teasing and bullying that takes place at school. Internal consistency was .78 for students and.84 for teachers.

Security Measures. This is an index of 9 school security measures (e.g., metal detectors, security cameras) taken from the School Crime Supplement to the National Crime Victimization Survey (NCES, 2005). Respondents were asked to respond *Yes, No,* or *Don't Know* to items such as, "Does your school take any of these measures (i.e. locker checks, metal detectors, etc.) to make sure students are safe?" The "*Yes*" responses on the Security measure index were summed and ranged from 0 to 9 on the student and teacher surveys. Students typically reported 5 of the nine listed security measures, while teachers reported 6 or 7. Internal consistency was .48 for students and .41 for teachers.

The relatively low internal consistency values for the Security Measures index (and for Daily Structure) suggest that observers were not highly consistent in their reports of the items we grouped together. As noted by Streiner (2003), internal consistency may be low when items are not manifestations of an underlying hypothetical construct, but are used to tap qualities that define the construct. These qualities, such as different forms of

security measures, may not be highly intercorrelated and so the set of items is better described as an index rather than a scale.

Willingness to Seek Help. This is an 8-item version of the same student scale used in Study One. The item dropped from the *Virginia High School Safety Study* version of the scale was, "The teachers at this school are genuinely concerned about me." Teacher perceptions of student willingness to seek help from school staff members for bullying and threats of violence were measured with a 6-item *Help-seeking* scale from the School Climate Bullying Survey (Cornell, 2011). Internal consistency was .89 for students and .82 for teachers.

Time Out of Class questions. Two questions were used to ask students how much time they have for lunch *(< 20 minutes, 20-29 minutes, 30-39 minutes, 40-49 minutes, 50-59 minutes, or 60+ minutes*) and how many times they change class in a typical day (*0* to *10 times*).

Procedure

Students completed the survey anonymously and online at school. The online survey form required participants to answer each item before moving to the next page. No compensation was provided to either students or teachers.

Data Analyses

In order to identify invalid responders, two screening questions were used: "I am being honest in this survey" and "I am telling the truth in this survey." Negative responses (*strongly disagree* or *disagree*) to either question flagged the respondent as an invalid responder (n = 281 from 178 schools). The prevalence rate of invalid responders

in the overall sample was 4%. The remaining students (n = 6,965) were labeled as valid responders.

Responses from invalid and valid responding students were compared with the external criterion of teacher reports. One invalid responder per school was randomly selected and matched, based on gender and race/ethnicity, with one valid responder from the same school. Invalid and valid student responses to the eight measures were correlated with the corresponding measures obtained from teachers. Because correlations will fluctuate due to sampling error, we employed a bootstrap resampling procedure using SAS 9.2 to obtain more stable correlation coefficient estimates. We conducted 200 random resamplings of invalid responders and matched valid students, and averaged the corresponding correlations. Our hypothesis was that valid and invalid responders would differ in their report of school conditions and that the correlations between valid responders and teachers would be greater than the correlations between invalid responders and teachers. One-tailed tests of differences between paired correlations used Fisher's r to z transformation using the Psych package (Revelle, 2010) in R (R Development Core Team, 2009). In addition, mean differences between responses of valid and invalid responders were compared using *t*-tests; effect sizes are shown using Cohen's d as a standardized measure of mean differences.

Results

Descriptive demographic comparisons of valid and invalid responders (Table 4) show a greater proportion of invalid male responders (61.2%) compared to valid male responders (50.3%), $\chi^2(1) = 12.84$, $p < .001$. Invalid responders also had a higher

proportion of non-white students (29.9% Black and 20.3% Other) compared to valid responders (22.5% Black and 14.4% Other), $\chi^2(2) = 20.57$, $p < .001$. A formal test to predict responder type (i.e., valid or invalid) was conducted using logistic regression analysis. Results show that invalid responders were more likely to be male (*OR* = 1.56) and non-white (*OR* = 1.72) students.

Results of the t-tests (see Table 6) indicated that valid responders consistently gave higher scores on all the scales compared to the invalid responders ($p < .05$). The effect sizes can be characterized as small ($d = 0.26$) to moderate ($d = 0.60$) based on Cohen's (1992) standards. For the time out of class items, valid responders indicated that they changed classes less frequently ($M = 4.76$) compared to the invalid responders (M = 5.04), $t = 2.58$, $p = .01$, $d = 0.29$. Valid respondents also said that they had more time for lunch ($M = 3.48$) compared to invalid responders ($M = 3.18$), $t = 2.96$, $p < .01$, $d = 0.33$).

There were higher correlations between teacher responses and valid student responders on two measures: Security Measures ($r = .35$ vs. $.13$, $p = .02$) and Willingness to Seek Help ($r = .15$ vs. $-.03$, $p = .05$). In addition, the correlation between teacher and valid student responders was higher on the time out of class item, "How many times do students change classes on a normal day" ($r = .70$ vs. $.47$, $p < .01$).

Discussion

These two studies support the need to consider the use of additional screening procedures for the validity of adolescent self-report surveys. Many widely used surveys such as the Youth Risk Behavioral Surveillance Survey (U.S. Department of Health and Human Services, 2009) do not use validity screening items to identify participants who

acknowledge that they are not answering truthfully or carefully. In this article, students who endorsed at least one screening item inappropriately were designated “invalid responders” and contrasted with “valid responders.” Although relatively few students may endorse such items inappropriately, their answers may exert a distorting effect on survey results.

Prevalence rate findings. In the first study, invalid responders endorsed significantly higher rates of risk behavior than valid responders, including items used in the Youth Risk Behavior Survey. For example, the rates of endorsement for risk behaviors such as smoking, drinking alcohol, and using marijuana were more than double among invalid responders compared to valid responders. The self-reported rate of carrying weapons was more than four times greater among invalid responders than valid responders. Across ten risk items and three validity items, nearly all of the comparisons were statistically significant with small to moderate effect sizes. These findings indicate that students who admit they are not answering a survey truthfully or carefully are generally (although not uniformly) inclined to endorse high-risk behaviors.

The 918 students identified as invalid responders was a relatively small proportion (approximately 11.8%) of the full sample of 7,801 students; nevertheless, their impact was substantial. Because prevalence rates for high-risk behaviors are generally low, even a small number of invalid responders can produce inflated rates. For example, the prevalence rates for marijuana use increased from 10.8% in the screened sample to 12.8% in the full, unscreened sample. This increase of 2% represents an 18.5% inflation in reported marijuana use. In reporting national trends, such a difference would attract

nationwide concern. In contrast, the smallest effects were observed for two items that may be less appealing for students to endorse, feeling sad (inflated just 2%) and considering suicide (inflated 10%).

The error in survey results that could be attributable to inconsistent and extreme responding is large enough to rival the typical reductions reported by many violence and risk prevention programs. For example, in their meta-analysis of anti-bullying programs, Ttofi and Farrington (2009) concluded that programs are effective in reducing bullying and victimization by about 20-23%. In the present study, validity screening alone reduced the self-reported prevalence of being bullied from 7.4% to 6.8%, a reduction of 8% and reduced the prevalence of bullying others from 4.8% to 3.7%, a reduction of 23%. In other words, the self-report method is vulnerable to measurement errors that can be about as large as expected treatment effects. Conceivably, some treatment effects, or failure to obtain treatment effects, could be attributed to changes in the attitudes of students toward completing the survey from baseline to follow-up.

There may be variations across surveys and survey sites in how frequently students give inappropriate responses to validity screening items. Using a different set of validity criteria, Cross and Newman-Gonchar (2004) found that the proportion of students in one high school who reported having been bullied was 45.7%, but after invalid surveys were removed from the sample, the proportion dropped to 25.0%, a reduction of more than 45%. Student attitudes toward completing the survey may be influenced by their attitudes toward their school and the authorities that administer the survey, as well as their understanding and appreciation of the purpose of the survey. Baly

and Cornell (in press) found that middle school students reported lower rates of bullying after observing an educational video that was designed to teach them the distinction between bullying and peer conflict between students of comparable strength or status.

It is noteworthy that relatively few participants—just 3.8%—in Study Two were identified as invalid responders compared to the 11.8% in Study One. In part, this may be attributable to the use of two rather than three validity screening items in Study Two (the use of the two truth two items in Study One would have generated an invalid response rate of 6.9%), but there were also differences in student selection procedures. The students in Study One included nearly all available students in the school (approximately 91% of school enrollment) because the survey was conducted on a school-wide basis. In contrast, Study Two participants were a more select group that consisted of 73% of a small group in each school identified by random number and invited by the school principal to participate in the survey. It is possible that the students in Study Two were more serious about their participation. There is a need to study factors that influence student willingness to complete surveys in a reliable manner.

Factor structure findings. Whereas the first set of analyses showed that invalid responders produced inflated prevalence rates for individual risk behavior items, the second set of analyses in Study One demonstrated that factor scales for school climate measures generated by invalid responders differed from those produced using valid responders. A comparison of factor loadings found that ten of 20 items failed to measure their respective factor with similar accuracy. Each of these items produced *higher* factor loadings in the invalid group than the valid group.

The higher factor loadings for the invalid group might seem like a counter-intuitive finding, but can be understood as a reflection of the greater tendency among invalid responders to provide extreme responses that are more homogeneous in terms of their position on the item scale and thereby enhance the magnitude of correlations among related items. For example, invalid responders were more likely to frequently endorse highly aggressive attitudes (e.g., “If you fight a lot, everyone will look up to you” and “Students who are bullied or teased mostly deserve it”) and these items loaded higher on the Aggressive Attitudes scale than when answered by valid responders. This phenomenon is similar to the finding by Fan and colleagues (Fan et al., 2002) that there was an inflated correlation between adoption status and high-risk behaviors because students who tended to falsely claim to be adopted also endorsed higher rates of risky behaviors. An important implication of these findings is that researchers may obtain inflated factor loadings and the appearance of a stronger factor structure because of the failure to exclude invalid responders who are giving extreme responses.

Study Two findings. The second study demonstrated that invalid responders can be less reliable informants than valid responders in reporting on school conditions. There were no statistically significant correlations between valid and invalid responders for any of the six school climate scales, but moderate correspondence (.39-.41) for the two questions concerning the number of class changes each day and minutes for lunch.

The invalid responders tended to report a consistently less favorable view of the school than valid responders. The invalid responders described teachers as less concerned with pressing students to work hard (Academic Press) and that the rules were less fair

(Experience of School Rules) in comparison to valid responders. The invalid responders reported a higher level of bullying and teasing at school (Prevalence of Bullying and Teasing), but that teachers were less likely to help them with problems in this area (Willingness to Seek Help). The invalid responders also described a less structured and secure school environment, with fewer security measures (Security Measures) and less consistency in enforcing school rules (Daily Structure). They reported more frequent class changes and less time for lunch. These findings suggest that invalid responders not only provide inflated reports of their own engagement in high risk behavior, but that they are negatively biased informants about school conditions.

The final analyses in Study Two concerned whether valid and invalid responders reported perceptions of school conditions that correlated with an independent criterion—in this case, teacher perceptions. For valid responders there was a consistent pattern of low, but statistically significant correlations with teacher perceptions for four of the eight school condition measures. The highest levels of agreement between valid responders and teachers were for the relatively more observable and objective measures, such as the number of times students change classes on a normal day (r = .70), the number of minutes that students have for lunch on a normal day (r = .65), and the scale listing the security measures in place at the school (r = .35). There was relatively less correspondence between invalid responding students and teachers in their perceptions of school conditions, with only three of the eight correlations statistically significant.

There were statistically significant differences between pairs of correlations for three measures, with valid responders more highly correlated with teacher perceptions

than were invalid responders in reporting the number of times students change classes on a normal day, the number of security measures in place at school, and student willingness to seek help for bullying and threats of violence. These findings suggest that invalid responders provide less accurate data in reporting on school conditions. It would be useful to test this conclusion using a stronger criterion than teacher perceptions, since teacher and student perceptions are often weakly correlated (Konold & Glutting, 2008).

Limitations and Directions for Further Study

A key limitation to this study is that these validity questions only identify students who were willing to acknowledge their dishonesty or lack of care in answering questions. There may be students who answer the validity questions appropriately but then answer other questions inaccurately. The validity questions in the present study offer only an incremental improvement in the quality of survey data, and there are other undetermined sources of error to consider. Research is needed to understand the value of these questions or to determine whether there are more useful questions or other ways to assess test-taking attitudes that improve the quality of survey data.

Adolescent personality inventories such as the Minnesota Multiphasic Personality of Adolescents (MMPI-A, Butcher et al., 1992) and the Millon Adolescent Clinical Inventory (MACI, Millon et al., 1993) have validity screening scales to measure test-taking attitudes including both defensiveness (under-reporting of problems) and exaggeration (over-reporting). Notably, the MACI includes two validity items to detect random or intentionally provocative reporting: (1) “I have not seen a car in the last ten years,” and (2) “I flew across the Atlantic 30 times last year.” The MACI manual

recommends that one item is endorsed, “the test results may be unreliable” and that if both items are endorsed, “the test should be considered invalid” (Millon et al., 2006, p. 56). However, there is no report of the frequency of unreliable or invalid responding in the manual and no published research on these items could be found.

Survey participants might endorse validity items inappropriately for a number of different reasons: in acknowledgement of not being accurate on the survey, as a consequence of careless responding, or because of some other response set such as yea-saying or nay-saying. Additional research that manipulates different forms of validity questions might shed light on the relative contribution of these different reasons.

Finally, it must be emphasized that we do not contend that all students who inappropriately endorsed these validity items were not telling the truth on all other items. For example, some of their reports of involvement in high-risk behaviors may have been accurate and indeed, the invalid responders did not uniformly and consistently endorse extreme levels of these behaviors, and some items (such as using marijuana) showed much higher levels of endorsement than others (such as attempting suicide). Some invalid responders might be accurately reporting a high rate of risk behavior. As a result, removal of participants classified as invalid responders based on their response to validity items could mean the removal of valid as well as invalid data. The key issue is whether the removal of invalid respondents has a net positive impact on data quality. The results of our two studies suggest that there is a net improvement in data by removal of invalid respondents, although of course further studies are needed.

External validity of self-report surveys. Researchers may be less inclined to recognize the serious consequences of adolescent misrepresentation on self-report surveys because of the great convenience and efficiency of collecting anonymous survey data. It would be practically impossible to gather direct observations on adolescent risk behavior and there is no other way to gather data so easily from large samples. Unfortunately, rater effects (e.g., rater mood and inclination to provide socially desirable responses) can have adverse effects on the traits being measured (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), and these effects are often greater than the influence of the behavioral trait being measured (Konold & Glutting, 2008). Moreover, behavioral ratings provided by parent-teacher (Konold & Pianta, 2007) and parent-student (Konold & Glutting, 2008) dyads are weakly correlated.

It is important for future studies to have external criteria with strong evidence of reliability and validity. The low internal consistency coefficients obtained for several scales and indices used in this study could have affected our results and limited our ability to show differential effects for valid and invalid responders.

The anonymity of survey research makes it possible to ask questions that adolescents presumably are free to answer honestly, but this anonymity prevents researchers from verifying the accuracy of the students' claims. Does this adolescent really consume alcohol and use marijuana on a weekly basis? In the face of the practical and methodological obstacles to obtaining verifiable data from any other source, researchers have accepted the convenient assumption that survey data are good enough and that errors are sufficiently random and non-systematic to be negligible. On the

contrary, the previously discussed research by Fan and colleagues (2002, 2006) provided a rare opportunity to check the accuracy of student survey data that revealed a systematic bias among a small group of students that was sufficient to generate false findings. As a result, seemingly important findings about the adjustment of adopted adolescents (Miller et al., 2000) had to be retracted. How many other illusory findings might there be in the survey research literature that went undetected because there was no opportunity to verify the accuracy of responses?

These concerns about the accuracy of adolescent survey data suggest that there is a need to provide stronger evidence for the external validity of surveys that are widely used on an anonymous basis. (There may be a similar need to examine the accuracy of adult survey data, too, but adolescents seem more prone to defiance.) The foundation for adolescent surveys should rest on a body of evidence that they provide valid information that meets minimal standards for external confirmation with independent criteria. This is not an extraordinary goal, since there is an expectation for external criterion-related validity in most psychological measures (AERA, APA, NCME, 1999). Surveys should not be given a pass on this requirement because of the practical and logistical problems in conducting such research. There is a need for studies that provide credible evidence for the accuracy of student reports before such instruments are adopted for widespread use and employed as indicators of national trends and measures of intervention effectiveness. Such studies would be labor-intensive and methodologically complex because they could involve follow-up interviews with the adolescent or with suitable informants who can verify survey claims.

A common objection to the external validation of surveys is that anonymity is essential to obtain information on sensitive topics, and anonymity prevents researchers from linking survey data to independent sources of information. However, the enormously fruitful lines of research using the Add Health survey (Carolina Population Center, n.d.) demonstrate that confidential surveys are viable. Moreover, there is some research that challenges the assumption that an anonymous survey is necessary in order for students to reveal sensitive information. Chan and colleagues (2005) studied students who were randomly assigned to take a bullying survey anonymously or to write their names on the survey. There were no statistically significant differences in rates of endorsement of behaviors that reflected bullying others and being victims of bullying (i.e., hitting, teasing, and lying about other students) between these two groups. O'Malley and colleagues (2000) examined differences between anonymous and non-anonymous adolescent reporting of drug use and illegal behaviors (i.e., stealing and weapon carrying) on the Monitoring the Future survey. In this study, one group answered the survey anonymously and the other group was required to report names and addresses to researchers, but was told that their answers would be confidential. Again there were little or no group differences in endorsement rates for sensitive information (O'Malley et al., 2000).

A related line of research should examine the conditions that affect participant attitudes and compliance in taking a survey. How does the presentation of the survey affect student response to it? For example, are students more receptive to surveys when they receive more complete explanations of the purpose of the survey and information

about its value? Do student attitudes toward the persons administering the survey affect the results? How do the attitudes of persons administering the survey (e.g., teachers who regard the survey as unimportant) affect student response patterns? With the increasing use of online survey administration, there is considerable work needed to understand the dynamics of student response to survey format and conditions of administration.

In conclusion, adolescent self-report surveys are a staple of psychological assessment in social science research because of their convenience and efficiency in collecting data from large samples on sensitive topics, but evidence for the accuracy of adolescent self-reports using external criteria is sparse. Our results indicate that a small, but noteworthy proportion of adolescents will admit that they are not answering questions truthfully or carefully, and the data produced by these surveys are systematically and substantially different from that of other adolescents. These findings point to the need for more systematic research on the validity of adolescent self-report and consideration of validity screening items as a means of improving data quality.

References

AERA, APA, NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American School Health Association. (1989). *National Adolescent Student Health Survey (NASHS). A report on the health of America's youth*. Oakland, CA: Third Party Publishing Company.

Arbuckle, J.L. (2007). *Amos16.0 User's Guide*, Chicago, IL: SPSS Inc.

Baly, M. W. & Cornell, D. G. (in press). Effects of an educational video on the measurement of bullying by self-report. *Journal of School Violence.*

Bandyopadhyay, S., Cornell, D. G., & Konold, T. R. (2009). Internal and external validity of three school climate scales from the School Climate Bullying Survey. *School Psychology Review, 38,* 338-355.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606. doi:10.1037/0033-2909.88.3.588

Branson, C., & Cornell, D. (2009). A comparison of self and peer reports in the assessment of middle school bullying. *Journal of Applied School Psychology, 25,* 5-27. doi:10.1080/15377900802484133

Brener, N.D., Kann, L., Kinchen, S.A., Grunbaum, J.A., Whalen, L., Eaton, D., Hawkins, J. & Ross, J.G. (2004). Methodology of the Youth Risk Behavior Surveillance System. *Morbidity and Mortality Weekly Report, 53,* 1-12.

Brener, N. D., McManus T., Galuska, D. A., Lowry, R., & Wechsler, H. (2003).

Reliability and validity of self-reported height and weight among high school students. *Journal of Adolescent Health, 32*(4), 281–287. doi:10.1016/S1054-139X(02)00708-5

doi: 10.1016/S1054-139X(02)00708-5

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp.136-162). Newbury Park, CA: Sage.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., et al. (1992). *MMPI-A, Minnesota Multiphasic Personality Inventory—Adolescent: Manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.

California Healthy *Kids Survey, 2007-09 Statewide Results*: *Main Report.* San Francisco, CA: WestEd Health and Human Development Program for the California Department of Education.

Carolina Population Center (n.d.). *Add Health*. Retrieved from http://www.cpc.unc.edu/projects/addhealth

Celis, W. (1994, August 31). Schools getting tough on guns in the classroom. *The New York Times*. Retrieved from http://www.nytimes.com/1994/08/31/us/schools-getting-tough-on-guns-in-the-classroom.html?n=Top%2FReference%2FTimes+Topics%2FPeople%2FC%2FCelis%2C+William+III&pagewanted=all

Centers for Disease Control and Prevention. [Youth Risk Behavior Surveillance —

United States, 2009]. Surveillance Summaries, [June 4, 2010]. MMWR 2010; 59 (No. SS-5).

Chan, H.F.J., Myron, R., Crawshaw, M. (2005). The efficacy of non-anonymous measures of bullying. *School Psychology International, 26,* 443.458.

Chromy, J. R. (1998) *The effects of finite sampling corrections on state assessment sample requirements. NAEP validity studies (NVS)* (ED460134). Washington, DC: American Institutes for Research. Retrieved from http://eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED460134&ERICExtSearch_SearchType_0=no&accno=ED460134

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi:10.1037/0033-2909.112.1.155

Cornell, D. G. (2011). *The School Climate Bullying Survey: Description and Research Summary*. Unpublished report, University of Virginia, Charlottesville, Virginia.

Cornell, D. G. (2006). *School violence: Fears versus facts.* Hillsdale, New Jersey: Erlbaum.

Cornell, D., & Brockenbrough, K. (2004). Identification of bullies and victims: A comparison of methods. *Journal of School Violence*, *3*(2), 63-87. doi:10.1300/J202v03n02_05

Cornell, D., & Gregory, A. (2008). *Virginia High School Safety Study: Descriptive Report of Survey Results from Ninth Grade Students and Teachers*. Charlottesville, VA: University of Virginia.

Cornell, D. G., & Loper, A. B. (1998). Assessment of violence and other high-risk behaviors with a school survey. *School Psychology Review, 27*, 317-330.

Donegan, B. (2008). The linchpin year. *Educational Leadership, 65,* 54-57.

Fan, X., Miller, B., Park, K., Winward, B., Christensen, M., Grotevant, H., & Tai, R. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, *18*(3), 223-244. doi:10.1177/152822X06289161

Fan, X., Miller, B., Christensen, M., Bayley, B., Park, K., Grotevant, H., & van Dulmen, M., & Dunbar, N. (2002). Questionnaire and interview inconsistencies exaggerated differences between adopted and non-adopted adolescents in a national sample. *Adoption Quarterly*, *6*(2), 7-27. doi:10.1300/J145v06n02_02

Furlong, M., Sharkey, J., Bates, M. P., & Smith, D. (2004). An examination of reliability, data screening procedures, and extreme response patterns for the Youth Risk Behavior Surveillance Survey. *Journal of School Violence*, *3*(2), 109-130. doi:10.1300/J202v03n02_07

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55. doi:10.1080/10705519909540118

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

Ingels, S. J. (1990) *Findings from the NELS:88 Base Year Student Survey*. (ERIC Document Reproduction Service No. ED319747)

Johnston, L. D., O'Malley, P. M., Bachman, J. G., & Schulenberg, J. E. (2010). *Monitoring the future national survey results on drug use, 1975-2009. Volume I: Secondary school students* (NIH Publication No. 10-7584). Bethesda, MD: National Institute on Drug Abuse.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25*(2), 137-155. doi:10.1207/s15327906mbr2502_1

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.

Konold, T.R., & Glutting, G.G. (2008). ADHD and method variance: A latent variable approach applied to a nationally representative sample of college freshmen. *Journal of Learning Disabilities*, *41*, 405-416.

Konold, T.R., & Pianta, R.C. (2007). The influence of informants on ratings of children's behavioral functioning. *Journal of Psychoeducational Assesment*, *25*, 222-236.

Levine, F. J., & Rosich, K. J. (1996). *Social causes of violence: Crafting a science agenda.* Washington, DC: American Sociological Association.

Maginnis, R. (1995). *Violence in the schoolhouse: At ten year update*. Washington, DC: Family Research Council.

Marsh, H.W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers of over generalizing Hu and Bentler's (1999) finding. *Structural Equation Modeling*, 11, 320-341.

Midgley, C., Maehr, M., Hruda L., Anderman, E., Freeman, K. (2000). *Manual for patterns of adaptive learning.* Ann Arbor: MI.

Miller, B.C., Fan, X., Christensen, M., Grotevant, H.D., & van Dulmen, M. (2000). Comparisons of adopted and nonadopted adolescents in a large, nationally representative sample. *Child Development, 71*(5), 1458-1473. doi:10.1111/1467-8624.00239

Millon, T., Millon, C., Davis, R., & Grossman, S. (1993). *MACI Manual, Second Edition.* Minneapolis, MN: NCS Pearson, Inc.

National Center for Education Statistics (2005). *School crime supplement to the national crime victimization survey 2005*. Retrieved from http://nces.ed.gov/Programs/Crime/surveys.asp.

National Crime Prevention Council. (1995). Strategy: Gun-free school zones. Retrieved from http://www.ncpc.org/ncpc/ncpc/?pg=2088-10796

National Education Association (2005). *Statistics: Gun violence in our communities*. Retrieved from http://www.neahin.org/programs/schoolsafety/gunsafety/statistics.htm

National School Boards Association (1993). *Violence in the schools: How America's school boards are safeguarding your children*. Alexandria, VA: Author.

O'Malley, P.M., Johnston, L.D., Bachman, J.G., & Schulenberg, J.E. (2000). A comparison of confidential versus anonymous survey procedures: Effects on reporting of drug use and related attitudes and beliefs in a national study of students. *Journal of Drug Issues*, *30*, 35-54.

Podsakoff, P.M., KacKenzie, S.B., Lee, J.Y., & Podsakoff, N.P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879-903.

R Development Core Team (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Revelle, W. (2010). *Psych*: *Procedures for psychological, psychometric, and personality research*. Retrieved from http://www.personality-project.org/R/psych.manual.pdf

Robers, S., Zhang, J., & Truman, J. (2010). *Indicators of school crime and safety: 2010*. (NCES 2011-002/NCJ 230812). Washington, DC: National Center for Education Statistics, U.S. Department of Education, and Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice.

Rosenblatt, J. A., & Furlong, M. J. (1997). Assessing the reliability and validity of student self-reports of campus violence. *Journal of Youth and Adolescence*, *26*(2), 187-202. doi:10.1023/A:1024552531672

Sharkey, J. D., Furlong, M. J., & Yetter G. (2006). An overview of measurement issues in school violence and school safety research. In S. R. Jimerson & M. J. Furlong (Eds.), *The handbook of school violence and school safety: From research to practice* (pp. 121-134). Mahwah, NJ: Lawrence Erlbaum Associates.

Singer, C. (2005). School-by-school reform. Retrieved from http://www.pbs.org/makingschoolswork/sbs/csp/tension.html

Solberg, M., & Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim Questionnaire. *Aggressive Behavior, 29,* 239-268. Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient Alpha does and doesn't matter. *Journal of Personality Assessment, 80,* 217-222. doi:10.1207/S15327752JPA8003_01

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation modeling. In K. S. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp.10-39). Newbury Park, CA: Sage.

Ttofi, M. M., & Farrington, D. P. (2009). What works in preventing bullying: Effective elements of anti-bullying programmes. *Journal of Aggression, Conflict and Peace Research, 1*, 13-24.

U.S. Department of Health and Human Services. (2009). United States High School Survey Data Users Manual. Rockville, MD: Author.

U.S. Office of Juvenile Justice and Delinquency Prevention (n.d.). Juvenile justice reform initiatives in the states: 1994-1996. Retrieved from http://www.ojjdp.gov/pubs/reform/ch1_c.html

Williams, F., & Cornell, D. (2006). Student willingness to seek help for threats of violence. *Journal of School Violence*, *5*(4), 35-49. doi:10.1300/J202v05n04_04

Witkin, G. (1991, April 8). Kids who kill; disputes once settled with fists are now settled with guns. *U.S. News & World Report*. *110*, 26 (7).

Table 1

Descriptive statistics for sample before and after validity screening

Risk items	Total Sample (N = 7,801) N (%)	After Validity Screening[a] (N = 6,883) N (%)	Total Sample Inflation (%)
I have been bullied at school in the past month.			
About Once a Week or Several Times per Week	574 (7.4)	466 (6.8)	8.82
Never or Once or Twice	7,185 (92.1)	6,388 (92.8)	0.76
I have bullied others at school in the past month.			
About Once a Week or Several Times per Week	371 (4.8)	258 (3.7)	29.73
Never or Once or Twice	7,317 (93.8)	6,538 (95.0)	1.28
During the past 30 days, on how many days did you smoke cigarettes?			
Agree or Strongly Agree	746 (9.6)	538 (7.8)	23.08
Disagree or Strongly Disagree	7,020 (90.0)	6,321 (91.8)	2.00
During the past 30 days, on how many days did you have at least one drink of alcohol?			
Agree or Strongly Agree	1,413 (18.1)	1,139 (16.5)	9.70
Disagree or Strongly Disagree	6,328 (81.1)	5,702 (82.8)	2.10
During the past 30 days, how many times did you use marijuana?			
Agree or Strongly Agree	999 (12.8)	743 (10.8)	18.52
Disagree or Strongly Disagree	6,774 (86.8)	6,117 (88.9)	2.42
During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club on school property?			
Agree or Strongly Agree	473 (6.1)	301 (4.4)	38.64
Disagree or Strongly Disagree	7,256 (93.0)	6,530 (94.9)	2.04
During the past 12 months, how many times were you in a physical fight on school property?			
Agree or Strongly Agree	945 (12.1)	727 (10.6)	14.15
Disagree or Strongly Disagree	6,801 (87.2)	6,114 (88.8)	1.83

During the past 30 days, on how many days did you not go to school because you felt you would be unsafe at school or on your way to or from school?			
Agree or Strongly Agree	524 (6.7)	359 (5.2)	28.85
Disagree or Strongly Disagree	7,186 (92.1)	6,455 (93.8)	1.85
During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?			
Agree or Strongly Agree	1,595 (20.4)	1,376 (20.0)	2.00
Disagree or Strongly Disagree	6,161 (79.0)	5,474 (79.5)	0.63
During the past 12 months, did you ever seriously consider attempting suicide?			
Agree or Strongly Agree	917 (11.8)	738 (10.7)	10.28
Disagree or Strongly Disagree	6,848 (87.8)	6,123 (89.0)	1.37

Note. Missing data account for differences in N.

[a]Respondents endorsing not telling the truth, giving false answers, or not paying attention were removed from the sample as part of validity screening.

Table 2

MGCFA factor loadings for school climate items of the SCBS

Factor Items	Valid Group (N = 6,814)		Invalid Group (N = 909)
Factor 1 (Willingness to Seek Help)			
Factor *r* (F1 and F2)	**.57**		**.29**
1. If another student was bullying me, I would tell one of the teachers or staff at school.	.63	=	.70
2. Students here try to stop bullying when they see it happening.	.51	=	.57
3. If another student brought a gun to school, I would tell one of the teachers or staff at school.	.50		.65
4. Teachers here make it clear to students that bullying is not tolerated.	.62		.72
5. If another student talked about killing someone, I would tell one of the teachers or staff at school.	.53		.70
6. If I tell a teacher that someone is bullying me, the teacher will do something to help.	.80	=	.82
7. There are adults at this school I could turn to if I had a personal problem.	.69		.78
8. Students tell teachers when other students are being bullied.	.61	=	.67
9. The teachers at this school are genuinely concerned about me.	.70		.80
Factor 2 (Aggressive Attitudes)			
Factor *r* (F2 and F3)	**.45**		**.36**
10. It feels good when I hit someone.	.74	=	.71
11. If you fight a lot, everyone will look up to you.	.68		.75
12. Sometimes you only have two choices – get punched or punch the other person first.	.69	=	.74
13. If you are afraid to fight, you won't have many friends.	.66		.73
14. If someone threatens you, it is okay to hit that person.	.77	=	.76
15. Students who are bullied or teased mostly deserve it.	.62		.73

Factor Items	Valid Group (N = 6,814)		Invalid Group (N = 909)
16. Bullying is sometimes fun to do.	.69		.73
Factor 3 (Prevalence of Bullying and Teasing)			
Factor *r* (F3 and F1)	**.33**		**.18**
17. Bullying is a problem at this school.	.63	=	.67
18. Students here often get teased about their clothing or physical appearance.	.77		.80
19. Students here often get put down because of their race or ethnicity.	.65	=	.72
20. There is a lot of teasing about sexual topics at this school.	.58	=	.67

Note. '=' denotes equality of unstandardized coefficients (not shown) as indicated by MGCFA analyses.

Table 3

Comparison of valid and invalid responders on risk items

	I am telling the truth on this survey.		The answers I have given on this survey are true.		I am not paying attention to how I answer this survey.		One or More Items Invalid	
Risk items	Agree or SA[a] N (column %)	Disagree or SD N (column %)	Yes N (column %)	No N (column %)	Agree or SA N (column %)	Disagree or SD N (column %)	Valid N (column %)	Invalid N (column %)
Victim	523 (7.2)	49 (11.0)	530 (7.1)	40 (21.5)	498 (6.9)	68 (13.0)	466 (6.8)	108 (11.9)
Not Victim	6759 (92.8)	398 (89.0)	6964 (92.9)	146 (78.5)	6692 (93.1)	454 (87.0)	6388 (93.2)	797 (88.1)
$\chi 2$ (ϕ)	8.8 (.03)*		55.0 (.09)**		26.6 (.06)**		30.8 (.06)**	
Bully	311 (4.3)	59 (13.6)	327 (4.4)	38 (20.7)	293 (4.1)	75 (14.4)	258 (3.8)	113 (12.7)
Not Bullying	6913 (95.7)	375 (86.4)	7098 (95.6)	146 (79.3)	6829 (95.9)	445 (85.6)	6538 (96.2)	779 (80.3)
$\chi 2$ (ϕ)	76.8 (.10)**		103.8 (.12)**		112.4 (.12)**		135.1 (.13)**	
Cigarettes	629 (8.6)	113 (25.5)	669 (8.9)	68 (36.2)	612 (8.8)	129 (24.5)	538 (7.8)	208 (22.9)
No Cigarettes	6662 (91.4)	331 (74.5)	6828 (91.1)	120 (63.8)	6580 (91.5)	398 (75.5)	6321 (92.2)	699 (77.1)
$\chi 2$ (ϕ)	136.6 (.13) **		157.0 (.14)**		144.3 (.14)**		210.1 (.16)**	
Alcohol	1273 (17.5)	134 (30.6)	1304 (17.5)	90 (47.9)	1237 (17.2)	167 (31.9)	1139 (16.6)	274 (30.4)
No Alcohol	6001 (82.5)	304 (69.4)	6168 (82.5)	98 (52.1)	5935 (82.8)	357 (68.1)	5702 (83.4)	626 (69.6)
$\chi 2$ (ϕ)	47.5 (.08)**		114.0 (.12)**		70.0 (.10)**		101.4 (.11)**	
Marijuana	861 (11.8)	133 (29.7)	900 (12.0)	83 (43.9)	829 (11.5)	160 (30.2)	743 (10.8)	256 (28.0)
No Marijuana	6435 (88.2)	315 (70.3)	6604 (88.0)	106 (56.1)	6368 (88.5)	370 (69.8)	6117 (89.2)	657 (72.0)
$\chi 2$ (ϕ)	120.7 (.13)**		168.6 (.15)**		154.2 (.14)**		213.0 (.17)**	

Carry Weapons	378 (5.2)	92 (20.8)	411 (5.5)	54 (29.7)	361 (5.0)	106 (20.3)	301 (4.4)	172 (19.2)
No Weapons	6880 (94.8)	350 (79.2)	7060 (94.5)	128 (70.3)	6800 (95.0)	416 (79.7)	6530 (95.6)	726 (80.8)
$\chi 2$ (ϕ)	177.1 (.15)**		181.9 (.15)**		198.6 (.16)**		300.4 (.20)**	
Physical Fights	827 (11.4)	108 (24.4)	869 (11.6)	63 (33.7)	791 (11.0)	140 (26.6)	727 (10.6)	218 (24.1)
No Phys Fights	6446 (88.6)	335 (75.6)	6612 (88.4)	124 (66.3)	6384 (89.0)	386 (73.4)	6114 (89.4)	687 (75.9)
$\chi 2$ (ϕ)	66.4 (.09)**		83.3 (.10)**		112.1 (.12)**		135.2 (.13)**	
Feel Unsafe	434 (6.0)	86 (19.6)	470 (6.3)	49 (26.9)	413 (5.8)	102 (19.6)	359 (5.3)	165 (18.4)
Feel Safe	6809 (94.0)	352 (80.4)	6980 (93.7)	133 (73.1)	6733 (94.2)	419 (80.4)	6455 (94.7)	731 (81.6)
$\chi 2$ (ϕ)	121.8 (.13)**		119.1 (.13)**		147.6 (.14)**		216.1 (.17)**	
Depressed	1491 (20.5)	98 (22.2)	1522 (20.3)	57 (30.6)	1436 (20.0)	143 (27.1)	1376 (20.1)	219 (24.2)
Not Depressed	5795 (79.5)	344 (77.8)	5969 (79.7)	129 (69.4)	5751 (80.0)	384 (72.9)	5474 (79.9)	687 (75.8)
$\chi 2$ (ϕ)	0.7 (.01)		11.9 (.04)*		15.4 (.05)**		8.2 (.03)*	
Suicide	824 (11.3)	88 (20.0)	849 (11.3)	59 (31.6)	795 (11.0)	108 (20.5)	738 (10.8)	179 (19.8)
No Suicide	6473 (88.7)	353 (80.0)	6650 (88.7)	128 (68.4)	6401 (89.0)	418 (79.5)	6123 (89.2)	725 (80.2)
$\chi 2$ (ϕ)	30.0 (.06)**		71.7 (.10)**		42.7 (.07)**		62.7 (.09)**	

Note. Missing data account for differences in N.

[a]SA = Strongly Agree; SD = Strongly Disagree.

* $p < .05$. ** $p < .001$.

Table 4

Comparison of responder types by gender and race/ethnicity

	Valid responders *n* = 6,965		Invalid responders *n* = 281		
Characteristic	*n*	*%*	*n*	*%*	χ
Gender					12.84***
Male	3,504	50.3	172	61.2	
Female	3,461	49.7	109	38.8	
Race/Ethnicity					20.57***
White	4,397	63.1	140	49.8	
Black	1,565	22.5	84	29.9	
Other	1,003	14.4	57	20.3	

*** $p < .001$.

Table 5

Comparison of bootstrapped correlation coefficients of valid and invalid student respondents with teacher responses

		Correlations between students and teachers		
		Student group		
Items	*n*	Valid	Invalid	z
Scales				
Academic press	159	.05	.14	0.87
Daily structure	156	.13	.01	1.11
Experience of school rules	156	.10	.10	0.04
Perceptions of teasing and bullying	155	.12	.17	0.33
Security measures	146	.35	.14	2.07*
Willingness to seek help	161	.15	-.03	1.68*
Time out of class items				
How many times do students change classes on a normal day?	156	.70	.47	3.67**
How many minutes do students have for lunch on a normal day?	156	.65	.55	1.52†

† $p < .10$. * $p < .05$. ** $p < .01$.

Table 6

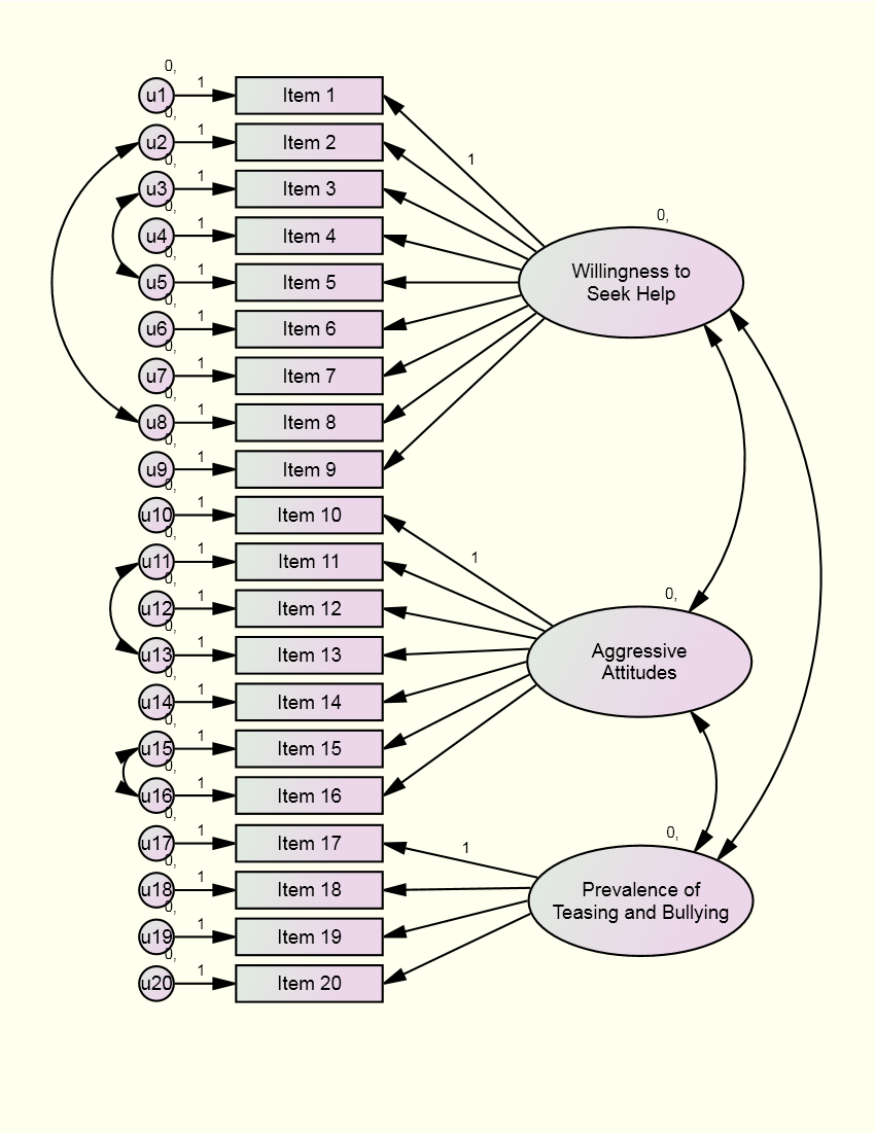
Mean differences between invalid and valid responders

	Invalid		Valid			
Items	*M*	*SD*	*M*	*SD*	*t*	*d*[a]
Scales						
Academic press	18.62	2.81	20.52	3.47	5.37***	0.60
Daily structure	15.89	1.28	16.38	2.22	2.39*	0.27
Experience of school rules	18.56	2.16	19.24	3.01	2.29*	0.26
Perceptions of teasing and bullying	9.61	1.34	10.26	2.21	3.14**	0.36
Security measures	4.43	1.16	4.83	1.30	2.77**	0.32
Willingness to seek help	20.90	2.78	21.96	3.72	2.89**	0.32
Time out of class						
How many times do students change classes on a normal day?	5.03	0.82	4.76	1.02	-2.58*	0.29
How many minutes do students have for lunch on a normal day?	3.18	0.68	3.48	1.07	2.96**	0.33

[a]*d* = Cohen's *d*.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Figure 1: Graphic representation of school climate SCBS measurement structure[1]



1. Complete item stems are provided in Table 3.