




OPEN

Large studies reveal how reference bias limits policy applications of self-report measures

Benjamin Lira^{1,7}, Joseph M. O'Brien^{2,7}, Pablo A. Peña³, Brian M. Galla⁴, Sidney D'Mello⁵, David S. Yeager², Amy Defnet⁶, Tim Kautz⁶, Kate Munkacsy⁶ & Angela L. Duckworth¹

There is growing policy interest in identifying contexts that cultivate self-regulation. Doing so often entails comparing groups of individuals (e.g., from different schools). We show that self-report questionnaires—the most prevalent modality for assessing self-regulation—are prone to *reference bias*, defined as systematic error arising from differences in the implicit standards by which individuals evaluate behavior. In three studies, adolescents ($N = 229,685$) whose peers performed better academically rated themselves lower in self-regulation and held higher standards for self-regulation. This effect was not observed for task measures of self-regulation and led to paradoxical predictions of college persistence 6 years later. These findings suggest that standards for self-regulation vary by social group, limiting the policy applications of self-report questionnaires.

Self-regulation refers to a diverse set of personal qualities, distinct from cognitive ability, that enable individuals to set and pursue goals. The terminology favored for self-regulation and its facets varies across the literatures of child development (e.g., effortful control, ego strength)^{1–3}, adult personality (e.g., Big Five conscientiousness)⁴, psychopathology (e.g., impulse control)⁵, and economics (e.g., temporal discounting)^{6,7}. Such diverse traditions in behavioral science have directed this attention because individual differences in self-regulation predict later life outcomes, including academic performance^{8–10}, physical and mental health^{11–13}, well-being and life satisfaction¹⁴, civic and social behavior^{12,15}, job performance¹⁶, earnings^{12,17–19}, and wealth^{12,17}. Moreover, the effects of self-regulation are independent of, and comparable in magnitude to, cognitive ability and family socioeconomic status (SES)^{8,12}.

A half-century of basic research suggests that self-regulation develops optimally in caring environments that encourage adaptive goal-relevant knowledge (e.g., strategies for managing attention), beliefs (e.g., that emotion and motivation can be regulated), and values (e.g., that self-regulation is important)²⁰. This development extends far beyond early childhood, when children are mostly in the company and care of parents. Indeed, adolescence may be particularly important for supporting self-regulation because of the rapid growth, learning, adaptation, and neurobiological development that mark this period of life^{21–23}. Further, impulsive choices in adolescence (e.g., to start smoking, to drop out of school) can alter life trajectories in ways that are difficult to reverse¹².

Schools are a natural target for policy because of their potential to provide equal access to environments that support the development of self-regulation^{24,25}. Not only is school where young people spend most of their waking hours outside the home, it is also where they experience a multitude of factors that have been shown to either scaffold or stymie the development of self-regulation, including adult role models^{26,27} and peers^{28,29}. Recently, a growing chorus of policymakers has urged schools to extend their purview beyond traditional academic coursework and into the domain of social-emotional skills such as self-regulation—a trend that is reflected in the expanded scope of federal and state standards and accountability systems^{30–32}.

In this investigation, we identify a pervasive measurement bias that, if not remedied, may thwart policymakers' efforts to evaluate, measure, and improve the effectiveness of schools that foster adolescent self-regulation. The possibility of this measurement bias has led to serious questions from policymakers about "whether we can make [self-regulation skills] visible, comparable, and therefore amenable to deliberate policy action in a similar way that traditional tests do with academic knowledge and skills"³³. As a result, education systems have been left with great interest in self-regulation and related constructs—but insufficient scientific guidance.

¹University of Pennsylvania, Philadelphia, USA. ²University of Texas at Austin, Austin, USA. ³University of Chicago, Chicago, USA. ⁴University of Pittsburgh, Pittsburgh, USA. ⁵University of Colorado-Boulder, Boulder, USA. ⁶Mathematica, Inc., Princeton, USA. ⁷These authors contributed equally: Benjamin Lira and Joseph M. O'Brien. ✉email: blira@sas.upenn.edu



Figure 1. Peers influence the standards by which an individual judges their own behavior, resulting in a “reference bias” effect that distorts cross-context comparisons of self-reported self-regulation. Illustration by Tom McQuaid.

The empirical starting point for our research is the mixed and often counterintuitive evidence regarding school effects on self-regulation. On one hand, Jackson et al.³⁴ show encouraging evidence that schools can differ in how much they improve students’ scores on a self-report measure of hard work, and these school differences predicted students’ later college enrollment and persistence. On the other hand, evaluations of charter schools show that they fail to raise self-reports of self-regulation, despite raising report card grades, standardized test scores, attendance rates, and college enrollment levels while reducing incarceration and unplanned pregnancies^{35–38}. Are high-performing schools whose cultures explicitly emphasize hard work and high expectations^{39,40} in fact having no impact on students’ self-regulation—or is there a problem in how self-regulation is measured?

We suggest that reference bias, the systematic error that arises when respondents refer to different implicit standards when answering the same questions⁴¹, is a legitimate threat to between-school comparisons and can help explain the conflicting evidence of school effects on self-regulation. Moreover, we contend that even within a school, comparisons of students are biased when different subgroups of students rely on different standards when answering the same questions. In the present policy context, reference bias is especially pernicious because it is difficult to detect and diagnose. Unlike social desirability bias, modesty bias⁴², faking, and response style biases⁴³, reference bias can emerge even when respondents answer truthfully, and it can coexist with otherwise strong validity associations at the individual level. This is because reference bias can distort inferences any time there are comparisons of self-regulation across *groups* who differ in their frames of references—for example, schools with very different peer cultures with respect to effort, or even subcultures within a school.

Why might self-report questionnaires be subject to reference bias? Dominant models in survey methodology identify a multi-stage cognitive response process: students first read and interpret the question; then they identify relevant information in memory, form a summary judgment, and translate this judgment into one of the response options; finally, they edit their response if motivated to do so^{44–46}. As illustrated in Fig. 1, a student may interpret a questionnaire item and its response options differently depending on their peers’ typical behaviors⁴⁷. If they have high-achieving classmates who, for example, study for hours each evening and consistently arrive prepared for class, they might judge themselves against higher standards and rate themselves lower in self-regulation than an equally industrious student whose lower-achieving peers study and prepare less. While schools might be effective in increasing self-regulated behavior, they might at the same time increase the standards, leading to lower self-reported self-regulation.

A well-established research literature has demonstrated that the subjective view students hold of themselves, both in general terms (i.e., self-esteem) and in the realm of academic performance (i.e., academic self-concept) depends upon peer comparisons^{42,48,49}. In particular, the Big Fish Little Pond Effect (BFLPE) refers to the lower academic self-concept of students in higher-achieving schools⁵⁰. A related and older literature on social comparison has demonstrated that in general, people spontaneously compare themselves to other people, especially to people who are superior to them in some way, which can lower their subjective appraisal of their own ability⁵¹. Finally, there is evidence that academic self-concept and standardized test scores are positively correlated

within countries but inversely correlated between countries—a phenomenon dubbed the attitude-achievement paradox^{42,52}. In sum, there is ample evidence for the influence of peers on inherently subjective constructs.

In contrast, evidence that reference bias distorts comparisons of self-regulation across social groups has been indirect. A handful of cross-cultural studies have yielded paradoxical findings (e.g., Asian countries such as Japan and South Korea ranking lower in self-reported conscientiousness than other countries that are typically thought to be less conscientious⁵³), but none of these studies directly measured standards for behavior, relying instead on experts' ratings of cultural stereotypes or indirect proxies for self-regulation (e.g., the average walking speed in downtown locations of a convenience sample of a country's residents, as a proxy for the nation's conscientiousness).

In the educational literature, studies that compare the test scores and average self-regulation scores for different schools have not ruled out unobserved confounds, such as the possibility that school factors (e.g., average family income) that increase test scores (e.g., due to investment in educational opportunities) also decrease self-regulation (e.g., by shielding children from responsibilities that could cultivate self-regulation). Therefore, the research literature to date has not been able to distinguish biases in self-reports from potentially true group differences in self-regulation.

In this investigation, we overcome these limitations by using three complementary methods to examine reference bias more directly than has been possible previously. Our approach is motivated by the basic finding that people judge themselves compared to salient and similar others⁴⁷. Therefore we exploit (Studies 1 and 2) or work around (Study 3) variation in people's reference groups.

In Study 1 (total $N = 206,589$ students in $k = 562$ Mexican high schools), we show that the reference bias effect appears even within the same school in a year-over-year comparison. When students are surrounded by higher-achieving peers relative to other students at the same school in a different year, they rate themselves lower in self-regulation. Study 2 addresses an additional confound that could remain in Study 1's analysis, which is the possibility that year-over-year fluctuations in test scores are not random but are due to choices made by families about the academic trajectory of the school. In Study 2 ($N = 21,818$ students in $k = 62$ U.S. secondary schools), we rule this out with an analysis rooted in the purported psychological explanation for reference bias, which is that people's self-judgments should be more influenced by the peers whose behaviors they observe rather than peers whose behaviors they do not observe. We show that reference bias is evident in data from a single school year only when administrative data showed that the peers shared classes and therefore had an opportunity to observe each other's self-regulated behavior. Furthermore, Study 2 examined the theorized, but typically unmeasured, explanation for reference bias: differences in students' implicit standards for self-regulation (i.e., how many hours of homework constitute “a lot of homework” and how often it means to “sometimes” forget what they need for class).

Studies 1 and 2 argue against school-level alternative explanations for reference bias but nevertheless allowed for the possibility that high-achieving peers reduce a student's real capacity for self-regulation. Study 3 ($N = 1278$ seniors in $k = 15$ U.S. high schools) addressed this possibility with a workaround: an objective behavioral task that involves no self-reports and therefore is not subject to biases due to differences in frames of reference. By matching self-regulation data collected in high school to records of college graduation, we show that there is no evidence of reference bias when a behavioral task is used. This evidence is bolstered by Study 3's use of a measure of school achievement that is independent of the high school peer group: graduation from college within 6 years after high school completion.

Study 1: Evidence for reference bias in a country-wide natural experiment

In 2012 and 2013, the Secretariat of Public Education administered questionnaires measuring grit (the passion and perseverance for long term-term goals⁵⁴) and collected data on academic performance from high school seniors in a nationally representative sample of 10% of high schools in Mexico. We analyzed data from the 1% of all schools that, by chance, were selected in both years. This enabled us to exploit exogenous variation in the academic performance of the 2013 high school cohort when compared to the performance of the 2012 cohort. Reference bias was quantified as the effect on self-reported grit uniquely attributable to peer academic performance (i.e., the cohort-wide averages of GPA, standardized math test scores, and standardized reading test scores, respectively, excluding said student from the average), after controlling for differences between schools, cohort year, and each student's own academic performance.

Methods. *Sample and procedure.* High school seniors in two representative random samples, each comprising 10% of schools in Mexico, completed standardized achievement tests of math and reading and, separately, self-report questionnaires late in the spring term of the 2011–2012 and 2012–2013 academic years, respectively. By chance, about 1% ($k = 562$) of high schools were included in both years. Our final sample includes 97.8% of the students in these high schools ($N = 206,589$) who completed a questionnaire measure of grit. There were slightly more girls than boys in our sample (53.49% female). On average, students in our sample were 17.61 years old ($SD = 0.79$).

Self-reported grit. The Technical Committee for Background Questionnaires at the National Center of Evaluation for Higher Education in Mexico (Centro Nacional de Evaluación para la Educación Superior) translated all 8 items of the Short Grit Scale⁵⁵ as well as its 5-point Likert-type response scale (1 = *Not at all like me* to 5 = *Very much like me*) into Spanish. The observed reliability was $\alpha = 0.62$. All reported reliabilities are Cronbach's alphas.

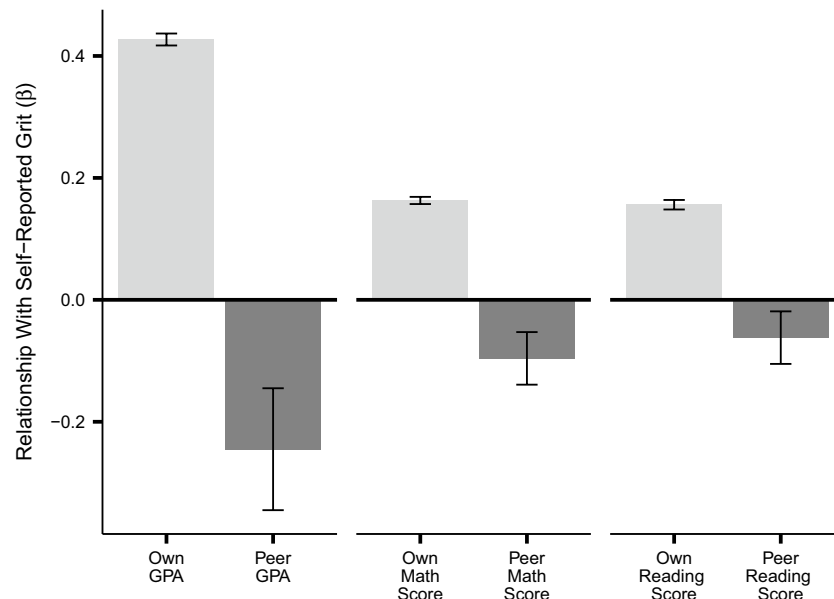


Figure 2. In Study 1, self-reported grit correlated positively with a student's own academic performance but inversely with the performance of their schoolmates. OLS models included demographic controls and school fixed effects. Error bars represent 95% confidence intervals. Model R^2 s for GPA, math score, and language score were 0.124, 0.071, and 0.071, respectively.

Grade point average (GPA). Students reported their overall, verbal, and math GPAs using a categorical scale which ranged from *less than 5.9* to *10* in half-point increments (i.e., < 5.9, 6.0–6.4, 6.5–6.9, etc.). We used the midpoint of the range in our analyses (i.e., 5.7, 6.2, 6.7, etc.). Although official GPAs were not available, meta-analytic estimates of the correlation between self-reported and objectively recorded GPA is $r = 0.82^{56}$. To avoid any issues with multicollinearity, we ran separate models for each GPA measure.

Standardized test scores. The Mexican Secretariat of Public Education provided standardized math and reading scores.

Analytic strategy. We used ordinary least squares (OLS) regression with clustered standard errors to predict self-reported grit from student's own and peer's academic performance:

$$G_{ist} = \alpha a_{ist} + \gamma b_{-ist} + \theta_s + \eta_t + \varepsilon_i$$

where G_{ist} is the self-reported grit for student i who was in 12th grade in school s at time t (2012 or 2013). Term a_{ist} is that student's own academic performance, operationalized as self-reported GPA, standardized math scores, or standardized verbal scores, respectively. Term b_{-ist} represents the average academic performance of students sharing a school with each student i , excluding student i . Term θ_s represents fixed effects for each student's school and captures ways in which schools might differ from each other—including such differences as teachers, curricula, school policies, and regional populations from which schools draw their members. Term η_t (fixed effect for year), captures how cohorts for each school systematically differ from each other. ε_i represents error.

Results. *Students surrounded by higher-performing classmates rate themselves lower in grit.* Consistent with prior research, among students in the same school, self-reported grit correlated positively with GPA ($\beta = 0.43$, $p < 0.001$), standardized math test scores ($\beta = 0.16$, $p < 0.001$), and standardized reading test scores ($\beta = 0.16$, $p < 0.001$). However, consistent with reference bias, self-reported grit correlated inversely with schoolmates' GPA ($\beta = -0.25$, $p < 0.001$), peer standardized math test scores ($\beta = -0.09$, $p < 0.001$), and peer standardized reading test scores ($\beta = -0.07$, $p = 0.004$). See Fig. 2 and Supporting Information for details.

Evidence for reference bias was consistent across demographic subgroups. Capitalizing on the size and representativeness of our sample, we explored moderators of reference bias. Regression coefficients for peer academic performance were not significantly different across subgroups defined by gender, mother's educational level, school type (public or private), or school size. See Tables S6 and S7 in Supporting Information for details.

Study 2: Replication and extension in a single large school district

In Study 2, we partnered with the nonprofit organization Character Lab to replicate and extend Study 1 with a sample of students in grades 8 through 12 in a large, diverse school district in the United States. This partnership enabled us to obtain official class schedules for each student, which we used to distinguish near- versus far-peers

as students who did or didn't share daily academic classes, respectively. Whereas GPA was self-reported in Study 1, in Study 2 we obtained GPA from official school records. As part of a larger survey administered by Character Lab, students completed a self-report questionnaire of conscientiousness (the tendency to be organized, responsible, and hardworking⁵⁷) as well as two questions we developed to directly assess self-regulation standards.

Methods. *Sample and procedure.* This study included data from $N = 21,818$ (50% female, $M_{age} = 15.60$, $SD_{age} = 1.54$) students attending $k = 62$ middle and high schools in a large public school district in the United States who completed surveys in either October 2019 or February 2020. This district was part of Character Lab Research Network (CLRN), a consortium of school partners committed to advancing scientific insights that help children thrive. According to school records, the race/ethnicity of our sample was: Hispanic/Latinx (41%), White (28%), Black (23%), and other (8%). About half (49%) of students were eligible for free and reduced-price meals.

Self-reported conscientiousness. Students completed 12 items from the Big Five Inventory-2⁵⁸ assessing conscientiousness (e.g., "I am someone who is persistent, works until the task is finished") using a 5-point Likert-type scale ranging from 1 = *Not like me at all* to 5 = *Totally like me*. The observed reliability was $\alpha = 0.83$.

Standards for hard work and preparedness. We included two questions to measure implicit standards for self-regulation. One question assessed norms for hard work: "If a student in your grade says they did 'a lot of homework' on a weeknight, how long would you guess they mean?" Eight response options ranged from 15 min (coded as 0.25 hours) to 3 or more hours (coded as 3 hours). The second question assessed norms for preparedness: "If a student in your grade says they 'sometimes' forget something they need for class, how often would you guess they mean?" Seven response options ranged from *once a month* to *three times or more per day* (coded as 66 times per month). We reverse-coded these values such that higher numbers indicated stricter standards for preparedness. These items were created for this study and used here for the first time.

Grade point average (GPA). From school administrative records, we calculated GPAs on a 100-point scale by averaging final grades in students' academic courses (English language arts, math, science, social studies) for the quarter in which students took the survey during the 2019–2020 school year.

Near-peer and far-peer GPAs. For each student, we designated near-peers as those students who took at least one academic course with the target student during the quarter in which they took the survey. We designated far-peers as students in the same school who did *not* share any academic courses. For the average student in our sample, 38% of schoolmates were near-peers and 62% were far-peers.

Analytic strategy. To examine whether self-regulation standards and conscientiousness related to students' own and peers' performance, we fit OLS regression models with standard errors clustered by school to estimate the following equation:

$$S_{is} = \alpha a_{is} + \gamma_1 b_{-is} + \gamma_2 c_{-is} + \delta x_{is} + \theta_s + \epsilon_i,$$

where S_{is} is a survey measure of conscientiousness or self-regulation standards for student i in school s , a_{is} is a student's own GPA, b_{-is} is the average GPA of students in the same school sharing at least one academic course with student i , c_{-is} is the average GPA of students in the same school but not sharing any academic courses with student i , x_{is} is a vector of student characteristics (age, gender, race/ethnicity, grade level, free or reduced-price meal status, English language learner status, special-education status, home language, and timing of the survey), θ_s represents school fixed effects, and ϵ_i is a random error term.

Results. *Reference bias replicates: students whose classmates perform better academically rate themselves as lower in conscientiousness.* As expected, this effect is driven by near-peers rather than far-peers. If implicit standards for self-regulation are determined by social comparison, reference bias should be driven by the individuals with whom individuals are in direct contact. As shown in Fig. 3, consistent with Study 1, self-reported conscientiousness was correlated positively with a student's own GPA ($\beta = 0.29$, $p < 0.001$), negatively with the GPA of near-peers ($\beta = -0.06$, $p < 0.001$), and not at all with the GPA of far-peers ($\beta = 0.01$, $p = 0.395$). See Table S10 in Supporting Information for details.

Students whose near-peers perform better academically hold higher self-regulation standards. As expected, standards for hard work were predicted by a student's own GPA ($\beta = 0.07$, $p < 0.001$) and the GPA of their near-peers ($\beta = 0.23$, $p < 0.001$), but not the GPA of their far-peers ($\beta = -0.03$, $p = 0.198$). The same pattern emerged for preparedness norms, which were predicted by students own GPA ($\beta = 0.05$, $p < 0.001$) and the GPA of their near-peers ($\beta = 0.14$, $p < 0.001$), but not far-peers ($\beta = -0.02$, $p = 0.080$). As in Study 1, the patterns of findings were generally similar across subgroups. See Tables S11–S16 in Supporting Information for details.

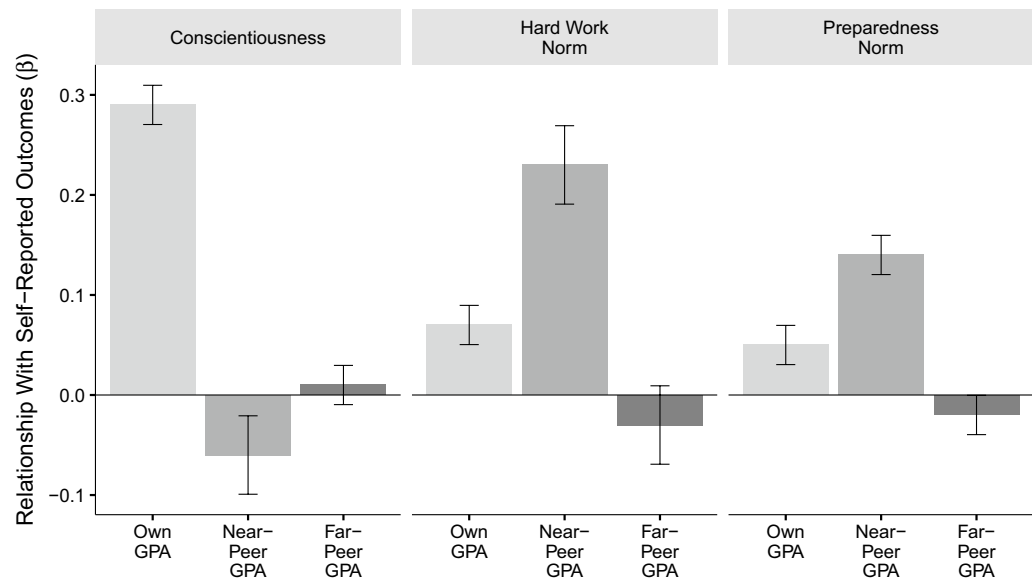


Figure 3. In Study 2, self-reported conscientiousness correlated positively with a student's own GPA and negatively with the GPA of near-peers. In contrast, standards for what constitutes hard work and preparedness correlated positively with both own and near-peer GPA. As expected, there was no effect of far-peer GPA. OLS models included demographic controls and school fixed effects. Error bars represent 95% confidence intervals. Model R^2 s for conscientiousness, hard work norms, and preparedness norms were 0.095, 0.159, and 0.059, respectively.

Study 3: In a longitudinal study of college graduation, evidence of reference bias in questionnaire but not in task measures of self-regulation

In Study 3, we sought evidence of discriminant validity. Unlike questionnaires, which require participants to make subjective judgments of their behavior, task measures assay behavior directly. In a prospective, longitudinal study of $N = 1278$ students attending $k = 15$ different college-preparatory charter schools in the United States, we tested the prediction that reference bias should be evident in questionnaire but not behavioral task measures of self-regulation. In their senior year of high school, students self-reported their grit and self-control (the ability to be in command of one's behavior and to inhibit one's impulses⁵⁷). In addition, they completed the Academic Diligence Task, a behavioral task in which students voluntarily allocate attention to either good-for-me-later math problems or fun-for-me-now games and videos. The Academic Diligence Task has previously been validated as indexing self-control and grit^{59,60}. Six years later, we used the National Student Clearinghouse database to identify students who successfully obtained their college diploma.

Methods. *Sample and procedure.* A few weeks before graduation, $N = 1278$ (55% female, $M_{age} = 18.01$, $SD_{age} = 1.01$) high school seniors responded to self-report questionnaires and task measures in school computer labs. Students attended $k = 15$ charter schools located in various urban centers in the United States. Between 76 and 98% of the students at each school participated in the study. Most students were socioeconomically disadvantaged (84% of students' mothers had less than a 4-year degree, 68% qualified for free or reduced-priced meals), and were mostly Latinx (46%) and African American (40%).

Self-reported grit. Students completed a 4-item version of the Grit Scale developed specifically for adolescents⁶¹. Students responded on a 5-point Likert-type scale ranging from 1 = *Not at all true* to 5 = *Completely true*. The observed reliability was $\alpha = 0.78$.

Self-control. Students completed four items from the Domain-Specific Impulsivity Scale^{59,62} assessing academic self-control (e.g., "I forgot something needed for school"). Students responded on a 5-point Likert-type scale ranging from *Not at all true* to *Completely true*. The observed reliability was $\alpha = 0.72$.

Academic Diligence Task (ADT). A subset ($n = 802$) of students in our sample completed the Academic Diligence Task, a behavioral assessment of self-regulation that has been validated in separate research⁵⁹. This computer-based task begins with screens explaining that practicing simple mathematical skills like subtraction can aid in further enhancing overall math abilities. Then, they completed three 3-min timed task blocks. In each, they chose between "Do math" and "Play game or watch movie." Clicking "Do math" displayed a math task involving single-digit subtraction with multiple-choice responses. On the other hand, clicking "Play game or watch movie" allowed students to play Tetris or watch entertaining videos. Students could freely switch between them during each block. See Supporting Information for details. The key metric from the ADT was the mean

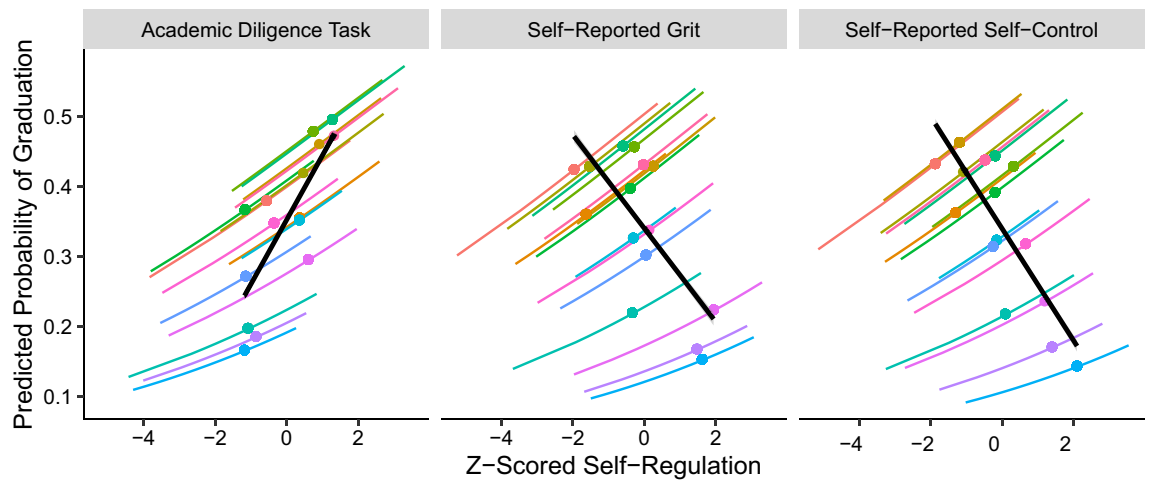


Figure 4. In Study 3, comparing students within schools (colored lines), higher self-regulation predicted higher odds of college graduation, whether measured by self-report questionnaires for grit and self-control or by a behavioral task called the Academic Diligence Task. When comparing schools to each other, however, higher self-reported grit and self-control scores predicted *lower* graduation rates, whereas the behavioral task *positively* predicted college graduation, as shown in the solid black lines. Plots show predicted probabilities of graduation from multilevel logistic regression models. AUCs for models predicting the academic diligence task, self-reported grit, and self-reported self-control were 0.694, 0.693, and 0.676, respectively.

number of problems correctly answered over the three blocks. Basic subtraction is very easy for most 12th grade students, so attentive engagement with the task resulted almost exclusively in correct answers: The median rate of correct responses was 98.3%. Due to positive skew and some clustering of data at 0 (i.e., students who did no math problems), we applied a square-root transformation to minimize bias from extremely high scores; this created an approximately normal distribution, which we used in subsequent calculations. Models using raw (i.e., non-transformed) ADT scores are shown in Table S19. Across the three blocks, the observed reliability was $\alpha = .78$.

General cognitive ability. During the online survey, students completed a brief (12-item) version of Raven's Progressive Matrices as an assessment of general cognitive ability⁶³. The ability variable was calculated as the sum of correctly answered questions out of 12, with any missing questions marked as incorrect. The observed reliability was $\alpha = 0.73$.

College graduation. We matched our data to the National Student Clearinghouse, a public database that includes enrollment and graduation data for over 97% of students in 2022^{64,65}. We coded six-year college graduation as 1 = *obtained degree within 6 years of enrollment* and 0 = *did not obtain degree within 6 years of enrollment*.

Analytic strategy. Because we were interested in both individual-level and school-level differences in self-regulation, we used multilevel modeling to analyze how the Academic Diligence Task and self-reported grit and self-control, predict college graduation. Specifically, we expected the ADT to positively predict college graduation at both the within- and between-school levels. We expected the relationship to be positive because prior research shows that students who obtain higher ADT scores tend to perform better academically⁵⁹. Moreover, we expected the relationship to be positive at *both levels* because, as a task measure, it does not involve comparative judgment and thus cannot be influenced by reference bias. In contrast, we expected self-reported grit and self-control to positively predict college graduation within a school but negatively between schools. We used a missing dummy variable coding approach to deal with missing data and included controls for general cognitive ability in our models.

Results. *Evidence of reference bias in longitudinal predictions of college graduation from self-reported, but not objectively measured, self-regulation.* As shown in Fig. 4, among seniors in the same high school, higher scores on self-report questionnaires of self-control ($b = 0.16$, $OR = 1.17$, $p = 0.022$) and grit ($b = 0.16$, $OR = 1.18$, $p = 0.020$) each predicted greater odds of earning a college diploma 6 years later. However, college graduation rates were actually lower for schools with higher self-reported self-control and grit scores ($b = -0.44$, $OR = 0.64$, $p = 0.001$; $b = -0.39$, $OR = 0.68$, $p = 0.005$, for self-control and grit, respectively).

This paradoxical pattern was not evident when self-regulation was assessed objectively using the Academic Diligence Task⁵⁹. Among seniors in the same school, college graduation was predicted by higher scores on the Academic Diligence Task ($b = 0.15$, $OR = 1.17$, $p = 0.031$). Likewise, when comparing across schools, college graduation rates were higher for schools whose students performed better on the Academic Diligence Task ($b = 0.46$, $OR = 1.58$, $p < 0.001$).

Taken as a whole these findings suggest that reference bias reversed the relationship between self-regulation and graduation across schools. See Supporting Information for summaries of multilevel logistic regression models, robustness checks, and a replication of the own versus peer performance models in Studies 1 and 2.

Discussion

The three studies in this investigation provide direct evidence for reference bias in self-reported self-regulation. In Study 1, high school seniors rated themselves lower in grit when their schoolmates earned higher GPAs and standardized achievement test scores. In Study 2, we replicated this effect using self-report questionnaires of conscientiousness and showed that it was driven by near-peers rather than by far-peers. Further, we showed that the GPA of near-peers (but not far-peers) correlates positively with self-regulation standards. Finally, in Study 3, we found that using self-report questionnaires of grit and self-control to predict college graduation 6 years later produced paradoxical results: Within a high school, students with higher self-reported self-regulation were more likely to graduate from college 6 years later, but across schools, average levels of self-regulation negatively predicted graduation. In contrast, an objective task measure of self-regulation—which indexed performance directly and did not ask students to judge themselves—positively predicted college graduation both within and across schools.

How big are reference bias effects? Studies 1 and 2 provide estimates in the range of $r = 0.06$ to 0.25 . All else being equal, a student in our samples whose peers' academic achievement is one standard deviation above the mean is predicted to rate their own self-regulation as 10–20% of a standard deviation lower. Assuming that higher standards for self-regulation depress self-report ratings while at the same time, via social norms and modeling, encourage more self-regulated behavior, these are actually lower-bound estimates. Consistent with this possibility, when we use a behavioral task to assess self-regulation, we observe results consistent with positive peer effects (Study 3), which have also been previously reported in the literature^{66–68}. Taken together, our findings suggest that reference bias effects, even across social groups in the same country, can be at least small-to-medium in size by contemporary benchmarks⁶⁹ and comparable to the effect sizes for faking on self-regulation questionnaires in workplace settings⁷⁰.

Several limitations of the current investigation suggest promising directions for future research.

First, we must be cautious about drawing strong causal inferences from the non-experimental data in our three field studies. In Study 1, variation in peer quality could have influenced self-reported self-regulation for reasons other than reference bias. Against this, we found direct evidence for near-peer influence on self-regulation standards provided in Study 2. However, in Study 2, there is the possibility of reverse-causality. For example, rather than near-peers determining self-regulation standards, it is possible that self-regulation standards determined patterns of enrollment (e.g., students with higher standards self-selecting into the same difficult classes). In Study 3, we cannot rule out the possibility that some unmeasured confound gave rise to contradictory within-school versus between-school results on self-report (but not objective task) measures of self-regulation. In sum, it is important to confirm our observational findings by experimentally manipulating peer groups and/or standards of self-regulation.

Second, there are limits to the external validity of our conclusions. In particular, we examined reference bias in adolescence, a developmental period in which sensitivity to peers is at its apogee⁷¹. The adolescents in our investigation lived in Mexico (Study 1) and the United States. (Studies 2 and 3). Further research on children and adults, in a wider sample of countries, and in contexts outside formal schooling, is needed to establish boundary conditions and moderators of reference bias. In general, effect sizes for reference bias are expected to be smaller when comparing social groups with more similar standards.

Third, we did not collect nuanced data on social networks (e.g., friendships, acquaintances). Indeed, our operationalization of peer groups was quite crude—students in the same grade and attending the same school in Study 1 and 3, and students in the same grade and school who share at least one academic class (i.e., near-peers) in Study 2. Given the increasing prevalence of social-network studies and the continued popularity of self-report questionnaires in behavioral science, it should be possible to identify the influence of prominent social referents and close friends on reference bias.

Finally, while we collected information about student's standards for self-regulation (in Study 2) and an objective measure of self-regulation (in Study 3), we have yet to collect both types of measures in the same sample. Doing so in a future study would enable us to test a mediation model in which peers influence standards for self-regulation which, in turn, diminish self-reported self-regulation relative to performance on a behavioral task of self-regulation. More generally, additional research is needed to establish the mediators, moderators, and boundary conditions of reference bias in the measurement of self-regulation.

Unfortunately, the problem of reference bias is not easily corrected. The most commonly suggested solution is anchoring vignettes⁷². This technique entails asking participants to rate detailed descriptions of hypothetical characters. These ratings are then used to adjust self-report questionnaire scores upward or downward depending on the stringency or leniency with which participants evaluated the hypothetical characters. Anchoring vignettes can increase the reliability and validity of self-reports⁷³ but do not always work as intended⁷⁴. They also increase the time, effort, and literacy required from survey respondents, which may limit their utility at scale^{73,75}.

A related possibility is to use behaviorally anchored⁷⁶ or act-frequency rating scales⁷⁷, which ask respondents to rate themselves on more specific, contextualized behaviors than is typical in traditional questionnaires. For example, while students at over-subscribed charter schools do not rate themselves as more self-regulated, they and their parents do report more “minutes of homework completed” in an open-ended question in the same questionnaire³⁸. In our view, such questions might mitigate response bias but probably do not eliminate it altogether. Why not? Because all subjective judgments rely, at least to some degree, on implicit standards that can differ (e.g., What level of effort is sufficient to consider yourself to be “doing homework?”).

As shown in Study 3, self-regulation can be assessed with behavioral tasks, which appear immune to reference bias. However, task measures have their own limitations, including a dramatically lower signal-to-noise ratio when compared to questionnaires and, relatedly, surprisingly modest associations with other measures of self-regulation^{46,78–81}.

Perhaps the best means of obviating reference bias is to take a multi-method, multi-informant approach to assessment, including trained observers who can rate behavior across multiple occasions¹². Observers who have seen hundreds, if not thousands, of cases typically have a wider reference frame than the individuals they are evaluating, which might explain why teacher ratings of behavior are more reliable and predictive of future outcomes than either parental reports or student self-reports⁸². The rarity of multi-method and multi-informant approaches suggests that, unfortunately, few researchers have the necessary resources or expertise to implement it, particularly at scale.

What are the implications of reference bias for researchers and policymakers?

Reference bias could suppress, or even reverse, the measured effects of interventions if the standards by which people judge their own behavior on pre- and post-questionnaires shift as a function of the intervention⁸³. In one study, participants were asked to rate their interviewing skills before training (*pre*). Afterward, participants rated themselves again (*post*) and, in addition, retrospectively estimated what their skills had been at baseline (*then*). Even though questionnaire items were identical for all assessments, *then* ratings were lower than *pre* ratings—suggesting that participants adopted higher standards as a result of the intervention. Moreover, third-party judges' ratings of performance matched *then-post* change better than *pre-post* differences⁸⁴.

The implications of reference bias extend beyond intervention research. Consider, for example, mean-level increases in conscientiousness from adolescence through midlife^{85–87}. If adults in their 50s hold higher standards for what it means to be courteous, rule-abiding, and self-controlled than teenagers, then age differences in conscientiousness may be even larger than we now think. In fact, to the extent that implicit standards and actual behavior are inversely correlated, reference bias should be expected to attenuate associations of self-regulation with groups of any kind.

While the importance of personal qualities like self-regulation is incontrovertible, the specter of reference bias argues against relying on self-report questionnaires when comparing students attending different schools, citizens who live in different countries, or indeed any of the members of any social group whose standards could differ from one another. Are you a hard worker? Responding to such a question requires looking *within* to identify the patterns of our behavior. In addition, the evidence for reference bias presented here suggests that knowingly or not, we also look *around* when we decide how to respond.

Ethics statement. All methods were carried out in accordance with relevant guidelines and regulations. Participants in Studies 2 and 3 completed written informed consent prior to participation in this study. Participants in Study 1 were completing country-mandated educational assessments, and thus did not complete written informed consent. We accessed this secondary dataset with authorization from the Mexican Secretariat of Education. Study 1 was approved by the Mexican Secretariat of Education. Study 2 was approved by Advarra IRB. Study 3 was approved by Stanford University IRB.

Data availability

The data that support the findings of Study 1 are available from the Mexican Ministry of Education but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Mexican Ministry of Education. Data for Study 2 and Study 3 are included in this published article's Supporting Information.

Received: 17 May 2022; Accepted: 31 October 2022

Published online: 10 November 2022

References

- Rothbart, M. K. Temperament, development, and personality. *Curr. Dir. Psychol. Sci.* **16**, 207–212. <https://doi.org/10.1111/j.1467-8721.2007.00505.x> (2007).
- Mischel, W., Shoda, Y. & Rodriguez, M. L. Delay of gratification in children. *Science* **244**, 933–938. <https://doi.org/10.1126/science.2658056> (1989).
- Freud, S. *Beyond the Pleasure Principle* 90 (The International Psycho-Analytical Press, 1922).
- Roberts, B. W. & Yoon, H. J. Personality psychology. *Annu. Rev. Psychol.* **73**, 489–516 (2022).
- Nigg, J. T. Annual research review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *J. Child Psychol. Psychiatry* **58**, 361–383. <https://doi.org/10.1111/jcpp.12675> (2017).
- Berns, G. S., Laibson, D. & Loewenstein, G. Intertemporal choice—toward an integrative framework. *Trends Cogn. Sci.* **11**, 482–488. <https://doi.org/10.1016/j.tics.2007.08.011> (2007).
- Heckman, J. J. & Kautz, T. Hard evidence on soft skills. *Labour Econ.* **19**, 451–464. <https://doi.org/10.1016/j.labeco.2012.05.014> (2012).
- Duckworth, A. L., Taxer, J. L., Eskreis-Winkler, L., Galla, B. M. & Gross, J. J. Self-control and academic achievement. *Annu. Rev. Psychol.* **70**, 373–399. <https://doi.org/10.1146/annurev-psych-010418-103230> (2019).
- Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C. & Domitrovich, C. E. Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Dev. Psychopathol.* **20**, 821–843. <https://doi.org/10.1017/S0954579408000394> (2008).
- Vedel, A. The Big Five and tertiary academic performance: A systematic review and meta-analysis. *Personal. Individ. Differ.* **71**, 66–76. <https://doi.org/10.1016/j.paid.2014.07.011> (2014).

11. Daly, M., Egan, M., Quigley, J., Delaney, L. & Baumeister, R. F. Childhood self-control predicts smoking throughout life: Evidence from 21,000 cohort study participants. *Health Psychol.* **35**, 1254–1263. <https://doi.org/10.1037/hea0000393> (2016).
12. Moffitt, T. E. *et al.* A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl. Acad. Sci.* **108**, 2693–2698. <https://doi.org/10.1073/pnas.1010076108> (2011).
13. Bogg, T. & Roberts, B. W. Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychol. Bull.* **130**, 887–919. <https://doi.org/10.1037/0033-2909.130.6.887> (2004).
14. Hofmann, W., Luhmann, M., Fisher, R. R., Vohs, K. D. & Baumeister, R. F. Yes, but are they happy? Effects of trait self-control on affective well-being and life satisfaction: Trait self-control and well-being. *J. Pers.* **82**, 265–277. <https://doi.org/10.1111/jopy.12050> (2014).
15. Hirschi, T. Self-control and crime. In *Handbook of Self-Regulation*, 537–552.
16. Barrick, M. R. & Mount, M. K. The big five personality dimensions and job performance: A meta-analysis. *Pers. Psychol.* **44**, 1–26. <https://doi.org/10.1111/j.1744-6570.1991.tb00688.x> (1991).
17. Duckworth, A. L., Weir, D., Tsukayama, E. & Kwok, D. Who does well in life? Conscientious adults excel in both objective and subjective success. *Front. Psychol.* <https://doi.org/10.3389/fpsyg.2012.00356> (2012).
18. Wiersma, U. J. & Kappe, R. Selecting for extroversion but rewarding for conscientiousness. *Eur. J. Work Organ. Psychol.* **26**, 314–323. <https://doi.org/10.1080/1359432X.2016.1266340> (2017).
19. Denissen, J. J. A. *et al.* Uncovering the power of personality to shape income. *Psychol. Sci.* **29**, 3–13 (2018).
20. Doebel, S. Rethinking executive function and its development. *Perspect. Psychol. Sci.* **15**, 942–956. <https://doi.org/10.1177/1745691620904771> (2020).
21. Casey, B. J. Beyond simple models of self-control to circuit-based accounts of adolescent behavior. *Annu. Rev. Psychol.* **66**, 295–319. <https://doi.org/10.1146/annurev-psych-010814-015156> (2015).
22. Dahl, R. E., Allen, N. B., Wilbrecht, L. & Suleiman, A. B. Importance of investing in adolescence from a developmental science perspective. *Nature* **554**, 441–450. <https://doi.org/10.1038/nature25770> (2018).
23. Steinberg, L. Cognitive and affective development in adolescence. *Trends Cogn. Sci.* **9**, 69–74. <https://doi.org/10.1016/j.tics.2004.12.005> (2005).
24. Bailey, R., Meland, E. A., Brion-Meisels, G. & Jones, S. M. Getting developmental science back into schools: Can what we know about self-regulation help change how we think about “No Excuses”? *Front. Psychol.* **10**, 1885. <https://doi.org/10.3389/fpsyg.2019.01885> (2019).
25. Hamilton, S. F. Chapter 6: The secondary school in the ecology of adolescent development. *Rev. Res. Educ.* **11**, 227–258. <https://doi.org/10.3102/0091732X011001227> (1984).
26. Leonard, J. A., Lee, Y. & Schulz, L. E. Infants make more attempts to achieve a goal when they see adults persist. *Science* **357**, 1290–1294. <https://doi.org/10.1126/science.aan2317> (2017).
27. Bandura, A. & Mischel, W. Modification of Self-Imposed delay of reward through exposure to live and symbolic models. *J. Pers. Soc. Psychol.* **2**, 698–705 (1965).
28. King, K. M., McLaughlin, K. A., Silk, J. & Monahan, K. C. Peer effects on self-regulation in adolescence depend on the nature and quality of the peer interaction. *Dev. Psychopathol.* **30**, 1389–1401. <https://doi.org/10.1017/S0954579417001560> (2018).
29. Doebel, S. & Munakata, Y. Group influences on engaging self-control: Children delay gratification and value it more when their in-group delays and their out-group doesn't. *Psychol. Sci.* **29**, 738–748. <https://doi.org/10.1177/0956797617747367> (2018).
30. Bertling, J. P., Marksteiner, T. & Kyllonen, P. C. General noncognitive outcomes. In *Assessing Contexts of Learning* (eds Kuger, S. *et al.*) 255–281 (Springer, 2016). https://doi.org/10.1007/978-3-319-45357-6_10.
31. U.S. Department of Education. Every Student Succeeds Act (ESSA) (2015).
32. Center for Disease Control and Prevention. Whole School, Whole Community, Whole Child (WSCC) (2021).
33. OECD. *Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills* (OECD, 2021).
34. Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A. & Kiguel, S. School effects on socioemotional development, school-based arrests, and educational attainment. *Am. Econ. Rev. Insights* **2**, 491–508 (2020).
35. West, M. R. *et al.* Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educ. Eval. Policy Anal.* **38**, 148–170. <https://doi.org/10.3102/0162373715597298> (2016).
36. Dobbie, W. & Fryer, R. G. The medium-term impacts of high-achieving charter schools. *J. Polit. Econ.* **123**, 985–1037. <https://doi.org/10.1086/682718> (2015).
37. Tuttle, C. C. *et al.* Understanding the Effect of KIPP as it Scales, Volume I, Impacts on Achievement and Other Outcomes. Tech. Rep, Mathematica Policy Research (2015).
38. Tuttle, C. C. *et al.* KIPP Middle Schools: Impacts on Achievement and Other Outcomes. Tech. Rep, Mathematica Policy Research (2013).
39. Angrist, J. D., Pathak, P. A. & Walters, C. R. Explaining charter school effectiveness. *Am. Econ. J. Appl. Econ.* **5**, 1–27. <https://doi.org/10.1257/app.5.4.1> (2013).
40. Dobbie, W. & Fryer, R. G. Getting beneath the veil of effective schools: Evidence from New York City. *Am. Econ. J. Appl. Econ.* **5**, 28–60. <https://doi.org/10.1257/app.5.4.28> (2013).
41. Heine, S. J., Lehman, D. R., Peng, K. & Greenholtz, J. What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *J. Pers. Soc. Psychol.* **82**, 903–918. <https://doi.org/10.1037/0022-3514.82.6.903> (2002).
42. Van de Gaer, E., Grisy, A., Schulz, W. & Gebhardt, E. The reference group effect. *Cult. Psychol.* **43**, 24 (2012).
43. Van Vaerenbergh, Y. & Thomas, T. D. Response styles in survey research: A literature review of antecedents, consequences, and remedies. *Int. J. Public Opin. Res.* **25**, 195–217. <https://doi.org/10.1093/ijpor/eds021> (2013).
44. Schwarz, N. & Oyserman, D. Asking questions about behavior: Cognition, communication, and questionnaire construction. *Am. J. Eval.* **22**, 127–160 (2001).
45. Tourangeau, R., Rips, L. J. & Rasinski, K. *The Psychology of Survey Response* (Cambridge University Press, 2000).
46. Duckworth, A. L. & Yeager, D. S. Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educ. Res.* **44**, 237–251. <https://doi.org/10.3102/0013189X15584327> (2015).
47. Morina, N. Comparisons Inform Me Who I Am: A general comparative-processing model of self-perception. *Perspect. Psychol. Sci.* **16**, 1281–1299 (2021).
48. Marsh, H. W. & Craven, R. G. Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspect. Psychol. Sci.* **1**, 133–163. <https://doi.org/10.1111/j.1745-6916.2006.00010.x> (2006).
49. Marsh, H. W. *et al.* The Big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educ. Psychol. Rev.* **20**, 319–350. <https://doi.org/10.1007/s10648-008-9075-6> (2008).
50. Marsh, H. W. The Big-Fish-Little-Pond effect on academic self-concept. *J. Educ. Psychol.* **79**, 280–295 (1987).
51. Gerber, J. P., Wheeler, L. & Suls, J. A social comparison theory meta-analysis 60+ years on. *Psychol. Bull.* **144**, 177–197. <https://doi.org/10.1037/bul0000127> (2018).
52. Bybee, R. & McCrae, B. Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *Int. J. Sci. Educ.* **33**, 7–26. <https://doi.org/10.1080/09500693.2010.518644> (2011).

53. Schmitt, D. P., Allik, J., McCrae, R. R. & Benet-Martínez, V. The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *J. Cross-Cultural Psychol.* **38**, 173–212. <https://doi.org/10.1177/0022022106297299> (2007).
54. Duckworth, A. L., Peterson, C., Matthews, M. D. & Kelly, D. R. Grit: Perseverance and passion for long-term goals. *J. Pers. Soc. Psychol.* **92**, 1087–1101. <https://doi.org/10.1037/0022-3514.92.6.1087> (2007).
55. Duckworth, A. L. & Quinn, P. D. Development and validation of the Short Grit scale (Grit-S). *J. Pers. Assess.* **91**, 166–174. <https://doi.org/10.1080/00223890802634290> (2009).
56. Kuncel, N. R., Credé, M. & Thomas, L. L. The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Rev. Educ. Res.* **75**, 63–82. <https://doi.org/10.3102/00346543075001063> (2005).
57. American Psychological Association. *APA Dictionary of Psychology* 1st edn. (American Psychological Association, 2007).
58. Soto, C. J. & John, O. P. Short and extra-short forms of the Big Five Inventory-2: The BFI-2-S and BFI-2-XS. *J. Res. Pers.* **68**, 69–81. <https://doi.org/10.1016/j.jrp.2017.02.004> (2017).
59. Galla, B. M. *et al.* The Academic Diligence Task (ADT): Assessing individual differences in effort on tedious but important school-work. *Contemp. Educ. Psychol.* **39**, 314–325. <https://doi.org/10.1016/j.cedpsych.2014.08.001> (2014).
60. Zammaro, G., Nichols, M., Duckworth, A. & D'Mello, S. Further Validation of Survey Effort Measures of Relevant Character Skills: Results from a Sample of High School Students. EDRE Working Paper2018-07. <https://doi.org/10.2139/ssrn.3265332> (2018).
61. Galla, B. M. *et al.* Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *Am. Educ. Res. J.* **56**, 2077–2115. <https://doi.org/10.3102/0002831219843292> (2019).
62. Tsukayama, E., Duckworth, A. L. & Kim, B. Domain-specific impulsivity in school-age children. *Dev. Sci.* **16**, 879–893. <https://doi.org/10.1111/desc.12067> (2013).
63. Raven, J. & Raven, J. Raven progressive matrices. In *Handbook of Nonverbal Assessment* (ed. McCallum, R. S.) 223–237 (Kluwer Academic, 2003).
64. Dynarski, S. M., Hemelt, S. W. & Hyman, J. M. The missing manual: Using National Student Clearinghouse data to track postsecondary outcomes. *Educ. Eval. Policy Anal.* **37**, 53S–79S (2015).
65. Schoenecker, C. & Reeves, R. The National Student Clearinghouse: The largest current student tracking database. *New Directions Community Coll.* **143**, 47–57. <https://doi.org/10.1002/cc.335> (2008).
66. Cialdini, R. B. Descriptive social norms as underappreciated sources of social control. *Psychometrika* **72**, 263–268. <https://doi.org/10.1007/s11336-006-1560-6> (2007).
67. Bandura, A. *Social Learning Theory* (General Learning Press, 1971).
68. Sacerdote, B. Peer effects in education: How might they work, how big are they and how much do we know thus far?. *Handb. Econ. Educ.* **3**, 249–277. <https://doi.org/10.1016/B978-0-444-53429-3.00004-1> (2011).
69. Funder, D. C. & Ozer, D. J. Evaluating effect size in psychological research: Sense and nonsense. *Adv. Methods Pract. Psychol. Sci.* **2**, 156–168. <https://doi.org/10.1177/2515245919847202> (2019).
70. Martínez, A. & Salgado, J. F. A meta-analysis of the faking resistance of forced-choice personality inventories. *Front. Psychol.* **12**, 732241. <https://doi.org/10.3389/fpsyg.2021.732241> (2021).
71. Steinberg, L. & Amanda, S. M. Adolescent development. *Annu. Rev. Psychol.* **52**, 83–110 (2000).
72. King, G., Murray, C. J. L., Salomon, J. A. & Tandon, A. Enhancing the validity and cross-cultural comparability of measurement in survey research. *Am. Polit. Sci. Rev.* **98**, 191–207. <https://doi.org/10.1017/S000305540400108X> (2004).
73. Primi, R., Zanon, C., Santos, D., De Fruyt, F. & John, O. P. Anchoring vignettes: Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid?. *Eur. J. Psychol. Assess.* **32**, 39–51. <https://doi.org/10.1027/1015-5759/a000336> (2016).
74. Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M. & Ispány, M. Promises and pitfalls of anchoring vignettes in health survey research. *Demography* **52**, 1703–1728. <https://doi.org/10.1007/s13524-015-0422-1> (2015).
75. Bertling, J. P., Borgonovi, F. & Almonte, D. E. Psychosocial skills in large-scale assessments: Trends, challenges, and policy implications. In *Psychosocial Skills and School Systems in the 21st Century: Theory, Research, and Practice. The Springer Series on Human Exceptionality* (eds Lipnevich, A. A. *et al.*) (Springer, 2016). <https://doi.org/10.1007/978-3-319-28606-8>.
76. Schwab, D. P., Heneman, H. G. & DeCotiis, T. A. Behaviorally anchored rating scales: A review of the literature. *Pers. Psychol.* **28**, 549–562. <https://doi.org/10.1111/j.1744-6570.1975.tb01392.x> (1975).
77. Buss, D. M. & Craik, K. H. The act frequency approach to personality. *Psychol. Rev.* **90**, 105–126 (1983).
78. Enkavi, A. Z. *et al.* Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proc. Natl. Acad. Sci.* **116**, 5472–5477. <https://doi.org/10.1073/pnas.1818430116> (2019).
79. Duckworth, A. L. & Kern, M. L. A meta-analysis of the convergent validity of self-control measures. *J. Res. Pers.* **45**, 259–268. <https://doi.org/10.1016/j.jrp.2011.02.004> (2011).
80. Sharma, L., Kohl, K., Morgan, T. A. & Clark, L. A. “Impulsivity”: Relations between self-report and behavior. *J. Pers. Soc. Psychol.* **104**, 559–575. <https://doi.org/10.1037/a0031181> (2013).
81. Friedman, N. P. & Gustavson, D. E. Do rating and task measures of control abilities assess the same thing?. *Curr. Dir. Psychol. Sci.* **31**, 262–271. <https://doi.org/10.1177/09637214221091824> (2022).
82. Feng, S., Han, Y., Heckman, J. J. & Kautz, T. Comparing the reliability and predictive power of child, teacher, and guardian reports of noncognitive skills. *Proc. Natl. Acad. Sci.* **119**, e2113992119. <https://doi.org/10.1073/pnas.2113992119> (2022).
83. Howard, G. S. Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Eval. Rev.* **4**, 93–106 (1980).
84. Howard, G. S. & Dailey, P. R. Response-Shift Bias: A source of contamination of self-report measures. *J. Appl. Psychol.* **64**, 144–150 (1979).
85. Roberts, B. W., Walton, K. E. & Viechtbauer, W. Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychol. Bull.* **132**, 1–25. <https://doi.org/10.1037/0033-2909.132.1.1> (2006).
86. Damian, R. I. & Spengler, M. Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *J. Pers. Soc. Psychol.* **117**, 674–695 (2018).
87. Roberts, B. W. & DelVecchio, W. F. The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychol. Bull.* **126**, 3–25. <https://doi.org/10.1037/0033-2909.126.1.3> (2000).

Acknowledgements

This research received support from the Bill & Melinda Gates Foundation, the Raikes Foundation, the William T. Grant Foundation, and a fellowship from the Center for Advanced Study in the Behavioral Sciences (CASBS) to the sixth author and grants from the John Templeton Foundation, the Walton Family Foundation, and National Science Foundation to the last author. This research was supported by the National Institute of Child Health and Human Development (Grant No. 10.13039/1000000071 R01HD084772-01). The authors wish to thank Donald Kametz, Laura Keane, and the schools and students who participated in the research.

Author contributions

A.L.D., D.S.Y., J.M.O., and P.A.P. conceptualized the study and developed the methodology; B.M.G. developed methodology; A.L.D., D.S.Y., and J.M.O. collected data; B.L., J.M.O., P.A.P., A.D., T.K., K.M., A.L.D., and D.S.Y. analyzed and interpreted data; B.L., A.L.D., D.S.Y., J.M.O., P.A.P., A.D., T.K., and K.M. wrote the paper; all authors revised and approved the final draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-23373-9>.

Correspondence and requests for materials should be addressed to B.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022