Child Youth Care Forum (2015) 44:853–873
DOI 10.1007/s10566-015-9310-5

ORIGINAL PAPER

Measuring Educational Outcomes for At-Risk Children and Youth: Issues with the Validity of Self-Reported Data

Amanda Cleveland Teye · Liliokanaio Peaslee

Published online: 22 February 2015

Abstract
Background Youth programs often rely on self-reported data without clear evidence as to the accuracy of these reports. Although the validity of self-reporting has been confirmed among some high school and college age students, one area that is absent from extant literature is a serious investigation among younger children. Moreover, there is theoretical evidence suggesting limited generalizability in extending findings on older students to younger populations.
Objective The purpose of this study is to examine the validity of academic and attendance self-reporting among children and youth.
Method This study relies on original data collected from 288 children and youth using Big Brothers Big Sisters enrollment and assessment data, paired with school-records from two local school divisions. Initially, we utilized percent agreement, validity coefficients, and average measures ICC scores to assess the response validity of self-reported academic and attendance measures. We then estimated the affects of several moderating factors on reporting agreement (using standardized difference scores). We also accounted for cross-informant associations with child reported GPA using a moderated multiple regression model.
Results Findings indicate that children and youth report their individual grades and attendance poorly. Particularly, younger and lower performing children are more likely to report falsely. However, there is some evidence that a mean construct measure of major subjects GPA is a slightly more valid indicator of academic achievement.
Conclusion Findings suggest that researchers and practitioners should exercise caution in using self-reported grades and attendance indicators from young and low-performing students.

A. C. Teye · L. Peaslee (✉)
Department of Political Science, James Madison University, 91 E. Grace Street, MSC 7705, Harrisonburg, VA 22801, USA
e-mail: peaslelx@jmu.edu

Keywords Academic achievement · Youth programs · Assessment · Outcome evaluation · Self-reported data · Validity

Introduction

Robust outcome assessment is critical for youth programs to demonstrate impacts on protective factors and risky behavior. In particular, researchers and practitioners are often interested in the impact youth programming has on student academic achievement and school attendance (Catalano et al. 2002; DuBois et al. 2002; DuBois et al. 2011; Hall et al. 2003; Thomson and Kelly-Vance 2001). Assessment frequently relies on self-reported data, including attendance, truancy, grades, and standardized test scores, as proxies for actual achievement and academic risk when records are unavailable. Although the validity of self-reporting has been confirmed among some high school and college-age students, one area that is absent from extant literature is a serious investigation among younger children. In order to better understand the risks of using self-reported data from younger children, this paper expands upon existing research by exploring the validity of several academic achievement and attendance measures among 288 elementary and middle school students served by a Big Brothers Big Sisters (BBBS) affiliate agency. Our findings suggest that child and youth self-reported grades and attendance are error prone, especially among academically at-risk populations. Understanding types and sources of error in self-reported data can enable researchers and practitioners to better account for biases that can affect diagnostic practices and outcome reporting.

Interpreting Error in Self-Reported Data

Researchers and practitioners often rely on multiple indicators of the same behavior or construct to validate data and, therefore, more accurately diagnose risk, target interventions, and measure program outcomes. However, there are often systematic differences between data sources. This may be due to problems with construct validity or because some informants are less valid or reliable reporters than others. Weems et al. (2010) posit that children may either under or over-report their emotional responses, such as anxiety sensitivity, relative to a parent for a variety of reasons, including child social desirability bias or low levels of parental awareness of children's emotional state.[1] Yet not all such "informant discrepancies" represent error. For example, a child self-report may diverge from those made by external observers (parents, teachers, clinicians) because he or she may attribute the behavior to a different causal factor (De Los Reyes and Kazdin 2005). Additionally, discrepancies between multiple informant reports might occur because of where that behavior is observed (e.g. children may act differently at home, at school, or in a laboratory setting). In fact, the discrepancies themselves may provide valuable information about the nature or causes of a child's social, emotional, or behavioral problems (De Los Reyes 2011; De Los Reyes et al. 2013; see also Kraemer et al. 2003). For example, we might expect to see variation between a child's self-reported academic ability and parent or teacher's estimations of their ability. These inconsistencies could reflect

[1] Social desirability has been defined as "the desire to revise a response before communicating it to a researcher to protect self-image or inaccurately project an image of academic performance" (Cole et al. 2012, p. 2).

differences in where ability, effort, or attitudes are observed or other contextual factors. On the other hand, we would not expect such variability when asking respondents to report an objective measure such as actual grade received or number of school days missed. Therefore, when discrepancies exist between self-reported grades or attendance and academic records, they are likely to result from some type of measurement error.

While all data is susceptible to measurement error, self-reports are particularly disposed to problems with validity and reliability. Inaccuracies in self-reported data stem from two sources: random and systematic error (Cole et al. 2012; Crockett et al. 1987; Kuncel et al. 2005). Random error occurs by chance; in other words, subjects are likely to alter responses across different assessment periods. Random error commonly results from unpredictable cognitive distortions. That is, asking students to assess school performance and attendance can be subject to problems with information retrieval or misinterpretation. This type of error is most likely to affect reliability rather than validity. In general, random error is not particularly concerning to researchers, largely due to the fact that associated variances should be randomly distributed across the study population. Therefore, when repeated measures are taken, the arithmetic mean error should be null. In a single sample, this may be seen as an equal distribution of errors in both directions—random fluctuations in respondent over-reporting and underreporting that have no true impact on the mean score (Cole et al. 2012). On the other hand, random error may be problematic for practitioners who use self-reported performance measures at baseline to target services. Screening respondents into ability groups based on self-reports is likely to result in some misclassification (Kuncel et al. 2005).

A more significant problem in analyzing self-reported data is the presence of systematic error. This can be understood as a consistent bias or deviation from the 'truth.' The presence of systematic error affects the validity of the data source. In general, it is possible to observe a negative or positive arithmetic mean error, revealing systematic under- or over-reporting within the sample. Although commonly attributed to instrumentation failure, systematic error can be driven by respondents providing false information. In self-reported data, such motivated distortion is likely driven by social desirability bias (Bowman and Hill 2011; Crowne and Marlowe 1964; Dobbins et al. 1993; Gonyea 2005; Kuncel et al. 2005; Martin and Nagao 1989; Mayer et al. 2007; Zimmerman et al. 2002).[2] Systematic error due to social desirability is not limited to self-reported academic data. It is a persistent risk when respondents want to avoid social stigma or enhance self-image across a variety of critical outcome measures. However, unlike some self-reported attitudinal or behavioral data, self-reported academic achievement and attendance can be compared with school records to verify accuracy. Systematic error in academic achievement data has not only been confirmed but is generally understood as being driven by social desirability. A review of the literature reveals that students regularly overreport their academic achievement (Bahrick et al. 1996; Cassady 2001; Dobbins et al. 1993; Escribano and Díaz-Morales 2014; Fetters et al. 1984; Frucot and Cook 1994; Goldman et al. 1990; Mayer et al. 2007; Sawyer et al. 1988; Zimmerman et al. 2002). This would be observed as a positive mean error distortion in the data. However, much less is known about the directionality of systematic error in under- or over-estimating attendance.

Regardless of the directional impact, systematic error is difficult to attribute, especially when the distribution of error is constant across the sample. When systematic error is variable, however, it can be more easily associated with a set of causal factors. Variability

[2] While most researchers attribute overreporting to social desirability, others have attributed inaccuracies to recall failure and biases created by the positive reconstruction of memory (Bahrick et al. 1996, 1993).

can be related to the true value of the measured outcome, indicating different true scores, or to respondent characteristics, indicating bias. We can understand the latter as moderators of response validity. In the following section we identify a range of moderating factors commonly used to predict response validity in the literature. These include respondent academic/cognitive ability level, gender, grade level, race and ethnicity, psychological characteristics, risky behavior, and the recall period for information retrieval. Simply put, we seek to identify whether certain subgroups within a sample of children and youth falsely report their grades or attendance more so than others.

Moderators of Response Validity

The most consistent moderator in predicting response validity in self-reported academic data is student performance level. Here, performance is defined as student achievement, measured by institutionally-reported data. Research has overwhelmingly shown that lower performing students are more likely to provide inaccurate data and tend to overestimate when compared with their higher achieving peers (Anaya 1999; Crocket et al. 1987; Cole et al. 2012; Dobbins et al. 1993; Goldman et al. 1990; Frucot and Cook 1994; Gonyea 2005; Kuncel et al. 2005; Mayer et al. 2007; Sawyer et al. 1988). These findings are consistent across different approaches to operationalizing performance, including using different GPA cutoffs or standardized test scores to distinguish high from low performers (Anaya 1999; Dobbins et al. 1993; Dunnette 1952; Schiel and Noble 1991; Shepperd 1993; Zimmerman et al. 2002).

While construct irrelevant variance due to student performance has been the primary finding in research, findings on other respondent characteristics have been more variable. Where tested, studies frequently find higher validity of self-reports among female respondents versus males (Arthur et al. 2002; Crockett et al. 1987; Fetters et al. 1984; Frucot and Cook 1994; Sawyer et al. 1988; Goldman et al. 1990). Hamilton (1981), on the other hand, found that female students had lower overall correlations between self-reported and actual GPA and SAT scores and also tended to overestimate more. Other studies have found no significant differences based on gender (Escribano and Díaz-Morales 2014; Kuncel et al. 2005; Mayer et al. 2007; Shaw and Mattern 2009; Zimmerman et al. 2002). This variability suggests that further research is needed to examine the moderating role of gender.

Another demographic characteristic that may impact response validity is age, which may also be conceptualized as student grade level. To date, most research examining the reliability and validity of self-reported grades has been conducted with high school and college students. In general, researchers have found relatively strong correlations between student self-reports and academic achievement among older students. In a meta-analysis of 37 independent samples, Kuncel et al. (2005) found strong response validity among college ($r_{obs} = .90$) and high school ($r_{obs} = .82$) respondents (actual values ranged from .45 to .98). Much less has been done, however, in assessing the validity of self-reports among elementary and middle school populations. One exception is a study of 7th and 8th graders by Crockett et al. (1987), which had findings similar to research with older students. In that study, correlations between self-reported and actual grades ranged from .70 to .84. As with older students, researchers noted the presence of systematic bias in a socially desirable direction. Lower achieving students were more likely to overreport scores; however, accuracy increased with age. Nevertheless, there is theoretical evidence suggesting limited generalizability in extending the findings of high school and college studies to younger

populations. Crockett et al. (1987) suggest that the "reliability of recall and the influence of social desirability may change over the course of development" (p. 384). Similarly, Ross (2006) explains, "young children may over-estimate because they lack the cognitive skills to integrate information about their abilities and are more vulnerable to wishful thinking" (p. 3). Although there is limited literature on grade reporting, research indicates that self-assessment in general is less reliable among younger children (Alexander et al. 1994; Blatchford 1997; Butler 1990; Grossman 2009; Kaderavek et al. 2004; Ross 2006). These findings suggest that self-reported grades may be particularly prone to error among younger children.

While not frequently reported, some research suggests that factors like race and ethnicity, psychological characteristics, and engagement in risky behavior may be moderators in predicting response validity. In particular, research has noted that validity of self-reported data is greater among White students than among Black or Hispanic students (Kuncel et al. 2005; Fetters et al. 1984; Shaw and Mattern 2009); however, this finding is variable and rarely tested. Additionally, exploration of the moderating role of psychological characteristics and engagement in risky behavior is limited and inconsistent. (Försterling and Binser 2002) suggests that students with high depressive symptoms are more likely to overinflate grades than those with moderate or low degrees of depression. Similarly, in a review of student–teacher evaluations of academic competence, there was evidence that student overrating was driven by low self-esteem and self-confidence (Connell and Ilardi 1987). On the other hand, Zimmerman et al. (2002) found that youth who reported GPAs accurately were more likely to report higher depressive characteristics while those who overreported had more positive self-perceptions. However, more research is needed on how these characteristics may impact response validity among children.

The time period between recall and grade verification may also impact the validity of self-reports. Intuitively, we would expect that students who are asked to recall grades right after report cards are issued would be more accurate than those who are asked to recall information a few months later. Moreover, there is some evidence that self-reports are more reliable over short time periods (Ross 2006; Shaw and Mattern 2009). For example, Bahrick et al. (1993) asked college freshmen to recall grades from high school and argue that inaccuracies are largely due to flawed reconstructive processing. They posit that recall error likely "occurs some time during the first 6 months, and that little or no reprocessing occurs after that time" (p. 8). These findings are challenged by other research that suggests that accuracy of academic self-reports may be improved by longer time frames for information retrieval (Talento-Miller and Peyton 2006). Still, more research is needed to confirm whether concurrent validity testing should be used over predictive or postdictive forms.

In sum, despite substantial research on the validity of self-reported academic achievement data, most of it has been limited to high school and college students. Very little work has been done to explore response validity among younger children. Additionally, moderating factors other than student ability—including gender, race and ethnicity, personality and behavioral attributes, and range of recall—are not well understood. Moreover, other important academic outcome indicators, like student attendance, are entirely left out. Without a serious examination these factors among children, researchers and practitioners are left with little guidance as to whether they should rely on self-reports as adequate substitutes for school records.

The purpose of this study, therefore, was to extend the research on older adolescents and young adults to children and young adolescents, the age cohorts typically served by BBBS mentoring programs as well as by a large number of other prevention-oriented youth

development programs. As such, first we investigated the validity of self-reported grades and attendance as compared with student academic records among our total sample population. We anticipated that the response accuracy and consistency would be weak to moderate. Next, we assessed the effect of range of recall in reporting, comparing the total study population with a 30-day sub-sample of participants. Here, we anticipated recall time would have a significant effect on reporting accuracy and consistency. Specifically, those responding in a shorter range of recall should have improved reporting accuracy and consistency. Finally, we sought to better understand which factors moderated the validity of academic self-reporting among children. We began by exploring the impact of moderators on individual academic subject grades and attendance. Our null hypotheses suggested there was no relationship between key moderators and subject grades/attendance. We concluded by assessing the effect of moderators on difference scores. Based on the extant literature, we expected that both age and performance would moderate accuracy of self-reports. Specifically, we predicted that younger students would have low response validity when compared with older students and that lower performing students would be likely to overreport or inflate their grades.

Methods

The study was part of a larger experimental research project funded by a grant award from the Office of Juvenile Justice and Delinquency Prevention (OJJDP) that isolated the impact

Table 1 Population Characteristics

Age		
Mean = 9.72 SD = 1.49		
Grade by group		
Elementary (k-4th)	192	67.8 %
Middle (5th–8th)	91	32.2 %
Child Gender		
Male	97	33.8 %
Female	190	66.2 %
Child ethnicity		
White	120	42 %
Black	41	14.8 %
Hispanic	118	41.4 %
Other	5	1.8 %
Time from reporting		
30 days or less	114	40.6 %
More than 30 days	167	59.4 %
Performance		
GPA > 2.0	40	15.6 %
GPA ≥ 2.0	216	84.4 %
Quarterly absenteeism (school reported)	Mean days = 5.08 SD = 4.424	

of training and support on the strength of mentoring relationships and mentee outcomes. Research was conducted in conjunction with a local BBBS affiliate agency in Virginia, along with two local school divisions. BBBS is a nationally recognized mentoring model that matches at-risk children and youth with adult mentors. Research was approved by the Institutional Review Board at James Madison University; all participants were treated in accordance with the ethical guidelines approved by the Board. Informed consent was obtained for all research subjects prior to participation in the study. Participants who did not want to take part in the study were allowed to continue on in the BBBS program. There are no known conflicts of interest associated with this publication. OJJDP had no role in study design, data collection, analysis, or interpretation. The first author assumes primary responsibility for the integrity of the data and the accuracy of data analysis. Both authors have contributed to the interpretation and discussion of results.

Participants

Data were collected from the total population of children enrolled in BBBS from December 2011 through April 2013. Of the 554 children enrolled during this period, 500 parents/guardians consented to participate in the research study. The sample was comprised of students enrolled in both the site-based and community-based programs across two school divisions. Basic demographic data were collected from parents/guardians during the initial application and enrollment periods prior to the match date. All data in this study were collected by agency Enrollment Specialists (ES) and Match Support Specialists (MSS). Table 1 presents descriptive demographic statistics for respondents in the sample population. Respondent mean age was 9.7 years, ranging from 7 to 15 years. Respondent grade level ranged from kindergarten to eighth grade, although the majority (67.8 %) of respondents were elementary aged (grades K through four). Over a third of the sample (33.8 %) was male. Forty-two percent were White, 14.8 % were Black, 41.4 % were Hispanic, and 1.8 % were other ethnicities. At baseline, 15.6 % of students had a school-reported GPA below 2.0. Mean absenteeism was 5.08 days. Not represented in Table 1, 67.3 % of parents reported that their children received free or reduced price lunch. 11.2 % of parents reported their child had some learning disability. Nearly thirty percent of students had an emotional disorder.[3]

Measures

Data collected for the study included self-reported measures gathered from a nationally adopted, agency administered Youth Outcome Survey (YOS), along with local agency enrollment forms, and academic records collected from two local school divisions. The 32-item YOS assesses youth across seven broad behavioral and self-efficacy constructs using Likert scales. These include Self-Reported Grades, Attendance and Truancy, Social Acceptance, School Competency, Future Aspirations, Parental Trust, and Risk Avoidance. Each child completed the original 32-item instrument along with 54 additional items added by researchers. These included individual risk and protective factors. For this paper, we

[3] Frequently reported disabilities were ADHD, speech and communication delays, and teacher-reported learning delays. Common emotional problems, as reported by parents, included having family problems, displaying anger or anxiety issues, or receiving counseling services.

used construct measures developed from the mean score across related items for three domains of self-efficacy and academic achievement (for self-reported and academic records). We also constructed scales for child individual and environmental risk. These measures are described in further detail below.

Grades

Self-reported grade measures in the original YOS were assessed on a 4-item scale in which respondents were asked to recall "marks they are getting in school" by circling the corresponding letter grade in four major subject areas: Mathematics, Reading/Language Arts, Social Studies, and Science. It should be noted that the Reading/Language Arts item was framed as a double-barreled question, even though students in both districts receive distinct grades for Language Arts and Reading. Letter grade options ranged from F (Not Good at All) to A (Excellent). In order to collect objective grade measures, researchers facilitated the development of data sharing agreements between BBBS and the two school divisions in its service region. These included quarterly subject-area grades in Mathematics, Science, Social Studies, Reading and Language Arts. We excluded grades for respondents that were reported on a 3-point scale [Excellent (E), Satisfactory (S) and Not Satisfactory (N)]. Both divisions use the E, S, N scale for kindergarten through second grade students. Because only one of the two school districts used a 12-point grade scale (A, A−, B+, B, B−, C+, C, C−, D+, D, D−, F), all grades were converted to solid letter grades on a 5-point scale (A, B, C, D, F).

Performance Level

Student performance was measured based on a major subjects GPA calculated as the mean across actual student grades in Language Arts, Social Science, Mathematics, and Science. GPA was then converted to a binary scale using 2.0 as a cut off to define higher performing respondents. Forty-five (15.6 %) students in the sample were considered lower performing while 216 (84.4 %) were higher performing students.

Attendance

To assess student attendance, students were asked in the original YOS, "How often, in the past 30 days have you been absent from school?" Response options for number of days absent were presented on a 4-point scale ranging from "Never" to "I did it 3 or more times in the last 30 days." School records on quarterly attendance were also collected. Number of days absent, in both districts, was reported as the number of days per quarter in which the student missed school (approximately 90 day periods). To achieve comparability with the self-reported attendance measure in the YOS, number of days absent was converted to a 4-point categorical scale ranging from "Never" (corresponding with zero absences per quarter) to "3 or more in 30 Days" (corresponding with seven or more absences per quarter).

Self-Efficacy Domains

We used three domains of self-efficacy to capture measures of child psychological characteristics based on constructs developed from the original YOS. A Social Acceptance

construct was measured with six items containing questions related to efficacy in peer relationships such as "popularity," "making new friends," and "doing things" with other kids ($\alpha = .618$). A School Competency construct measured academic self-efficacy with six items including questions such as "I have trouble figuring out answers in school" and "I often forget what I learn" ($\alpha = .641$). A Future Aspirations construct assessed educational expectations toward graduating from high school, going to college, and graduating from college ($\alpha = .845$).

Child Individual and Environment Risk

We also constructed indices to measure student risk levels at baseline. Data were collected from parental enrollment forms, additional items added to the YOS, and school records on wide variety individual and environmental risk factors. Individual risk was calculated by summing reported risk across three broad areas: academic challenges, problem behavior, and mental health concerns (exhibited depressive symptoms). Depressive symptoms were assessed in the YOS using an adapted 20-item tool developed by Radloff (1977) and Radloff and Locke (1986).[4] Similarly, environmental risk was calculated by summing across reported risk in three broad areas: economic adversity, family risk/stress, and peer difficulties. Based on the procedure outlined by Herrera et al. (2013), we defined two dichotomous variables that classified students by individual risk (low/high) and environmental risk (low/high). In the sample, 36.5 % of the students were classified as high individual risk at baseline and 34.3 % were classified as high environmental risk.

Range of Recall

Although there are no national BBBS guidelines for addressing the issue, a clear challenge in YOS administration is the assumption that children will recall performance accurately over long time periods. Thus, in reporting grades, respondents are not given a clear time frame for reflection, such as "thinking back to your last report card." Despite this ambiguity, we assumed students would recall their most recent formal grade assessment. A unique baseline period for assessing response validity of self-reported grades and attendance was established for each respondent based on time of entry into the BBBS program. For example, if a student was matched with a mentor and given a baseline YOS in February 2013 we used school data from the period ending January 2013, 1 month prior to program entry. Among the entire sample population, dependent upon time of program entry, recall time was as recent as a week or as distant as 12 weeks. The average period over which respondents were asked to recall academic performance was 45.1 days, (SD = 36.69) with a maximum of 141 days. We also explored a second measures for recall to assess response validity, using respondents who reported within 30 days of receiving a report card (n = 107; mean = 14.2 days; SD = 8.89). We defined this shorter range of postdictive responding to assess the possible disadvantage of including students with longer recall periods.

[4] In tests conducted among adults, authors found psychometric tests of the depressive inventory indicate scale reliability and validity. Cronbach's alpha ranged from .85 in community samples in .9 in psychiatric samples. Test–retest reliability show moderate correlations ($r = .51$–$.67$).

Procedure

Of the 500 children enrolled in the research study, 330 students were assessed with the YOS at baseline; 170 students ages five to seven were ineligible for YOS assessment due to nationally recommended age restrictions. Although BBBS of America recommends that its agencies only assess children ages nine and above, researchers broadened the parameters to include 8 year-old respondents. Researchers also permitted Match Support Specialists use of individual discretion in determining inclusion for children with special needs, limited English proficiency and low comprehension.[5] Surveys were administered in-person by twelve MSS at the BBBS site. Assistance with reading and responding to questions was offered to each respondent; 81.4 % accepted assistance. In addition, each MSS was instructed to assess whether respondents understood questions asked and knew how to answer appropriately. Four students were identified for low comprehension and were dropped from the study.

Administrative records from the two school divisions were available for 288 students; however, some school records were incomplete. Thus, from the original 330 assessed on the YOS, the final sample was reduced to 257 valid observations in Math, 259 in Science, 256 in Social Science, 241 in Language Arts, and 197 in number of days absent. Quarterly academic data were collected from Spring 2010 through Spring 2013.

Data Analysis

Basic measures of response validity for student self-reports were assessed using percent agreement along with percent of respondents under- and over-reporting. Percent agreement was expressed as the number of accurate responses divided by the total number of observations. This measure, however, has been widely noted as an inadequate measure of validity on its own, due to the fact that it does not correct for accurate answers that may be occurring by chance. As a result, percent agreement is likely to overestimate accuracy. Measures like Kappa and Intra-Class Correlations (ICC), commonly used to assess inter-rater reliability, account for chance agreement and may provide a more realistic assessment of accuracy. Moreover, ICC can be specified to provide an overall measure of consistency, rather than absolute agreement. Thus, independent of the absolute accuracy, we assessed whether students reported in the general direction (meaning higher or lower grades) correctly. Therefore, in addition to percent agreement, we report a consistency average measures ICC, specifying a mixed-effects model with list-wise deletion for missing data.[6] We also report a validity coefficient (Pearson's product-moment correlation coefficient).[7] We anticipated that both consistency and accuracy of reporting among children and youth would fall below criteria for excellent/strong validity. Beyond an initial postdictive validity assessment on the full sample population, we conducted a comparative postdictive assessment accounting for range of recall.

[5] The Director of Programs reviewed all determinations prior to survey administration, which would mitigate any issues with multiple rater consistency in sampling.

[6] Cicchetti (1994) provides commonly-cited ICC cutoffs for qualitative ratings of agreement. Values less than .40 are considered as weak, values between .40 and .59 as fair, values between .60 and .74 as good, and values .75 and higher as excellent.

[7] Our criteria for assessing the validity coefficient was .00–.3 as weak, .3–.59 as moderate, and .6 or above as strong.

Next, we explored the differential effects key moderators may have had on reporting accuracy across academic subjects and days absent using simple one-tailed Fisher's Exact tests. For ethnicity, and other multi-level ordinal indicators, we used a Goodman–Kruskal Tau test. Here, the dependent variables for analysis were binary accuracy scores in each academic subject and number of days absent. Finally, the paper presents equivalent multiple regression models: a model predicting difference between self-reported and actual grades and a model predicting student reported grades. Defining the dependent variable for analysis in the first model involved a multi-step process. First, we established internal consistency reliability across major subjects, using Cronbach's alpha scores. School-reported student grades across four major subjects (Language Arts, Social Science, Science, and Math) were highly consistent ($\alpha = .839$). Grades reported by students at baseline (program entry) were generally less consistent, but acceptable ($\alpha = .677$). Once we established internal consistency, we constructed two GPA measures defined as the average across Language Arts, Mathematics, Social Studies, and Science grades. The mean actual GPA was 2.74 (SD = .9). The mean self-reported GPA was 3.0 (SD = .714). Next, we computed a gain score. As recommended, this variable was constructed by subtracting the standardized reported GPA from the standardized actual GPA, referred to as the DIZ score (De Los Reyes and Kazdin 2004). Scores of zero indicated absolute accurate reporting; positive scores indicated student overreporting. The DIZ variable ranged from −2.86 to 3.9. The mean score was .0162 (SD = .96) (n = 261). As shown in Fig. 1, we observed a roughly normal, continuous DIZ distribution (skewedness = .338; kurtosis = 1.22). In this model, performance level, gender, race and ethnicity, grade level (elementary/middle), range of recall, efficacy constructs, and risk scales were specified as key predictors. Among these factors, we hypothesized that performance level and grade level would be key predictors of agreement.

However, many note the theoretical and practical limitations of relying solely upon difference (DIZ) scores for assessing reporter agreement (Weems et al. 2010; Laird and Weems 2011; Laird and De Los Reyes 2013). Specifically, Weems et al. (2010) note that when using DIZ scores as an outcome variable we risk minimizing the "patterns of associations" among key moderators that can be captured by looking at equivalent

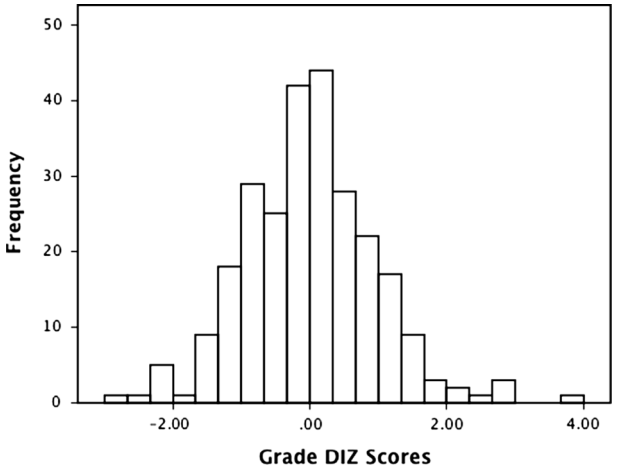

Fig. 1 Frequency of distribution of standardized grade DIZ Scores

regression models for individual reporting sources (p. 394). Difference scores cannot capture countervailing patterns of error (equal proportions of over and under-reporting within subgroups of a key theoretical moderator); therefore, they may mute the effect of important predictors. Thus, we incorporated a moderated regression model of student reported GPA intended to identify cross-informant associations.

Results

Assessing Academic Response Validity Overall

While in most cases we were able to reject a null hypothesis of no relationship between reported and actual academic grades, overall results from the postdictive validity assessment on the entire sample population indicate low to moderate validity coefficients ($r = .284$–$.558$) and fair to good internal consistency in reporting academic subjects (ICC = .440–.716) (See Table 2). Thus, students were more likely to report the general direction of their grades accurately (higher or lower) than they were their actual grades. As anticipated, Language Arts had the lowest correlational strength ($r = .284$) and lowest consistency in reporting (ICC = .440) among the academic subjects. Results for major subjects GPA were more promising. Results indicated a moderate, significant correlation between reported and actual GPAs ($r = .539$) and good internal consistency (ICC = .689).

Results for reporting number of days absent were less positive. Students in the entire population underestimated their absences by a mean difference of 2.19 (based on a 4-point categorical scale). Percent agreement for both samples was roughly 23 %. Both the validity coefficient ($r = .169$) and ICC consistency values (ICC = .308) were weak as well.

Assessing Academic Response Validity Based on Time of Recall

Results from the postdictive validity tests comparing the entire sample population to those enrolled within 30 days of their baseline period demonstrate no improvement in mean

Table 2 Total sample population postdictive criterion validity and reliability

Subject	N	Frequency			Validity		
		% Under-reporting	% Agreement	% Over-reporting	Mean Diff.	*R*	ICC/ (C.I.)
Language Arts	235	23.8 %	35.7 %	40.4 %	.3	.284**	.440/ (.276–.567)
Mathematics	251	15.5 %	44.2 %	40.2 %	.37	.468**	.636/ (.533–.716)
Social Studies	250	26.4 %	43.6 %	30 %	.12	.484**	.650/ (.551–.727)
Science	253	18.2 %	49.4 %	32.4 %	.23	.558**	.716/ (.637–.779)
Days absent	193	59.6 %	23.3 %	17.1 %	−2.19	.169**	.308/ (.048–.478)
Major subjects GPA	255	31 %	16.9 %	52.2 %	.214	.539**	.689/ (.603–.756)

difference, percent agreement, correlation, and general consistency of responses (See Table 3). Students in the 30-day sub-sample overestimated their grades by a mean difference ranging from .26 to .49 compared to a mean difference ranging from .12 to .37 in the entire sample population. Percent agreement in academic subjects among the entire sample population ranged from 35.7 to 49.4 %. With the exception of Language Arts, this was consistently higher than the 30-day sub-sample. In addition, we did not observe improvements in accuracy or consistency of reporting absences within the 30-day sub-sample. Based on these results, we can conclude low to moderate postdictive validity for academic reporting by subject area, regardless of the recall period. To confirm, we ran an independent *t* test, testing difference in GPA gain by range of recall. Results indicated a significant difference in reporting accuracy ($p = .039$). Those reporting after 30 days were significantly more accurate than those reporting within a 30-day window. The difference in GPA gain between the groups was .184. Thus, we reject our null of no effect of range of recall on response validity.

Isolating the Impact of Moderating Factors on Response Validity

Impact of Individual Moderators by Academic Subject Area and Attendance

To assess the factors that affect response validity by academic subject area and attendance, we computed simple one-tailed Fisher's Exact and Goodman–Kruskal Tau tests across theoretically-relevant characteristics: gender, grade-level, race and ethnicity, performance level, and range of recall. Our null hypotheses stated that there is no relationship between our dependent variables (grades/number of days absent) and these characteristics. Results are reported in Table 4. We found no relationship between student gender and accuracy of reporting across the four academic subjects and number of days absent. While our descriptive statistics indicate girls tend to report with higher accuracy than boys, gender is not a statistically significant predictor. Based on the literature, we anticipated that older

Table 3 Comparative postdictive criterion validity and reliability 30 day sub-sample

Subject	N	Frequency			Validity		
		% Under-reporting	% Agreement	% Over-reporting	Mean Diff.	*R*	ICC/(C.I.)
Language arts	101	23.8 %	35.6 %	40.6 %	.302	.161	.267/ (−.074–.512)
Mathematics	107	15 %	37.4 %	47.7 %	.48	.398**	.567/ (.365–.705)
Social studies	106	25.5 %	36.8 %	37.7 %	.233	.452**	.454/ (.189–.633)
Science	105	17.1 %	44.8 %	38.1 %	.292	.525**	.687/ (.539–.787)
Days absent	83	56.6 %	23.5 %	19.3 %	−.769	.124	.208/ (−.244–.488)
Major subjects GPA	107	28 %	14 %	57.9 %	.329	.538**	.689/ (.543–.788)

Table 4 Percent accuracy of self-reported grades by respondent characteristics

Characteristic		Language Arts		Mathematics		Social studies		Science		Days absent	
		% Agreement	Sig.	% Agreement	Sig.	% Agreement	Sig.	% Agreement	Sig.	% Agreement	Sig.
Gender	Male	26.5 %	.295	32.7 %	.094	34.7 %	.265	40.8 %	.319	14.7 %	.492
	Female	30.4 %		41.4 %		39.3 %		44.5 %		16.5 %	
Grade by Group	Elementary	31.4 %	.164	35.7 %	.108	34.7 %	.08	44.9 %	.245	18.2 %	.081
	Middle	24.7 %		44.1 %		44.1 %		39.8 %		11 %	
Ethnicity*	White	30.8 %	.336	44.2 %	.032*	41.7 %	.4	46.7 %	.733	18.6 %	.404
	Black	25.6 %		46.5 %		32.6 %		37.2 %		11.9 %	
	Hispanic	26.9 %		28.6 %		34.5 %		42 %		15.5 %	
	Other	57.1 %		57.1 %		57.1 %		42.9 %		0 %	
Performance	Low <2.0 GPA	23.6 %	.001**	35.7 %	.014*	26.4 %	.00**	36.4 %	.00**	10 %	.252
	High >2.0 GPA	41.8 %		50 %		59 %		60.7 %		18.1 %	
Time lapse from Reporting	<30 days	31.6 %	.352	35.1	.130	34.2 %	.119	41.2 %	.216	17.5 %	.289
	> 30 Days	28.7 %		42.5		41.9 %		46.7 %		14.4 %	

* p is significant at .05 level; ** p is significant at .01 level or lower

students would have greater ability to recall accurately. However, at 95 % confidence, results indicate there is no significant relationship between grade level and reporting accuracy across all five measures. Middle school students, however, do report more accurately in Social Studies and number of days absent using a 90 % confidence level. There was a significant relationship between student race and ethnicity and Mathematics reporting. Hispanic students reported less accurately (28.6 %) than Whites (44.3 %), Blacks (46.5 %), and Other (57.1 %) students. This trend did not persist across other academic subjects and days absent. Despite lack of evidence across other moderators, we were able to reject a null of no relationship for student performance across all four academic indicators. Higher performing students were consistently more accurate in reporting academic grades. For example, 60.7 % of higher performing students accurately reported their Science grades versus 36.4 % of lower performing students.

Collective Impact of Moderators on GPA Difference Scores

Prior to standardizing, student-reported major subjects GPA was higher (mean = 3.0; SD = .71) than actual major subjects GPA (mean = 2.74; SD = .9), paired t (260) = −5.39, $p = .000$. The correlation matrix demonstrated a moderate association between reported and actual GPAs ($r = .539$). Student reports were positively associated with the DIZ score ($r = -.484$) while actual scores were negatively associated with the DIZ score ($r = -.477$).

To identify factors predicting agreement between actual and child reported grades, we used a multiple regression model of standardized difference scores. Here, student performance level, gender, ethnicity, grade level, recall time, three domains of efficacy measured in the YOS, and our risk indicators are used as predictors. Results presented in Table 5 demonstrated that performance, grade level, and school competence were significant predictors of agreement. Specifically, students with lower performance in major subjects (GPA < 2.0) were more likely to overreport their grades (t = −8.832; $p = .000$). Thus, we can reject the null of no effect of performance on reporting agreement. Holding all other variables constant, the predicted DIZ score among students with higher performance was .063 compared with lower performers at 1.29. Moreover, standardized coefficients indicated performance had the greatest effect (standardized β = −.505) on agreement. Elementary aged children were also more likely to overreport in comparison to middle

Table 5 Summary of regression analysis predicting differences (diz scores)

Variables in the equation	β	*t*	*P*
(Constant)	−.198	−.561	.575
Performance (<2.0 GPA)	−1.234	−8.832	.000**
Gender (male)	−.143	−1.408	.161
Grade level (elementary)	−.305	−2.906	.004**
Ethnicity (white)	.106	1.113	.267
Days from baseline (<30 days)	−.178	−1.863	.064
Social acceptance	−.124	−1.387	.167
School competence	.569	5.848	.000**
Future aspirations	.053	.721	.472
Individual risk (low)	.029	.307	.759
Environmental risk (low)	−.017	−.168	.867

$R^2 = .353$; $p = .000$

school students (t = −2.91; p = .004; standardized β = −.171). For example, holding all other variables constant, the predicted DIZ score for middle school age students was .992 a difference of a .0305 compared to elementary-aged reporters. Thus, we reject the null of no effect of grade level on reporting agreement. While both groups were likely to overreport, elementary aged students were more so. Inversely, those with higher self-efficacy related to School Competence were more likely to overreport than those with low to moderate perceived School Competence (t = 5.84; p = .000; standardized β = .373).

However, as discussed, predicting agreement using standardized difference scores is a limited approach to understanding patterns of reporting error. These limitations were observed in looking at the correlations between standardized reported and actual GPAs across key moderating predictors. For example, the correlation between standardized reported and actual GPAs among elementary aged children was (r = .488) and for middle school ages was (r = .603). The correlation for higher performers was (r = .506) and for lower performers was (r = .123). Thus, a moderated multiple regression of standardized reported GPAs was used to test a variety of interaction (product-term) effects that may not have been detected in DIZ model. The final model, presented in Table 6, shows only significant predictors and moderators (r^2 change indicated a .134 improvement when including interaction terms, p = .000). Here, standardized actual GPAs and continuous predictors were mean-centered to reduce multicollinearity. However, standardized actual GPA was ultimately removed from the final model due to persistently high multicollinearity (VIF scores above 4) with the performance level indicator. Similar to the DIZ model, results indicated performance, grade level and School Competency were all predictive of child reported grades. Results indicated the interaction terms for performance with actual GPA (t = 7.462; p = .000) and school competence with actual GPA (t = −2.906; p = .004) were significant predictors of reported GPA. Other interaction terms tested such as actual GPA with gender, recall time, and ethnicity were not significant despite compelling correlational values. For example, among boys the correlation between reported and actual GPAs was (r = .450) and for girls was (r = .576). Among White children the correlation was (r = .633) while for other ethnicities it was (r = .454).

Given that findings indicated performance level was most predictive of reporting agreement, we directed further investigation to exploring the directionality of systematic reporting error by performance type. Figure 2 displays the simple interaction effect of performance by actual GPA. Note, in a simple bivariate regression model, actual GPA was not predictive of reported GPA among low performing students (r^2 = .015; F = .660, p = .421). However, among higher performing students, actual GPA was predictive of reported GPA (r^2 = .256; F = 73.65, p = .000). For every one unit change in actual grade, we estimated a .687 unit change in reported grades (p = .000). As seen in Fig. 2, the explained variability in reported GPA among higher performers is 25.6 % compared to 1.5 % among lower performers.

Table 6 Summary of moderated regression analysis predicting child reported grade

Variables in the equation	β	t	p
(Constant)	−.383	−3.011	.003**
Performance (<2.0 GPA)	.479	3.62	.000**
Grade (elementary)	−.31	−3.031	.003**
School competence	.665	6.93	.000**
Performance * actual GPA	.58	7.462	.000**
School competence * actual GPA	−.264	−2.906	.004**

R^2 = .430; p = .000; R^2 change = .134; p =. 000

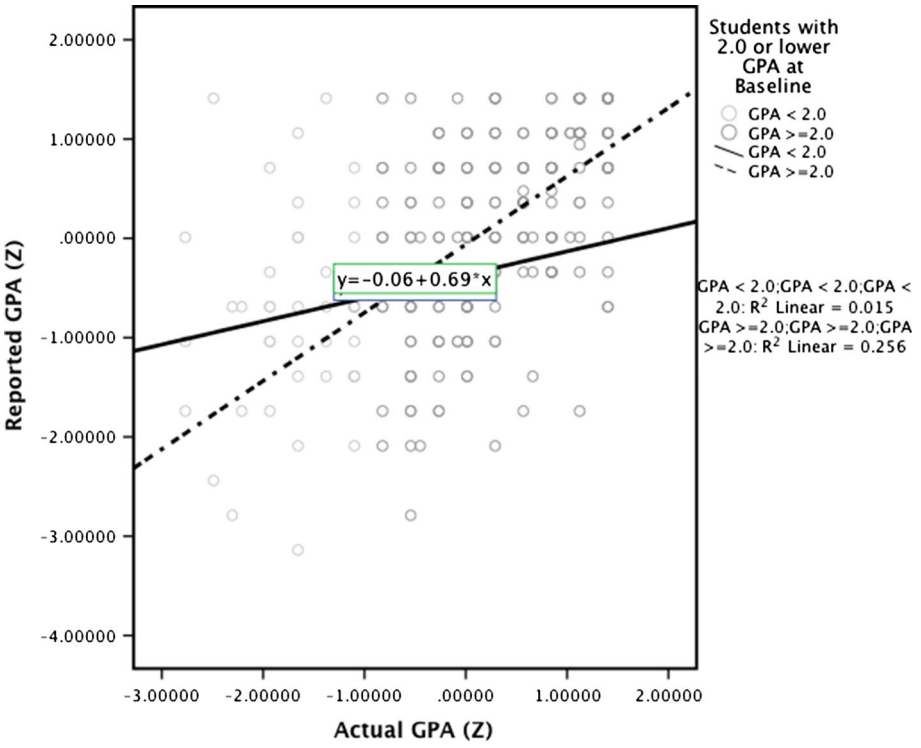

Fig. 2 Performance level as a product-term moderator

Discussion

While previous studies have found strong to excellent academic response validity among high school and college students, this research focused on children ages 7 to 15. In this study, response validity was weak to moderate across five indicators: Language Arts, Mathematics, Social Science, Science, and number of days absent. ICC findings, assessing general consistency of responses, were slightly more positive. Yet, both measures indicate that response validity among children and youth is much lower than among high school and college-age students. Among the academic subjects, student-reported Language Arts grades had the lowest correlational strength and consistency among academic subjects. While this may be due to double-barreled question framing, some have noted that lower correlational strength may be driven by lack of clarity in feedback provided to students in English compared with other academic subjects (Blatchford 1997). Despite poor to moderate results in individual academic subject reporting, overall major subjects GPA results indicated good response validity. How can we understand this in light of less acceptable individual subject correlations? Major subjects GPA effectively adjusts for random error in individual grade reporting by cumulating what may be some random over- and under-reporting among respondents. Thus, while students may be underreporting more in Social Studies and Language Arts, they could also be overreporting in Math and Science. This would lead to a more accurate GPA measure, despite generally inaccurate reporting. Others have confirmed this trend (Escribano and Díaz-Morales 2014; Försterling and Binser 2002).

Results were consistent across the entire sample population with a mean recall period of 45.1 days as well as a smaller sub-sample with a 14.1 day average. Although intuition would suggest students reporting within 30 days of their most recent report card would produce more valid results, we found some preliminary evidence that a longer range of recall may promote greater accuracy. There is some rational which would suggest these results are driven by social desirability bias. Students may be less likely to report recent grades accurately but more likely to report historic grades accurately when social pressure diminishes. We recommend further investigation of this finding.

Results from both the DIZ multiple regression model and moderated child-reported GPA model indicated the key predictor of systematic inaccuracy in academic grade reporting was student performance. Students with lower grades tend to inflate their grades (as would be more socially desirable), whereas students who perform well tend to be more accurate. Additionally, those with higher School Competence tend to report less accurately. This confirms previous research suggesting self-efficacy is a key moderator for response accuracy. In this case, however, higher efficacy related to intellectual ability can override school reports. Finally, this study confirms the intuition that age matters; our findings indicated middle school students reported more accurately than elementary students. It is unclear whether this distinction is driven by higher levels of social desirability bias among younger children or issues related to memory and cognitive development. We would recommend further research to explore this issue. For now, practitioners should note this finding as an added challenge in relying solely upon self-reported academic ability among children. We also concluded that gender and ethnicity did not impact response validity. We did not detect systematic differences among either moderator.

Research limitations stem from reliance on an inflexible nationally-adopted assessment instrument and some issues with study design. Question framing in the original YOS, particularly the double-barreled Language Arts question and the attendance measure, lacked clarity as confirmed by results. The YOS attendance measure assesses categorical frequency of absenteeism within a 30-day window. Attendance information is not commonly reported to students or parents in this format. To compensate, we transformed a continuous school record of attendance into a 4-level categorical measure, which inherently reduces information. Rather, we would propose adding an open-ended or interval measurement to the YOS. As well, the range of recall for academic reporting is not clearly specified to respondents. We recommend clearly aligning recall periods with benchmarks that respondents are more likely to remember, such as "on your most recent report card." These problems are likely to exacerbate potential issues with error in reporting. Despite findings that indicate no improvement in reporting accuracy when reporting within a 30-day period, span of recall remains a major concern, especially among younger children. Further research should look to confirm findings using a concurrent reporting design.

In closing, self-reports among children and youth are widely used as outcome and diagnostic measures. Our findings, however, indicate that such data is a weak proxy for actual academic achievement, particularly for younger, underperforming students. This presents a challenge for youth development and prevention-oriented programs that rely on self-reported measures either to target support services or to measure program impact. Practitioners should be cautious in adopting self-reported measures for academic achievement as the primary indicator for outcomes reporting, especially for lower performing students. Moreover, there is evidence to suggest that younger respondents are not accurate reporters of individual risky behaviors like school absence. If students are not likely to report absences accurately, we might infer reduced likelihood for accurately reporting other socially unacceptable behaviors. However, unlike academic achievement,

many of these behaviors are difficult to verify with objective measures. Thus, while self-reports may be necessary when looking at attitudinal or behavioral measures, when possible, youth programs should attempt to collect academic records either from participant-provided report cards or directly from their schools.

Acknowledgments This project was funded by a Mentoring Best Practices Research Grant award from the Office of Juvenile Justice and Delinquency Prevention (Grant No. Q215F120107). The authors would like to acknowledge the hard work and dedication of the staff at Big Brothers Big Sisters Harrisonburg Rockingham County, their graduate assistants in the James Madison University Master of Public Administration program, and Dr. Gary Kirk, who was a Principle Investigator on the project from 2011–2013.

References

Alexander, K. L., Entwisle, D. R., & Bedinger, S. D. (1994). When expectations work: Race and socioeconomic differences in school performance. *Social Psychology Quarterly, 57*, 283–299. http://spq.sagepub.com/

Anaya, G. (1999). Accuracy of self-reported test scores. *College and University, 75*(2), 13–19.

Arthur, M. W., Hawkins, J. D., Pollard, J., Catalano, R. F., & Baglioni, A. J, Jr. (2002). Measuring risk and protective factors for substance use, delinquency, and other adolescent problem behaviors: The communities that care youth survey. *Evaluation Review, 26*, 575. doi:10.1177/0193841X0202600601.

Bahrick, H. P., Hall, L. K., & Berger, S. A. (1996). Accuracy and distortion in memory for high school grades. *Psychological Science, 7*, 265–271. doi:10.1111/j.1467-9280.1996.tb00372.x.

Bahrick, H. P., Hall, L. K., & Dunlosky, J. (1993). Reconstructive processing of memory content for high versus low test scores and grades. *Applied Cognitive Psychology, 7*, 1–10. doi:10.1002/acp.2350070102.

Blatchford, P. (1997). Students' self assessment of academic attainment: Accuracy and stability from 7 to 16 years and influence of domain and social comparison group. *Educational Psychology: An International Journal of Experimental Educational Psychology, 17*(3), 345–359. doi:10.1111/j.2044-8279.1997.tb01235.x.

Bowman, N. A., & Hill, P. L. (2011). Measuring how college affects students: Social desirability and other potential biases in college student self-reported gains. *New Directions for Institutional Research, 150*, 73–85. doi:10.1002/ir.390.

Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. *Child Development, 61*, 201–210. doi:10.1111/j.1467-8624.1990.tb02772.x.

Cassady, J. C. (2001). Self-reported GPA and SAT: A methodological note. *Practical Assessment, Research & Evaluation 7*(12). http://pareonline.net/Home.htm

Catalano, R. F., Berglund, M. L., Ryan, J. A. M., Lonczak, H. S., & Hawkins, J. D. (2002). Positive youth development in the United States: Research findings on evaluations of positive youth development programs. *Prevention & Treatment, 5* (15). doi: 10.1037/1522-3736.5.1.515a

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment 6*(4), 284–290. http://www.apa.org/pubs/journals/pas/

Cole, J. S., Rocconi, L., & Gonyea, R. M. (2012). Accuracy of Self-Reported Grades: Implications for Research." Paper presented at the annual meeting of the Association for Institutional Research, New Orleans, Louisiana. http://cpr.iub.edu/uploads/AIR%202012%20Cole%20Rocconi%20Gonyea.pdf

Connell, J., & Ilardi, B. C. (1987). Self-system concomitants of discrepancies between children's and teachers' evaluations of academic competence. *Child Development, 58*, 1297–1307. doi:10.2307/1130622.

Crockett, L. J., Schulenberg, J. E., & Peterson, A. C. (1987). Congruence between objective and self-reported data in a sample of young adolescents. *Journal of Adolescent Research, 2*(4), 383–392. doi:10.1177/074355488724006.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*, New York: Wiley.

De Los Reyes, A. (2011). More than measurement error: Discovering meaning behind informant discrepancies in clinical assessment of children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 40*, 1–9. doi:10.1080/15374416.2011.533405.

De Los Reyes, A., & Kazdin, A. E. (2004). Measuring informant discrepancies in clinical child research. *Psychological Assessment, 16*, 330–334. doi:10.1037/1040-3590.16.3.330.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant descrepencies in the assessment of child psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*(4), 483–509. doi:10.1037/0033-2909.131.4.483.

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. A. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, *9*, 123–149. doi:10.1146/annurev-clinpsy-050212-185617.

Dobbins, G. H., Farh, J. L., & Werbel, J. D. (1993). The influence of self-monitoring on inflation of grade-point averages for research and selection purposes. *Journal of Applied Social Psychology, 23*(4), 321–334. doi:10.1111/j.1559-1816.1993.tb01090.x.

DuBois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002). Effectiveness of mentoring programs for youth: A meta-analytic review. *American Journal of Community Psychology, 30*(2), 157–197.

DuBois, D. L., Portillo, N., Rhodes, J. E., Silverthorn, N., & Valentine, J. C. (2011). How effective are mentoring programs for youth? A systematic assessment of the evidence. *Psychological Science in the Public Interest, 12*(2), 57–91. doi:10.1177/1529100611414806.

Dunnette, M. D. (1952). Accuracy of students' reported honor point averages. *Journal of Applied Psychology, 26*, 20–22. doi:10.3102/00346543075001063.

Escribano, C., & Díaz-Morales, J. F. (2014). Are self-reported grades a good estimate of academic achievement?/Son las notas auto-informadas una buena estimación del rendimiento académico? *Estudios de Psicología: Studies in Psychology, 35*(1), 168–182. doi:10.1080/02109395.2014.893650.

Fetters, W. B., Stowe, P. S., & Owings, J. A. (1984). *Quality of responses of high school students to questionnaire items*. Washington, DC: National Center for Education Statistics. http://nces.ed.gov.

Försterling, F., & Binser, M. J. (2002). Depression, school performance, and the veridicality of perceived grades and causal attributions. *Personality and Social Psychology Bulletin, 28*(10), 1441–1449. doi:10.1177/014616702236875.

Frucot, V. G., & Cook, G. L. (1994). Further research on the accuracy of students' self-reported grade point averages, SAT scores, and course grades. *Perceptual and Motor Skills, 79*, 743–746.

Goldman, B. A., Flake, W. L., & Matheson, M. B. (1990). Accuracy of college students' perceptions of their SAT scores, high school and college grade point averages relative to their ability. *Perceptual and Motor Skills, 70*, 514.

Gonyea, R. M. (2005). Self-reported data in institutional research: Review and recommendations. *New Directions for Institutional Research, 127*, 73–89. doi:10.1002/ir.156.

Grossman, J. B. (2009). Evaluating Mentoring Programs. *Public/Private Ventures*. Retrieved April 15 2014. http://ppv.issuelab.org

Hall, G., Yohalem, N., Tolman, J., & Wilson, A. (2003). *How afterschool programs can most effectively promote positive youth development as a support to academic achievement: A report commissioned by the Boston After-School for All Partnership.* National Institute on Out-of-School Time. Retrieved July 6 2014. http://www.vamentoring.org

Hamilton, L. C. (1981). Sex differences in self-report errors: A note of caution. *Journal of Educational Measurement, 18*(4), 221–228. doi:10.1111/j.1745-3984.1981.tb00855.

Herrera, C., DuBois, D. L., & Grossman, J. B. (2013). The role of risk: Mentoring experiences and outcomes for youth with varying risk profiles. *MDRC*. www.mdrc.org

Kaderavek, J. N., Gillam, R. B., Ukrainetz, T. A., Justice, L. M., & Eisenberg, S. N. (2004). School-age children's self-assessment of oral narrative production. *Communication Disorders Quarterly, 26*(1), 37–48. doi:10.1177/15257401040260010401.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry, 160*, 1566–1577. doi:10.1176/appi.ajp.160.9.1566.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*(1), 63–82. doi:10.3102/00346543075001063.

Laird, R. D., & De Los Reyes, A. (2013). Testing informant discrepancies as predictors of adolescent psychopathology: Why difference scores cannot tell you what you want to know and how polynomial regression may. *Journal of Abnormal Child Psychology, 41*, 1–14. doi:10.1007/s10802-012-9659-y.

Laird, R. D., & Weems, C. F. (2011). The equivalence of regression models using difference scores and models using separate scores for each informant: Implications for the study of information discrepancies. *Psychological Assessment, 23*, 388–397. doi:10.1037/a0021926.

Martin, C. L., & Nagao, D. H. (1989). Some effects of computerized interviewing on job applicant responses. *Journal of Applied Psychology*, *74*, 72–80. http://psycnet.apa.org

Mayer, R. E., Stull, A. T., Campbell, J., Almeroth, K., Bimber, B., Chun, D., & Knight, A. (2007). Overestimation bias in self-reported SAT scores. *Educational Psychology Review, 19*(4), 443–454. doi:10.1007/s10648-006-9034-z.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385–401. doi:10.1177/014662167700100306.

Radloff, L. S., & Locke, B. Z. (1986). The community mental health assessment survey and the CES-D Scale. In M. M. Weissman, J. K. Myers, & C. E. Ross (Eds.), *Community surveys of psychiatric disorders* (pp. 177–189). New Brunswick, NJ: Rutgers University Press.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Research, Evaluation & Assessment, 11*(10), 1–13. http://pareonline.net

Sawyer, R., Laing, J. & Houston, W. (1988). Accuracy of self-reported high school courses and grades of college-bound students. *ACT Research Report Series, 88*(1), ii-32). Iowa City, IA: American College Testing Program. www.act.org

Schiel, J., & Noble, J. (1991). Accuracy of self-reported course work and grade information of high school sophomores. *ACT Research Report Series. 91*(6). Iowa City, IA: American College Testing Program. www.act.org

Shaw, E. J. and Mattern, C. D. (2009). Examining the accuracy of self-reported high school grade point average. *College Board Research Report No. 2009-5.* http://research.collegeboard.org

Shepperd, J. A. (1993). Student derogation of the Scholastic Aptitude Test: Biases in perceptions and presentations of College Board scores. *Basic and Applied Social Psychology,* 14, 455–473. http://www.psych.ufl.edu

Talento-Miller, E., & Peyton, J. (2006). Moderators of the accuracy of self-report grade point average. *Graduate Management Admission Council Research Reports RR-06-IO.* McLean, Virginia. Retrieved June 30 2014 from http://www.gmac.com

Thompson, L. A., & Kelly-Vance, L. (2001). The impact of mentoring on academic achievement of at-risk youth. *Children and Youth Services Review, 23*(3), 227–242. doi:10.1016/S0190-7409(01)00134-7.

Weems, C. F., Taylor, L. K., Marks, A., & Varela, R. E. (2010). Anxiety sensitivity in childhood and adolescence: Parent reports and factors that influence associations with child reports. *Cognitive Therapy and Research, 34*, 303–315. doi:10.1007/s10608-008-9222-x.

Zimmerman, M. A., Caldwell, C. A., & Bernat, D. H. (2002). Discrepancy between self-report and school-record grade point average: Correlates with psychosocial outcomes among African American adolescents. *Journal of Applied Social Psychology, 32*(1), 86–109. doi:10.1111/j.1559-1816.2002.tb01421.x.