What I See My Role Models Do: Elucidating the Mechanisms of Reference Bias

Benjamin Lira
University of Pennsylvania

Reference bias arises when individuals rely on different implicit reference points (i.e., standards) to answer the same questionnaire item. To date, little is known about the mechanisms and boundary conditions of reference bias. Here, I examine evidence of the relative influence of role models versus friends for reference bias in a wider variety of traits than have been examined in prior research. At four time points over 2 academic years, N = 1,071 adolescents completed self-report questionnaires on two intrapersonal traits (i.e., academic self-control and grit), and two interpersonal traits (i.e., gratitude and interpersonal self-control). Separately, their teachers rated them on these same traits. At each time point, students also nominated peers who exemplified each trait and, in addition, named their close friends. I operationalized reference bias as the effect of peers' GPA on students' self-reported personality when controlling for their own GPA. I found evidence of reference bias for grit and academic self-control but not for gratitude or interpersonal self-control. Compared with these interpersonal traits, grit and academic self-control were more related to GPA and more reliably observed by teachers. Evidence of reference bias was stronger for peer-nominated role models than friends, and between- rather than within-people.

Keywords: reference bias, social networks, personality traits

> Comparison is the thief of joy.
>
> *– Attributed to Theodore Roosevelt*

Self-report questionnaires are by far the most common assessment tool in psychological science, particularly in the field of personality (Paulhus & Vazire, 2007). Why? Self-report questionnaires are extremely reliable–an order of magnitude more so than performance tasks (Enkavi et al., 2019). Relatedly, self-report questionnaires are remarkably predictive of positive future outcomes, such as earnings and health (e.g., Bogg & Roberts, 2004; Denissen et al., 2018; Duckworth et al., 2012; Heckman & Kautz, 2012; Lundberg, 2019). Finally, questionnaires are also far cheaper than most other assessment techniques, such as observations or performance tasks.

Despite these advantages, questionnaires have limitations. For instance, questionnaires are fakeable, rendering them less useful in high-stakes situations (Krosnick, 1999; Sackett, 2011). However, even when respondents are doing their best to be accurate, their responses can still be biased. In this investigation, I analyzed the mechanisms for reference bias, defined as systematic error that arises when people rely on different implicit standards (i.e., reference points) to answer the same questionnaire items. (Duckworth & Yeager, 2015; Heine et al., 2008; Heine et al., 2002; West et al., 2016). Specifically, I capitalized on a large sample of adolescents to examine evidence for (1) evidence for reference bias across different kinds of personality measures, (2) the relative contribution of friends and role models to reference bias, and (3) the relative size of reference bias between respondents versus within respondents.

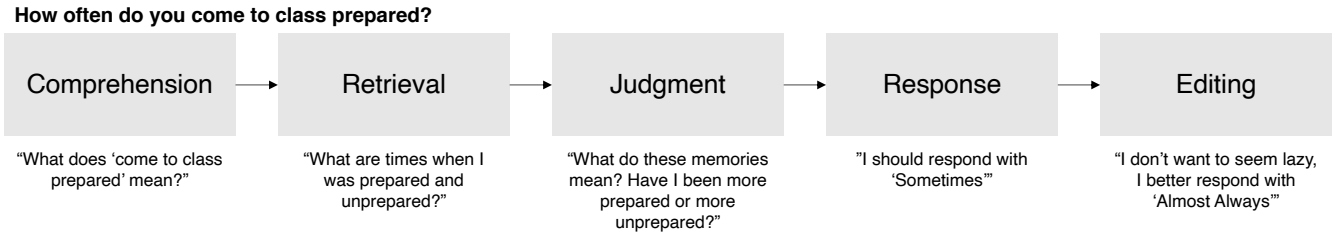
Reference Bias in Questionnaires

Questionnaires are extremely reliable (Enkavi et al., 2019). On its face, that statement may seem false. After all, when responding to any single item, participants are imprecise: They might provide a response to an item that is either too high or too low compared with reality. But reliability refers to the absence of *unsystematic* error (i.e., noise; Kahneman et al., 2021). If responses are purely noisy, participants are equally likely to overestimate and underestimate their true levels of the trait for each item. Therefore, when a plurality of items are averaged together, the errors cancel out, and the average of the imprecisely answered items will converge on the true score, provided that measurement is only noisy and not biased. This concept is known as the *principle*

Benjamin Lira
Correspondence concerning this article should be addressed to Benjamin Lira, Psychology Department, University of Pennsylvania, 3675 Market Street, Philadelphia, Pennsylvania, 19146. E-mail: blira@upenn.edu

Figure 1

Questionnaire Response as a Multi-Stage Cognitive Process



Note. Adapted from Duckworth and Yeager (2015), Schwarz and Oyserman (2001), and Tourangeau et al. (2000)

of aggregation (Rushton et al., 1983). Questionnaires are remarkably reliable because every questionnaire score is composed of the aggregation of multiple items, each of which aggregates multiple lived experiences. Task measures—in which participants must display a given behavior at request for measurement to happen—are, relatively speaking, a snapshot of a person's performance in a single situation at a single point in time, resulting in much more random variation around a person's true score (i.e., higher noise, lower reliability; Enkavi et al., 2019).

In contrast, bias refers to *systematic* error: the difference between true scores and what is reported is correlated with some external factor. For example, when answering a questionnaire for a job application, participants who are more motivated to get the job might be more inclined to overstate their conscientiousness. This means that bias is *directional* rather than random, leading to self-reports that are consistently larger or smaller than the true score. By extension, averaging a collection of biased items will still result in a biased estimate: while noise cancels out when there are enough items, bias remains (Kahneman et al., 2016).

As shown in **Figure 1,** the dominant models in survey methodology suggest that answering questionnaires is a multi-stage cognitive process (Duckworth & Yeager, 2015; Schwarz & Oyserman, 2001; Tourangeau et al., 2000). First, participants need to comprehend what the item is asking. This involves semantic (i.e., what do the words mean?), syntactic (i.e., how should I parse the sentence?), and pragmatic (i.e., what does the administrator want me to say?) processes (Tourangeau, 2018). Second, after reading and interpreting the question, respondents must retrieve the relevant information from memory. Third, respondents must collect these retrieved memories and integrate them into a summary judgment. Fourth, the resulting judgment must be translated into one of the response options (e.g., "always" or "almost always"). Finally, the response can be edited if there are incentives to do so (i.e., faking; Sackett, 2011).

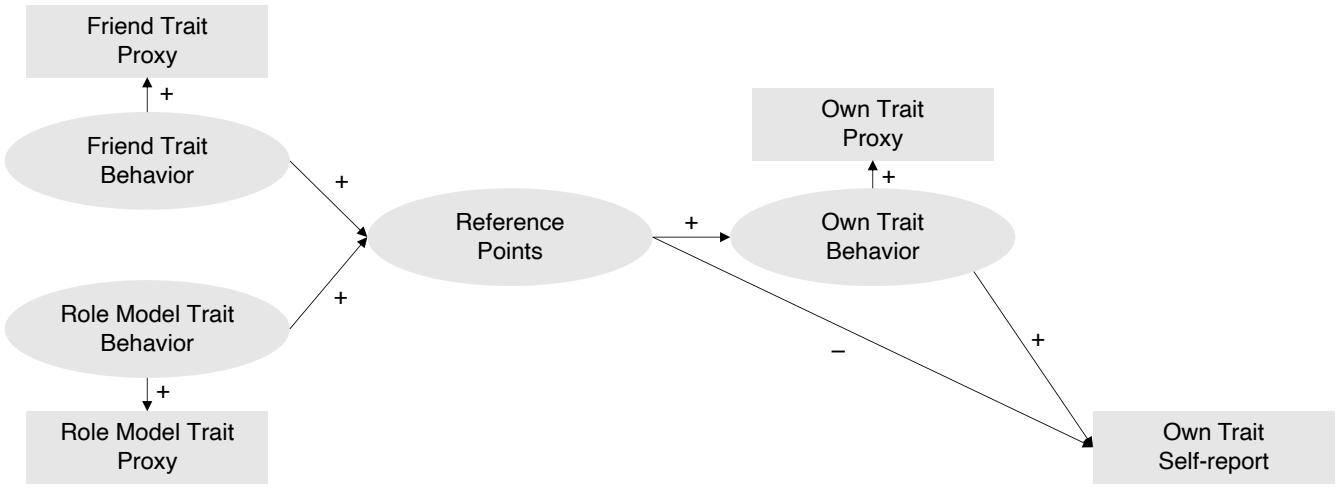
Often, participants use different implicit reference points to judge their behavior when responding to questionnaire items—a systematic error known as reference bias. Take John and Mary, for example, who both see the same questionnaire item asking them whether or not they usually come to class prepared. To answer such a question, they must remember when they came prepared or unprepared to class (see **Figure 2**; arrow leading from own trait behavior to own trait self-report), and integrate them into a summary judgment by comparing that against their reference points for preparedness (arrow leading from reference points to own trait self-report). John has lower reference points for preparedness, and therefore thinks that bringing a notebook and a pen counts as being prepared. Mary, on the other hand, has higher reference points for preparedness, and thinks that she is *unprepared* unless she has completed homework, studied the assigned readings, and reviewed last class's material. The higher the reference point, the lower the individual will report preparedness (indicated by the negative sign). Thus, even if John and Mary usually *behave* in the same way, they will judge that same behavior differently, resulting in divergent responses. These diverse reference points might be influenced by the characteristics of the peers surrounding John and Mary. Likely, Mary thinks about preparedness this way because she is surrounded by other people who tend to prepare more than the people John is surrounded by (arrows leading from role model trait behavior and friend trait behavior to reference points), as would be suggested by research on peer descriptive norms (Cialdini, 2007).

In my view, reference bias permeates the cognitive process of questionnaire answering (See **Figure 1**), but it mostly affects how a judgment is formed. Different reference points may lead to different interpretations of the question (e.g., what does "hard work" mean). The reference points used for judgment may bias memory processes so that certain contents are more likely to be recovered than others. However, even when two people interpret the same question and recall the same information, they may come to different conclusions if they use different reference points to judge themselves. Since reference bias can be introduced relatively early in the process, it can be problematic even when respondents have no incentives to edit their responses.

Reference bias may account for otherwise puzzling find-

Figure 2

Theoretical Model for Reference Bias



Note. Boxes indicate observed variables. Ellipses indicate unobserved latent variables.

ings in cross-cultural psychology. While cultural experts agree that people in Asian countries are more conscientious than people in the United States (Heine et al., 2002; Peng et al., 1997), Asian citizens consistently rate themselves the least conscientious in cross-national comparisons (Mõttus et al., 2012). Could this conflicting evidence be explained by measurement artifacts? Heine et al. (2008) showed that objective country-level proxies for conscientiousness, such as the efficiency of post offices and walking speed, are inversely correlated with country-level averages of self-reported conscientiousness. Similarly, conscientiousness correlates inversely with country-level gross domestic product (GDP), and life expectancy (Oishi & Roth, 2009). If these objective markers index conscientiousness, this evidence supports the thesis that questionnaires are biased because national cultures correlate with different reference points for judging conscientiousness.

Recently, reference bias has been demonstrated across cultures within a country, specifically, in the context of evaluating the effects of charter schools on student self-regulation and educational outcomes. Some evaluations of charter schools suggest that they improve self-report measures of hard work and future college persistence (Jackson et al., 2020). However, lottery evaluations show no improvements in self-reports of self-regulation, despite showing *increases* in positive outcomes such as report card grades, standardized test scores, attendance rates, and college enrollment levels, as well as lower levels of incarceration and unplanned pregnancies (Dobbie & Fryer, 2015; Tuttle et al., 2013; Tuttle et al., 2015; West et al., 2016). There are two potential explanations for this data: Either charter schools are changing behavior without changing personality or the effects of charter schools focused on character development include raising the reference points by which students evaluate their character, thus obscuring real changes in students' personality.

The nascent literature on reference bias requires unproven albeit reasonable assumptions to be true. Studies that suggest reference bias from group-level comparisons (e.g., graduation rates correlate inversely to school-level self-reported self-regulation) could be confounded at the school level (Dobbie & Fryer, 2015; Tuttle et al., 2013; Tuttle et al., 2015; West, 2016; West et al., 2016). Cross-cultural psychology has relied on experts' opinions, which might be biased (Heine et al., 2002; Peng et al., 1997), or on country-level far-proxies for behavior, which might be noisy and lack validity (Heine et al., 2008). Very recently, Lira et al. (2022) have shown that reference bias still occurs within schools, obviating these group-level concerns. However, little is known about mechanisms: Where do the reference points explaining reference bias come from?

Where Do Our Reference Points Come From? Friends or Role Models

Differing reference points leading to reference bias likely depend on social comparisons and social norms. Peers influence our behavior: Depending on who surrounds us, different social norms and social modeling influences might alter how we behave (Bandura, 1971; Cialdini, 2007; Sacerdote, 2011). Peers also influence how we perceive ourselves independent of influences on behavior (Marsh, 1987; Morina, 2021). As shown in **Figure 2,** the effect of peers on self-perception likely plays a key role in reference bias: The peers we compare ourselves to shape our reference points, and these reference points, in turn, shape our questionnaire

responses. Aside from the fact that peer composition determines reference points for comparison, little is known about what kinds of peers should be more related to reference bias. Past research has operationalized reference bias in terms of differences across countries (e.g., Heine et al., 2008) or schools (e.g., West et al., 2016); or in terms of the influence that broad peer groups (i.e., schoolmates, students sharing core classes) have on self-reports (Lira et al., 2022). To date, there is no evidence on the influence that more specifically defined peers, such as friends or role models, might have in the questionnaire response process.

Friends are an obvious potential mechanism for reference bias, especially during adolescence. Developmentally, adolescence is a period of heightened sensitivity to friends and other same-aged peers (Casey, 2015; Dahl et al., 2018; Steinberg, 2005). Moreover, adolescents spend increasing amounts of time with friends (Larson & Richards, 1991), suggesting that friends might be central in determining reference points. Because of this, friends' behavior might be more available and therefore more memorable (Schwarz et al., 1991).

On the other hand, role models might also be important for reference bias. Research on social norms (Cialdini, 2007) and social learning (Bandura, 1971) suggests that those whom we see as exemplars are likely to affect how we behave and influence what kind of behavior we consider to be appropriate. Because role models embody a given trait, their behavior might be more relevant to self-evaluations, thus having a stronger impact on how we respond to personality questionnaires.

Current Investigation

In this investigation, I examine the mechanism of reference bias. I leverage social network data on friendship and role models to identify the comparative relevance of friends and role models for reference bias. Also, I examine evidence for reference bias in a wider set of traits than have been studied before (i.e., self-regulation and conscientiousness). Finally, only recently has reference bias been established within schools, effectively controlling for potential confounds at the school level. However, it is unclear whether reference bias holds within people—that is, whether the same person will judge themselves differently when there are changes in their peer groups across time. I capitalized on a repeated-measures design to explore whether a particular student will rate themselves lower in positive personality traits in time points where they nominated higher-achieving peers compared to how they rated themselves when they nominated lower-achieving peers.

As shown in **Figure 2,** I operationalize reference bias as the effect of peer GPA on students' self-reported personality while controlling for their own GPA. If there is no reference bias in questionnaire responding, the academic performance of peers—be they role models or friends—should have no bearing on how a student self-reports their own personality traits. Adding controls for demographic characteristics and a student's own GPA removes variation in questionnaire responses that would be explained by differences in behavior. If we observe negative effects of peer GPA on questionnaire self-reports, this evidence would be consistent with reference bias. If, on the other hand, we observed positive effects of peer GPA on questionnaire self-reports, this evidence would be consistent with a self-enhancing effect, where students inflate their perceptions of their own personality when surrounded by higher-achieving peers (cfr. Cialdini et al., 1976).

Method

Participants and Procedure

This study included data from N = 1,071 (47.8% female, 50.9% male, 1.3% unreported; M_{age} = 15.6; SD_{age} = 13.7) students attending two public high schools in the United States who completed surveys from November 2014 to June 2016. According to school records, the race/ethnicity of our sample was: Black (49.0%), White (29.6%), Asian (14.0%), Hispanic/Latinx (4.3%), and other (3.2%). More than half (60.4%) of students were eligible for free and reduced-price meals.

These data were collected as part of a larger survey assessing character development during adolescence. Students completed virtual surveys in their school's computer laboratory and were supervised by their regular teachers. Waves of data collection were scheduled about a month before the end of each semester, and each one took around 45 minutes, scheduled during a single class period.

Measures

Self-Reported Personality Traits

Academic and Interpersonal Self-Control. Students completed the Domain-Specific Impulsivity Scale (Tsukayama et al., 2013), assessing academic self-control (e.g., "I came to class prepared.") and interpersonal self-control (e.g., "I stayed calm even when others bothered or criticized me.") using a 5-point Likert-type scale ranging from 1 = *Never* to 5 = *Always*. In each wave of data collection, the observed reliability ranged from α = .74 to .76 for academic self-control and α = .78 to .80 for interpersonal self-control

Grit. Students completed five items from the Short Grit Scale (Duckworth & Quinn, 2009), assessing their passion and perseverance for long-term goals (e.g., "I stayed committed to my goals, even if they took a long time to complete.") using a 5-point Likert-type scale ranging from 1 = *Never* to 5 = *Always*. In each wave of data collection, the observed reliability ranged from α = .72 to .77.

Gratitude. Students completed five items assessing the experience and expression of gratitude (Malin et al., 2017) (e.g., "I appreciated when other people helped me.") using a 5-point Likert-type scale ranging from 1 = *Never* to 5 = *Always*. In each wave of data collection, the observed reliability ranged from α = .68 to .72.

See alphas for each scale and time point as well as for the item-level average across time points in **Appendix D**.

Teacher-Reported Personality Traits

At every wave of data collection, four teachers (i.e., English language arts, math, science, social studies teachers) rated each student on each of the personality measures. For each personality trait, teachers saw a list of four or five descriptions of the trait and rated each student on a 5-point scale ranging from 1 = *Never true* to 5 = *Always true*. See **Appendix C** for the teacher rating prompts. I reasoned that if personality traits were more observable, teachers would be more likely to agree with each other when rating the same student. Therefore, I operationalized the observability of a particular trait as the interrater reliability of the four teacher ratings using the intraclass correlation coefficient (Shrout & Fleiss, 1979).

Role Model and Peer Nominations

In the first wave of data collection, students nominated two classmates as friends and two classmates as role models for each of the personality measures, except for interpersonal self-control. In subsequent waves, students nominated two classmates as friends and one classmate as a role model of each of the personality traits. See **Appendix B** for the peer nomination prompts.

Grade Point Average (GPA)

Using school administrative records, I calculated GPAs on a 100-point scale by averaging final grades in students' academic courses (English language arts, math, science, social studies) for each of the time points in which students took surveys during the 2019-2020 school year. GPA values below 50 were very rare (n = 26, 0.59%), likely indicated reporting errors, and were set as missing data.

Analytic Strategy

I used ordinary least squares (OLS) regression to predict self-reported personality from student's own and peer's academic performance:

$$\overline{y_i} = \alpha\overline{a_i} + \gamma\overline{b_{-i}} + \delta x_i + \epsilon_i$$

where $\overline{y_i}$ is the average of student i personality self-reports across the four waves of data collection. Term $\overline{a_i}$ is that students' averaged core GPA from school records. Term $\overline{b_{-i}}$ is the average core GPA of the student's friends or role models, depending on the model. Term x_i is a vector of student characteristics (i.e., an indicator for the school student i attends, and dummy codes for gender, ethnicity, English language learner status, special education status, and eligibility for free or reduced-priced meals). ϵ_i is an error term. Models not accounting for student characteristics (i.e., excluding vector x_i are available in **Appendix F**. Negative estimates for γ would suggest that nominating peers with higher GPAs biases questionnaire responding downward, consistent with reference bias.

To model reference bias within-persons, I fit a similar OLS regression model predicting self-reported personality from student's own and peers' academic performance, but without collapsing measures across time and adding fixed-effects for each student, effectively removing all between-student variance. To account for the repeated measures at the student level, I use cluster-corrected standard errors. I estimated the following regression equation:

$$y_{it} = \alpha a_{it} + \gamma b_{-it} + \theta_i + \epsilon_{it}$$

where y_{it} is student i personality self-report at time t. Term a_{it} is that student's core GPA from school records during time t. Term b_{-it} is the average core GPA of the student's friends or role models at time t, depending on the model. Term θ_i represents fixed effects for each student. ϵ_{it} is an error term. Models not accounting for student characteristics (i.e., excluding vector x_i) are available in **Appendix F**. Negative estimates for γ would suggest that nominating peers with higher GPAs biases questionnaire responding downward, consistent with reference bias.

Results

Descriptive Statistics

Six-month test-retest reliability for personality traits was high. It ranged between .60 and .63 for academic self-control, .54 and .60 for grit, .53 and .65 for gratitude, .65 and .69 for interpersonal self-control, respectively.

As shown in **Table 1,** personality traits of academic self-control, grit, gratitude, and interpersonal self-control correlated positively with each other, both for self-reports (range of rs = .33 to .63, ps < .001), and even more so for teacher-reports (range of rs = .77 to .94, ps < .001). Both self-reports and teacher-reports correlated with GPA, but teacher-reports were more highly correlated with academic performance (range of rs = .61 to .85, ps < .001) than were self-ratings (range of rs = .13 to .82, ps < .001).

Across the four waves of data collection, students nominated an average of 5.27 friends and 3.20 role models for each trait (except for interpersonal self-control, for which they nominated 1.85 role models). Many of these nominations were reciprocated, especially for friends. On average,

Table 1

Bivariate Correlations and Descriptive Statistics

Variable	1	2	3	4	5	6	7	8	9
1. GPA Composite		.37***	.30***	.16***	.25***	.76***	.79***	.56***	.57***
Self-Reported Personality Composites									
2. Academic Self-Control	.39***		.63***	.50***	.64***	.41***	.39***	.35***	.35***
3. Grit	.32***	.62***		.57***	.48***	.27***	.28***	.23***	.20***
4. Gratitude	.13***	.47***	.55***		.42***	.18***	.20***	.21***	.13***
5. Interpersonal Self-Control	.30***	.63***	.46***	.41***		.35***	.32***	.33***	.42***
Teacher-Reported Personality Composites									
6. Academic Self-Control	.82***	.45***	.28***	.14***	.38***		.92***	.78***	.83***
7. Grit	.85***	.43***	.30***	.15***	.34***	.94***		.78***	.75***
8. Gratitude	.61***	.39***	.26***	.19***	.36***	.79***	.79***		.80***
9. Interpersonal Self-Control	.62***	.40***	.21***	.12***	.45***	.84***	.77***	.80***	
M	79.49	3.80	3.80	4.25	3.65	3.41	3.30	3.64	3.62
SD	10.03	0.56	0.57	0.49	0.67	0.96	0.92	0.76	0.91

*Note. n*s ranged from 1,017 to 1,071. Values in the table represent composite scores averaging across the four waves of data collection. Values above the diagonal are bivariate correlations controlling for student demographic characteristics (i.e., dummy codes for gender, ethnicity, English language learner status, special education status, eligibility for free or reduced-priced meals, and school). *** $p < .001$, ** $p < .01$, * $p < .05$

39.3% of students nominated friends who nominated them back within a given time point, a figure that grows to 47.8% when collapsing across time points. In contrast, nominations for role models were far less likely to be reciprocated: 9.71% of ties within time points and 13.6% of ties collapsing across time points were reciprocated, respectively. Students' friend peer networks and role model peer networks were distinct: In only 18-23% of cases, students nominated as friends were also nominated as a role model by the same student. **Figure 3** shows an example of an academic self-control and friendship networks in a single school in Time 1. See **Appendix E** for details on student social networks.

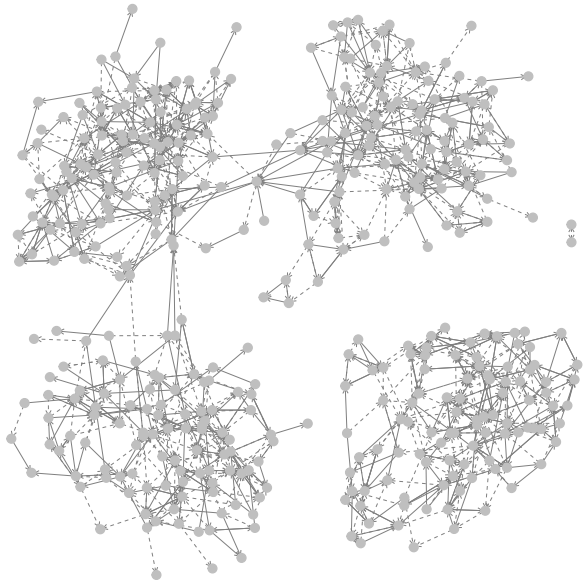
Evidence For Reference Bias in Academically Relevant and Observable Personality Traits

Consistent with prior research, students who earned higher GPAs, self-reported greater academic self-control, grit, gratitude, and interpersonal self-control, respectively (βs ranged from .18 to .42, ps < .002). See estimates on **Tables F1 - F4**.

In contrast, when the role models of academic self-control and grit had higher achievement, students rated themselves *lower* in academic self-control and grit (β = -.15, p < .001; β = -.08, $p < .001$). One standard deviation in the average GPA of the peers whom students nominated as role models of academic self-control and grit was associated with a respective 0.15 and 0.08 *SD* decrease in how students rated their own academic self-control and grit (see **Figure 4**). In contrast, there was no evidence that the academic performance of role models relates to self-ratings on the personality traits of gratitude and interpersonal self-control, respectively. Re-

Figure 3

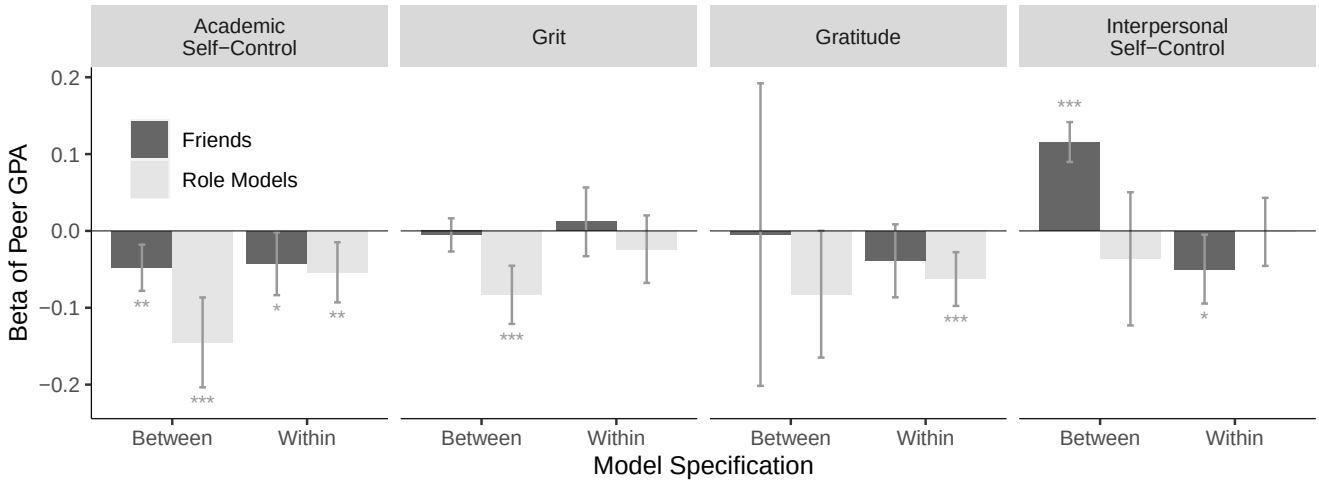
Illustrative Friendship and Academic Self-Control Role Model Network for One School in Time 1



Note. Arrows point from nominator to nominee. Bidirectional arrows indicate reciprocal ties. Filled arrows represent friendship networks, dashed arrows represent academic self-control role model nominations.

Figure 4

Effects of Role Model and Peer GPA on Students Self-Reported Personality



Note. Each bar corresponds to a OLS regression model. Models include controls for own GPA and student characteristics. Error bars represent 95% confidence intervals. *** $p < .001$, ** $p < .01$, * $p < .05$

sults were equivalent when controls for demographics are not included (see **Tables F1 - F4**). See **Appendix F** for complete models, models not controlling for demographic characteristics, and robustness checks.

As shown in **Tables F1 - F4,** academic self-control and grit (β = .42 and .37, respectively, $ps < .001$) were more strongly associated with a student's own core GPA than gratitude and interpersonal self-control (range of β = .18 to .23, $ps < .001$). Academic self-control and grit were also more related to GPA, both in terms of bivariate correlations and unique variance explained in a model predicting GPA from self-reported personality (see **Table 2**).

Teachers were significantly more consistent in rating students' academic self-control and grit compared with the other personality traits. As indicated by the non-overlapping 95% confidence intervals shown in **(Table 2)**, academic self-control and grit were significantly more reliably observed by teachers (ICC = .84 and .83) than were gratitude and interpersonal self-control (ICCs = .73, .81, and .71, respectively).

Another explanation for the effect could be differences in reliability. If grit and self-control were measured more reliably than gratitude and interpersonal self-control, it would lead to more precise estimates of reference bias and thus a higher probability of observing a significant effect. This does not seem to be the case: Interpersonal self-control (average α = .74) was *more* reliable than grit (average α = .70), suggesting that results were not explainable by differences in the reliability of the self-report measures. See **Table 2** for average alphas across the four time points and **Table D1** details on scale reliability. Finally, correlations between self- and teacher-ratings of personality—indexing consensual validity—were higher for interpersonal self-control (r = .45, $p < .001$), than for grit (r = .30, $p < .001$). See **Table 2** for correlations between teacher- and self-reports of personality.

Evidence for Reference Bias From Role Models, But Not Friends

As shown in **Figure 4,** reference bias effects on academic self-control and grit were diminished when defining peers as friends rather than as role models of a trait. The size of reference bias for academic self-control shrinks to a third of its size, reference bias for grit shrinks to a statistically non-significant 6.26% of its size. Interestingly, interpersonal self-control shows the opposite effect, with students rating themselves higher in these traits when their friends have higher GPAs, with effect sizes of comparable magnitude to the original reference bias effects for role models (0.12 *SDs* higher self-ratings for interpersonal self-control for every *SD* increase in their friends' core GPA). Results were equivalent when controls for demographics were not included (see **Tables F1 - F4**). See **Appendix F** for complete models, models not controlling for demographic characteristics, and robustness checks.

Reference Bias Is a Between-Person Phenomenon

As shown in **Figure 4,** reference bias effects shrink when estimated within-people. Relative to our main specification, reference bias for academic self-control shrinks to 40% of its size, and reference bias for grit shrinks to a statistically non-significant 32% of its size.

Table 2

Intraclass Correlation Coefficients for Interrater Reliability of Teacher Ratings

Personality Trait	r_{GPA}	β_{GPA}	ICC	p	[95% CI]	α	$r_{\text{teacher-report}}$
Academic Self-Control	.39	.30	.84	<.001	[.84-.85]	.75	.45
Grit	.32	.14	.83	<.001	[.83-.84]	.70	.30
Gratitude	.13	-.12	.73	<.001	[.72-.74]	.70	.19
Interpersonal Self-Control	.30	-.03	.81	<.001	[.81-.82]	.74	.45

Note. r_{GPA} is the bivariate correlation between the self-reported trait and GPA, β_{GPA} is the standardized regression coefficient in a model with each trait and student characteristics predicting GPA. Intraclass correlation coefficients were calculated with the ICC_{3k} formula described in Shrout and Fleiss (1979), $r_{\text{teacher-report}}$ is the bivariate correlation between the self-reported and the teacher-reported trait.

The lack of within-person effects was not explainable because of lack of variation in the outcome. Intraclass correlation coefficients for each of the self-reports range from .53 to .64, and for each operationalization of role-model peer GPA from .18 to .27. Results were equivalent when controls for demographics were not included, and similar when using an alternative specification without student fixed effects. See **Appendix F** for complete models, models not controlling for demographic characteristics, and robustness checks.

Discussion

Reference bias is an understudied limitation of questionnaires. Capitalizing on a longitudinal dataset of over 1,000 students, we find evidence of reference bias for academic self-control and grit but not for gratitude or interpersonal self-control. Counterintuitively, an individual's reference points appear to be more influenced by their peer role models than by their close friends. Finally, reference points appear to differ across people more than than within people over time.

Contrary to popular adages stating that we are the average of the people we spend the most time with, differences in the academic performance of friends were associated with smaller differences in how students rated themselves compared with the academic performance of role models. Moreover, these two groups were mostly non-overlapping, suggesting that friendship and role model networks are distinct. If students spend more time with friends than role models, we would expect that they would have more availability of information regarding friends rather than role models. The fact that role model GPA was a stronger predictor than friend GPA, suggests that reference points are determined more by the *relevance* rather than the availability of peer-related information.

Reference bias was observed only in the intrapersonal traits of grit and academic self-control, not in the interpersonal traits of gratitude and interpersonal self-control. This finding suggests that our operationalization of reference bias is as valid as the proxies used to index behavior. Reference points are likely influenced by domain-specific behavioral cues in our peers, and reference points for traits less related to GPA are less impacted by it. Thus, if these results point to limitations of our operationalization method rather than to substantive mechanisms of reference bias, it would be premature to rule out the possibility of reference bias for interpersonal traits.

Reference bias was also moderated by observability. Grit and academic self-control showed more reference bias as well as higher observability as indexed by interrater reliability. This suggests that reference points are informed by what we perceive in our peers, consistent with the social learning and social norms literatures (Bandura, 1971; Cialdini, 2007).

Finally, the smaller effect sizes when measuring reference bias within students suggest that reference points are not mercurial. Perhaps it is not sensible to think that reference points change every semester, as students shift whom they spend time with or whom they think exemplifies a particular trait. Changing peer groups might influence reference points through an *updating* rather than *replacing* process: that is, shifting peer groups may provide evidence to adjust our reference points rather than completely replacing them. Over the relatively short run of 2 years, and in a relatively stable environment as high school, there simply might not be enough changes in peer composition to meaningfully alter reference points. If this is accurate, we would expect to see within-person reference bias by extending the time horizon or by examining developmental stages where more pronounced shifts in peer groups and identity might occur, such as significant life transitions (e.g., starting college), or meaningful personality change interventions (e.g., undergoing psychotherapy).

Limitations and Future Directions

Several limitations of this work should be considered and addressed in future research.

First, the observational nature of this study precludes a causal interpretation of our results. While we speculate that peer networks impacted student self-reported personality, we cannot rule out the possibility that real differences in stu-

dents' personality caused differences in peer nomination patterns and in the academic performance of these peers. To causally test reference bias, future studies should manipulate peer groups or participants' reference points.

Second, we had to rely on core GPA as a proxy for a peer's personality. This probably explained why reference bias was stronger for the more academically relevant traits of academic self-control and grit. Ideally, when estimating reference bias effects for interpersonal traits, other more valid proxies for relevant trait behaviors should be used. However, there is no easy solution to this problem, because other alternatives have their own set of limitations. For example, tasks measuring these traits are often more unreliable (Enkavi et al., 2019) and are also mildly correlated with self-reports, at best (Saunders et al., 2022; Wennerhold & Friese, 2020). In that sense, GPA is uniquely useful in that it aggregates a large amount of performance data across different situations in a more extended period of time (Galla et al., 2019).

Third, our operationalization of peer networks was also limited, in terms of the kinds of peers nominated and the sparsity of the nominations (each student only nominated one or two students per nomination at each time point). Future research could include other kinds of peer nominations that might be potentially relevant (e.g., perceptions of popularity) while including more peers in each category. Including more peers would also help reduce the noise in the estimates: The core GPA of one or two peers is a noisy measure of the characteristics of a student's peer group. As noise biases the estimates toward zero, reference bias effects could be larger, especially in the case of role models, because participants nominated fewer of them in comparison to the number of friends they nominated.

Fourth, our study did not include measures of reference points or task measures of personality. A direct measure of the reference points for comparison would allow us to disambiguate if the smaller effects observed for friends compared to role models are explainable by how these peers differentially affect reference points. Including task measures for the same trait would provide a discriminant test for reference bias. Because task measures do not rely on judgment and interpretation, they should be unaffected by reference bias. Moreover, if the effects of peer achievement on task measures of personality were positive, that would mean that reference bias is obscuring a true positive effect on behavior, suggesting that effect sizes for reference bias are, in effect, larger.

Implications

What are the implications of reference bias for researchers and practitioners? Differences in peer composition are related to biased responses of self-report questionnaires, suggesting that research results that rely on these responses, be they comparative, correlational, or intervention-focused, might be biased. *Comparisons* will be biased when there are systematic differences in reference points or peer composition of the groups being compared. For example, the cross-national comparisons showing low conscientiousness in Asian countries (Mõttus et al., 2012) or the null differences in self-regulation of charter school students (Dobbie & Fryer, 2013) likely reflect differences in reference points rather than real personality differences between the groups. Additionally, reference bias suggests that *correlations* between self-reports of personality and future outcomes are likely underestimated, because those with higher reference points will systematically rate themselves lower on personality traits. Finally, reference bias might be relevant in the context of *intervention* research. If interventions change behavior as much as they change the reference points by which behavior is judged, reference bias will obscure positive intervention effects. In practice, reference bias suggests that questionnaires can have limitations even in low-stakes situations. If people hold different reference points, their responses will still be biased even if they are not editing their responses. This suggests that despite the importance of personality and character development, it might be premature to use these scales to inform policy decisions outside of research settings (Duckworth & Yeager, 2015).

Reference bias, however, does not preclude the potential utility of questionnaires. Rather, we should be cautious of how we interpret questionnaire data and triangulate it with evidence from different methods possessing complementary sets of strengths. Behavioral tasks that ask participants to display behavior directly do not require participants to judge their behavior and thus are immune to bias arising from differences in judgment and interpretation. Questionnaires provide reliability because, all in all, people are remarkably accurate aggregators and synthesizers of their own experiences, translating them into questionnaire responses that can be collected cost-effectively and at large scale. Experienced observers might be less affected by judgment biases because their broader experience observing multiple people might allow them to have more general reference frames for evaluation (Feng et al., 2022). Moreover, observers allow for the recording of behavior in its natural context, maximizing ecological validity. Finally, sophisticated methods for translating naturalistic data (e.g., text posted on social media, smartphone data, etc.) into psychological measurement are rapidly increasing in quality and popularity (see Tay et al., 2020). Perhaps these measures can be leveraged to complement questionnaires in a portfolio of measurement.

More broadly, we cannot make sense of the world in absolute terms. Comparisons inform our self-perception (Morina, 2021) and all judgment requires us to make comparisons (Mussweiler, 2003). We all must draw on our limited experience of a tiny sliver of the world to make sense of the whole of it. Throughout history, this tiny sliver has kept

growing, with the average person being exposed to a wider set of different people. Perhaps as we are exposed to more people and the diversity of the set of people we are exposed to increases, not only do our wider reference frames make us better at answering questionnaires, but also hopefully allow us to perceive ourselves more objectively, understand the world more broadly, and behave in ways conducive to our and others' well-being. If comparison is inevitable, perhaps only *narrow* comparison is the thief of joy.

References

Bandura, A. (1971). *Social learning theory*. General Learning Press.

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin*, *130*(6), 887–919. https://doi.org/10.1037/0033-2909.130.6.887

Casey, B. J. (2015). Beyond Simple Models of Self-Control to Circuit-Based Accounts of Adolescent Behavior. *Annual Review of Psychology*, *66*(1), 295–319. https://doi.org/10.1146/annurev-psych-010814-015156

Cialdini, R. B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika*, *72*(2), 263–268. https://doi.org/10.1007/s11336-006-1560-6

Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in Reflected Glory: Three (Football) Field Studies. *Journal of Personality and Social Psychology*, *34*(3), 366–375.

Dahl, R. E., Allen, N. B., Wilbrecht, L., & Suleiman, A. B. (2018). Importance of investing in adolescence from a developmental science perspective. *Nature*, *554*(7693), 441–450. https://doi.org/10.1038/nature25770

Denissen, J. J. A., Bleidorn, W., Hennecke, M., Luhmann, M., Orth, U., Specht, J., & Zimmermann, J. (2018). Uncovering the power of personality to shape income. *Psychological Science*, *29*(1), 3–13.

Dobbie, W., & Fryer, R. G. (2013). Getting Beneath the Veil of Effective Schools: Evidence From New York City. *American Economic Journal: Applied Economics*, *5*(4), 28–60. https://doi.org/10.1257/app.5.4.28

Dobbie, W., & Fryer, R. G. (2015). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy*, *123*(5), 985–1037. https://doi.org/10.1086/682718

Duckworth, A. L., Weir, D., Tsukayama, E., & Kwok, D. (2012). Who Does Well in Life? Conscientious Adults Excel in Both Objective and Subjective Success. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00356

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251. https://doi.org/10.3102/0013189X15584327

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit scale (Grit–S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Feng, S., Han, Y., Heckman, J. J., & Kautz, T. (2022). Comparing the reliability and predictive power of child, teacher, and guardian reports of noncognitive skills. *Proceedings of the National Academy of Sciences*, *119*(6), e2113992119. https://doi.org/10.1073/pnas.2113992119

Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., D'Mello, S. K., Finn, A. S., & Duckworth, A. L. (2019). 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 [Publisher: American Educational Research Association]. *American Educational Research Journal*, *56*(6), 2077–2115. https://doi.org/10.3102/0002831219843292

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451–464. https://doi.org/10.1016/j.labeco.2012.05.014

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us?: The case of conscientiousness. *Psychological Science*, *19*(4), 309–313. https://doi.org/10.1111/j.1467-9280.2008.02085.x

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918. https://doi.org/10.1037/0022-3514.82.6.903

Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., & Kiguel, S. (2020). School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights*, *2*(4), 491–508.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise. How to Overcome the High, Hidden Cost of Inconsistent Decision Making. *Harvard Business Review*, 38–46.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Harper Collins.

Krosnick, J. (1999). Survey Research. *Annual Review of Psychology*, *50*, 537–567.

Larson, R., & Richards, M. H. (1991). Daily Companionship in Late Childhood and Early Adolescence: Changing Developmental Contexts. *Child Development*, *62*, 284–300.

Lira, B., O'Brien, J., Peña, P., Galla, B. M., D'Mello, S., Yeager, D. S., Defnet, A., Kautz, T., Munkacsy, K., & Duckworth, A. L. (2022). Large Studies Reveal How Reference Bias Limits Policy Applications of Self-Report Measures. *Under Review*.

Lundberg, S. (2019). Noncognitive Skills as Human Capital. *Education, Skills, and Technological Change: Implications for Future US GDP Growth* (pp. 219–250). University of Chicago Press.

Malin, H., Liauw, I., & Damon, W. (2017). Purpose and Character Development in Early Adolescence. *Journal of Youth and Adolescence*, *46*(6), 1200–1215. https://doi.org/10.1007/s10964-017-0642-3

Marsh, H. W. (1987). The Big-Fish-Little-Pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295.

Morina, N. (2021). Comparisons Inform Me Who I Am: A General Comparative-Processing Model of Self-Perception. *Perspectives on Psychological Science*, *16*(6), 1281–1299.

Mõttus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., Ah–Kion, J., Amoussou–Yéyé, D., Bäckström, M., Barkauskiene, R., Barry, O., Bhowon, U., Björklund, F., Bochaver, A., Bochaver, K., de Bruin, G. P., Cabrera, H. F., Chen, S. X., Church, A. T., … Ng Tseung, C. (2012). Comparability of Self–Reported Conscientiousness across 21 Countries. *European Journal of Personality*, *26*(3), 303–317. https://doi.org/10.1002/per.840

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*(3), 472–489. https://doi.org/10.1037/0033-295X.110.3.472

Oishi, S., & Roth, D. P. (2009). The role of self-reports in culture and personality research: It is too early to give up on self-reports. *Journal of Research in Personality*, *43*(1), 107–109. https://doi.org/10.1016/j.jrp.2008.11.002

Paulhus, D. L., & Vazire, S. (2007). The self-report method [Publisher: Guilford;]. *Handbook of research methods in personality psychology*, *1*(2007), 224–239.

Peng, K., Nisbett, R. E., & Wong, N. Y. C. (1997). Validity problems comparing values across cultures and possible solutions. *Psychological Methods*, *2*(4), 16.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. [Publisher: American Psychological Association]. *Psychological bulletin*, *94*(1), 18.

Sacerdote, B. (2011). Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far? *Handbook of the Economics of Education* (pp. 249–277). Elsevier. https://doi.org/10.1016/B978-0-444-53429-3.00004-1

Sackett, P. R. (2011). Faking in Personality Assessments. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New Perspectives on Faking in Personality Assessment* (pp. 330–344). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0091

Saunders, B., Milyavskaya, M., & Inzlicht, M. (2022). *Longitudinal evidence that common neurocognitive assessments of self-regulation do not predict everyday goal pursuit* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/fscgj

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of Retrieval as Information: Another Look at the Availability Heuristic. *Journal of Personality and Social Psychology*, *61*(2), 195–202.

Schwarz, N., & Oyserman, D. (2001). Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction. *American Journal of Evaluation*, *22*(2), 127–160.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, *86*(2), 420–428.

Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences*, *9*(2), 69–74. https://doi.org/10.1016/j.tics.2004.12.005

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and Validity Issues in Machine Learning Approaches to Personality Assessment: A Focus on Social Media Text Mining. *European Journal of Personality*, *34*(5), 826–844. https://doi.org/10.1002/per.2290

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, *26*(2), 169–181. https://doi.org/10.1108/QAE-06-2017-0034

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Tsukayama, E., Duckworth, A. L., & Kim, B. (2013). Domain-specific impulsivity in school-age children. *Developmental Science*, *16*(6), 879–893. https://doi.org/10.1111/desc.12067

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP Middle Schools: Impacts on Achievement and Other Outcomes* (tech. rep.). Mathematica Policy Research.

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., Coen, T., & Goble, L. (2015). *Understanding the Effect of KIPP as it Scales: Volume I, Impacts on Achievement and Other Outcomes* (tech. rep.). Mathematica Policy Research.

Wennerhold, L., & Friese, M. (2020). Why Self-Report Measures of Self-Control and Inhibition Tasks Do Not Substantially Correlate (S. Vazire & S. Vazire, Eds.). *Collabra: Psychology*, *6*(1), 9. https://doi.org/10.1525/collabra.276

West, M. R. (2016). Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. *Evidence Speaks Reports*, *1*(13), 7.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, *38*(1), 148–170. https://doi.org/10.3102/0162373715597298

Appendix A
List of Self Report-Items

Academic Self-Control

In the last month...

1. I came to class prepared.
2. I followed directions.
3. I got to work right away instead of waiting around until the last minute.
4. I paid attention, even when there were distractions.
5. I stayed focused when doing independent work.

Grit

In the last month...

1. I finished whatever I started.
2. I stuck with a project or activity for more than a few weeks.
3. I tried very hard even though I failed.
4. I stayed committed to my goals, even if they took a long time to complete.
5. I kept working hard even if I felt like quitting.

Gratitude

In the last month...

1. I appreciated when other people helped me.
2. I showed that I appreciated the good things I have in my life.
3. I expressed appreciation by saying thank you.
4. I did something nice for someone else as a way of saying thank you.
5. I had so much in life to be thankful for.

Interpersonal Self-Control

In the last month...

1. I stayed calm even when others bothered or criticized me.
2. I allowed others to speak without interruption.
3. I was polite to classmates.
4. I controlled my temper.
5. I behaved well even when I was upset.

Appendix B
Peer Nomination Prompts

T1 Prompt

Friends. In the last month, I have spent time with these students from my school (this could be time spent in school, out of school, or online). Name 2 students in your grade.

Grit. Imagine a difficult skill that takes a long time to master, like juggling, playing the piano, or learning a new language. Now, thinking about the other kids in your class, who is most likely to do whatever it takes to master that difficult skill? Name 2 students in your grade.

Academic Self-Control. Imagine that you are given a boring assignment in class. The teacher is working with one student for a long time. Who is most likely to work hard on the assignment even when the teacher is not looking? Name 2 students in your homeroom.

Gratitude. Imagine that a new student arrives in your class and brings everyone candy for Halloween. Who is most likely the first person to go up to this new student and show their thanks? Name 2 students in your grade.

T2 - T4 Prompt

Friends. In the last month, I have spent time with these students from my school (this could be time spent in school, out of school, or online). Name 2 students in your class.

Grit. Write the name of one student who will work towards a super challenging long-term goal with passion and perseverance.

Academic Self-Control. Write the name of one student who will stay focused in class, even when there are distractions.

Interpersonal Self-Control. Write the name of one student who will stay calm, even when angry or upset.

Gratitude. Write the name of one student who will show that they are thankful for what they are given.

Appendix C
Teacher Rating Prompts

Teachers saw an introductory page stating "Now, we are going to ask you some questions about the individual students you teach." They then rated each student they taught using a 5-point Likert scale ranging from 1 = *Never True* to 5 = *Always True*. Please rate [Name of the child] on the following 5 character strengths:

During the past month...

Grit

1. Finished whatever s/he started.
2. Stuck with a project or activity for more than a few weeks.
3. Tried very hard even though s/he failed.
4. Stayed committed to her/his goals, even if they took a long time to complete.
5. Kept working hard even if s/he felt like quitting.

Academic Self-Control

1. Came to class prepared.
2. Followed directions.
3. Got to work right away instead of waiting around until the last minute.
4. Paid attention, even when there were distractions.
5. Stayed focused when doing independent work.

Interpersonal Self-Control

1. Stayed calm even when others bothered or criticized her/him.
2. Allowed others to speak without interruption.
3. Was polite to classmates.
4. Controlled her/his temper.
5. Behaved well even when s/he was upset.

Gratitude

1. Appreciated when other people helped her/him.
2. Showed that s/he cared and appreciated the good things that s/he have in my life.
3. Expressed appreciation by saying thank you.
4. Did something nice for someone else as a way of saying thank her/him.
5. Had so much in life to be thankful for.

Appendix D
Cronbach's Alphas for Each Time Point

In addition to what is reported in the main text, we calculated Cronbach's alpha for each of the four individual time points, as well as for the person-level average across the 4 time points. See Table D1.

Table D1

Cronbach's Alphas for Each Time Point

Variable	Mean	Overall	T1	T2	T3	T4
Academic Self-Control	.820	.756	.758	.761	.737	.761
Grit	.800	.740	.731	.724	.736	.770
Gratitude	.747	.695	.684	.691	.693	.716
Interpersonal Self-Control	.842	.790	.778	.791	.791	.796

Appendix E
Detailed Descriptives on Peer Nominations

Number of Nominations

As shown in Table E1, students nominated on average nominated around 1.3 peers as friends and role models in Time 1. In Time 2 and onwards, students nominated around the same number of friends, but role model nominations were roughly reduced by half. Note that students were not prompted to nominate role models for interpersonal self-control in Time 1.

Table E1

Number of Nominations in Each Time Point

Variable	T1	T2	T3	T4	Total
Friends	1.39	1.45	1.22	1.20	5.27
Academic Self-Control Role Model	1.34	0.70	0.58	0.58	3.20
Grit Role Model	1.36	0.70	0.58	0.58	3.22
Gratitude Role Model	1.34	0.69	0.58	0.58	3.19
Academic Self-Control Role Model	0.00	0.69	0.58	0.58	1.85

Reciprocal Nomination

As shown in Table E2, students were far more likely to be nominated back by students they nominated as friends (40%), than by students they nominated as role models (10%).

Table E2

Fraction of Reciprocal Nominations in each Time Point

Variable	T1	T2	T3	T4	Total
Friends	.39	.41	.36	.40	.48
Academic Self-Control Role Model	.17	.05	.06	.05	.13
Grit Role Model	.16	.09	.10	.12	.16
Gratitude Role Model	.16	.09	.09	.10	.17
Academic Self-Control Role Model	—	.09	.08	.08	.10

Overlapping Peers

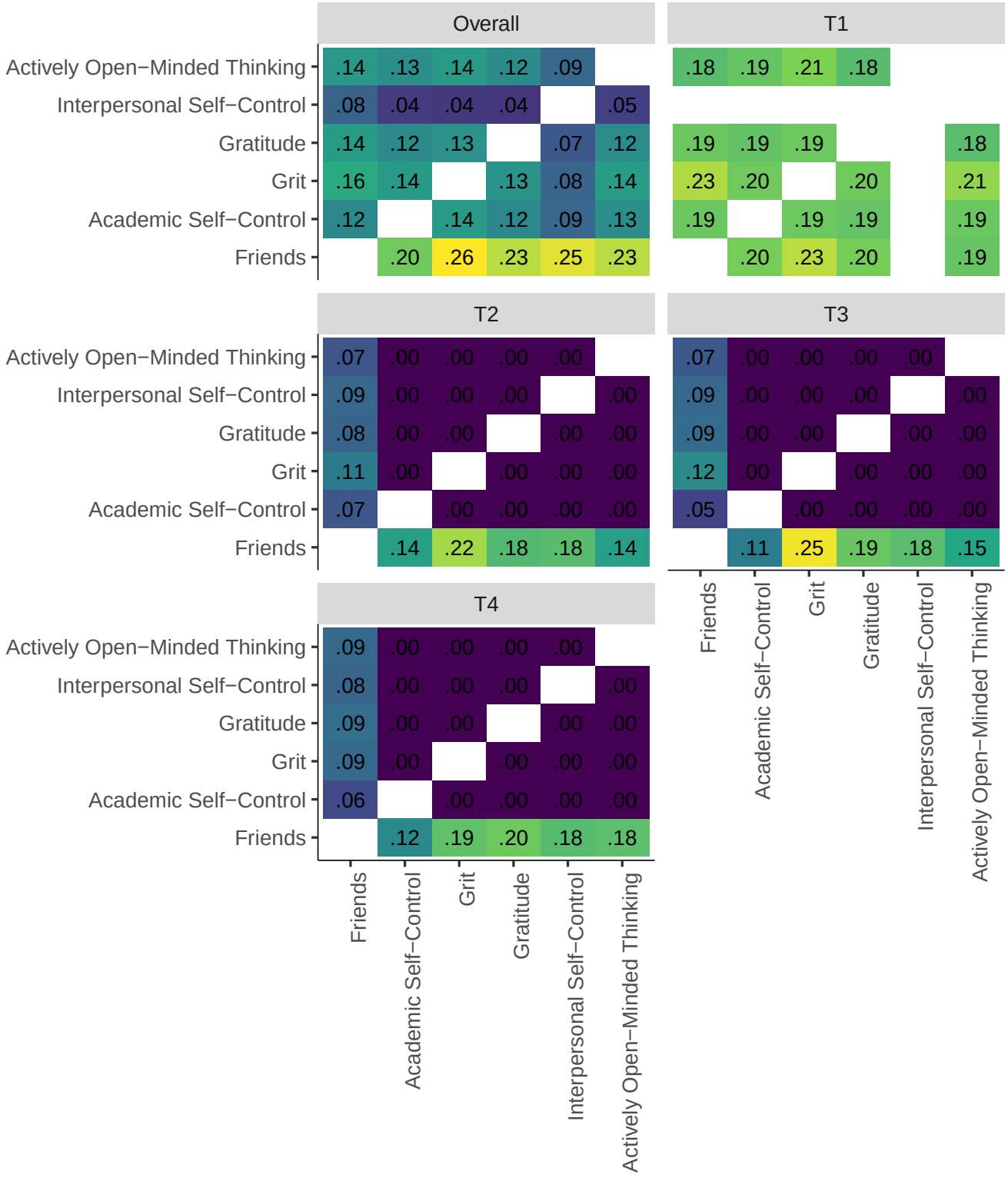
As shown in Figure E1, peer networks for role models and close friends were mostly non-overlapping. Each matrix in the figure represents overlapping nominations within each time point, or in the aggregate across all time points (labelled "overall"). The matrices are not symmetrical because the number of overlapping nominations is divided by the total number of nominations in each category.

Correlations Between Nomination Popularity and Character Traits

As shown in Table E3, there were mild positive correlations between the frequency with which students were nominated as friends or role-models and their own personality characteristics, as indexed by self- and teacher-reports.

Figure E1

Overlapping Nominations in Each Time Point



Note. Each cell represent the fraction of the nominated peers in the x-axis variable that were also nominated in the y-axis variable

Table E3

Bivariate Correlations Between Nomination Popularity and Self-Reported and Teacher-Reported Personality

Variable	1	2	3	4	5
Popularity Measures					
1. Friendship					
2. Academic Self-Control	.25***				
3. Grit	.31***	.55***			
4. Gratitude	.36***	.42***	.44***		
5. Interpersonal Self-Control	.26***	.15***	.10***	.14***	
Self-Reported Personality					
6. Academic Self-Control	-.08***	.20***	.13***	.07***	.05**
7. Grit	-.01	.14***	.16***	.10***	.03
8. Gratitude	.05**	.06***	.10***	.12***	.02
9. Interpersonal Self-Control	-.08***	.12***	.05**	.07***	.14***
Teacher-Reported Personality					
10. Academic Self-Control	.02	.32***	.26***	.18***	.12***
11. Grit	.04	.33***	.29***	.20***	.11***
12. Gratitude	.04	.19***	.17***	.13***	.19***
13. Interpersonal Self-Control	-.01	.20***	.15***	.11***	.20***
M	1.27	0.76	0.76	0.77	0.45
SD	1.45	1.39	1.49	1.23	0.87

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

Appendix F
Robustness Checks

Tables F1 to F4 show OLS models predicting personality from own and peer GPA, with peers operationalized as role models or friends, with and without controls for demographics, as well as within-person.

Figure F1 shows a subgroup analysis where our main model specification is ran separately across time points. The dashed line represents the model using the composites.

Table F2 shows a graphical representation of the different model specifications and how reference bias estimates react to different analytical decisions.

Figure F1

Subgroup Analysis for Each Time Period

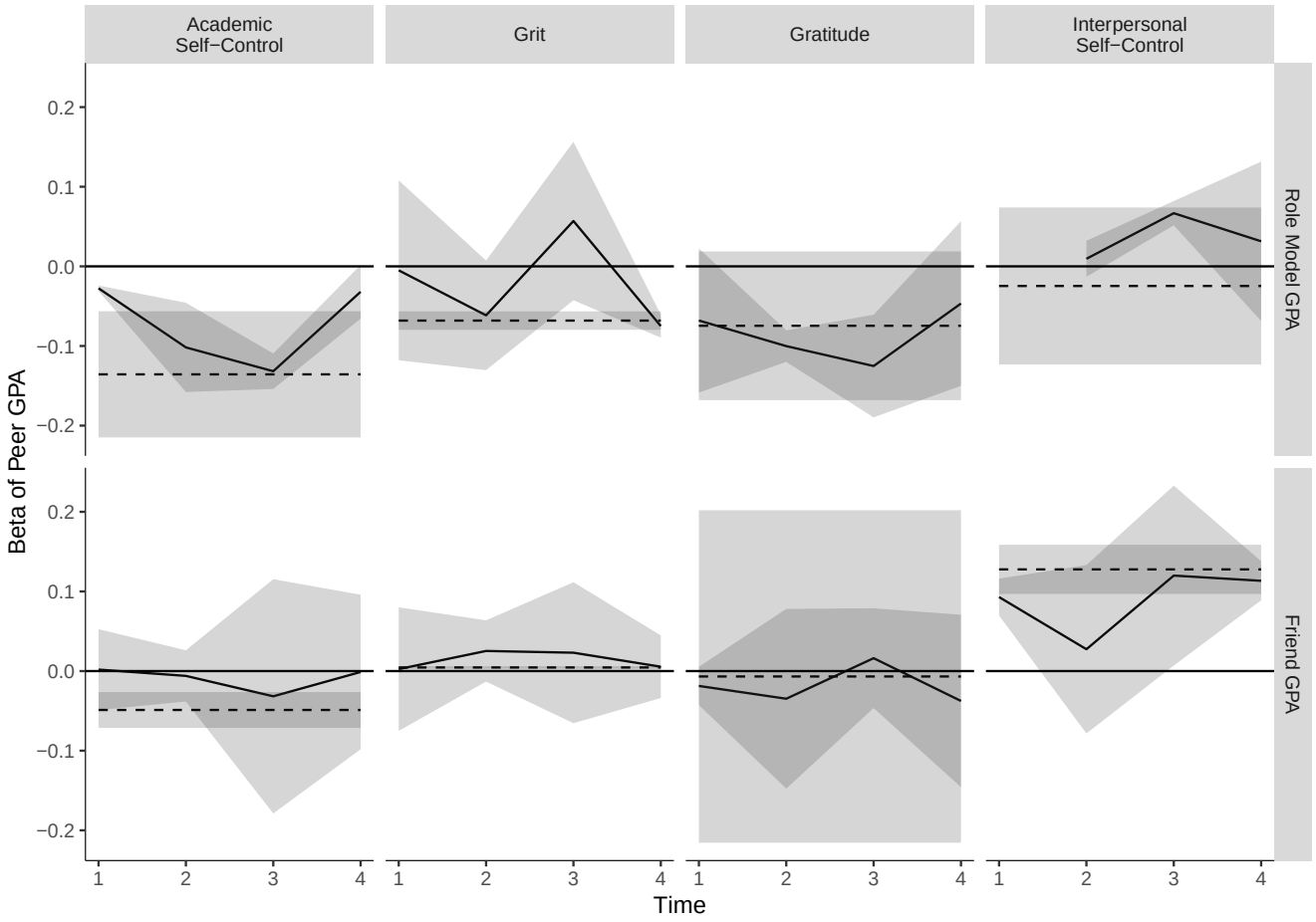


Table F1

OLS Models Estimating Self-Reported Academic Self-Control From Peer GPA, Own GPA, and Demographics

	Academic Self-Control									
	Role Models					Friends				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Role Model Core GPA	−.145***	−.146***	−.107***	−.105***	−.034					
	(.030)	(.036)	(.022)	(.022)	(.023)					
Friend Core GPA						−.043*	−.054**	−.048**	−.036***	−.041
						(.021)	(.020)	(.015)	(.007)	(.023)
Own Core GPA	.415***	.447***	.370***	.392***	.373***	.153***	.142***	.393***	.405***	.361***
	(.035)	(.007)	(.028)	(.025)	(.026)	(.033)	(.034)	(.021)	(0.000)	(.028)
School	.030***	−.029***	.034	−.003	−.003			.008	−.051***	.035
	(.006)	(.001)	(.064)	(.053)	(.051)			(.009)	(0.000)	(.061)
Race/Ethnicity										
Caucasian	−.088***		−.055					−.096*		−.059
	(.010)		(.075)					(.039)		(.072)
Asian	.076		.069					.036		.057
	(.084)		(.082)					(.055)		(.082)
Hispanic	−.119***		−.117					−.040		−.080
	(.021)		(.136)					(.070)		(.131)
American Indian	−.837***		−.856**					−.937***		−.879***
	(.014)		(.270)					(.008)		(.251)
Multi-Racial	−.318**		−.231					−.273***		−.199
	(.102)		(.247)					(.038)		(.220)
Demographics										
Female	.001		.001					−.024*		−.010
	(.015)		(.027)					(.011)		(.026)
English Language Learner	.087***		.094**					.080***		.078**
	(.013)		(.029)					(.010)		(.029)
Special Education Student	.017		.037					.032		.043
	(.014)		(.026)					(.019)		(.026)
Eligibile for Free or	−.044		−.028					−.035		−.027
Reduced-Priced Meals	(.035)		(.026)					(.037)		(.026)
Constant	−.001	−.006***	.003	0.000	0.000			.027	.020***	.003
	(.018)	(.001)	(.041)	(.034)	(.033)			(.019)	(0.000)	(.040)
Fixed effects for Student	No	No	No	No	Yes	No	No	No	No	Yes
Composites across time?	Yes	Yes	No	No	No	Yes	Yes	No	No	No
Observations	936	941	2,624	2,631	2,894	2,287	1,991	959	965	2,885
R^2	.196	.179	.150	.134	.127	.683	.698	.185	.168	.140
Adjusted R^2	.186	.176	.146	.133	.126	.568	.584	.174	.166	.137

Note. Values in parenthesis are standard errors. Reference group for race/ethnicity is Black. *** $p < .001$, ** $p < .01$, * $p < .05$

Table F2

OLS Models Estimating Self-Reported Grit From Peer GPA, Own GPA, and Demographics

	Grit									
	Role Models					Friends				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Role Model Core GPA	−.083***	−.107***	−.039	−.055*	−.018					
	(.019)	(.002)	(.022)	(.022)	(.024)					
Friend Core GPA						.012	−.024	−.005	−.032*	.001
						(.023)	(.022)	(.011)	(.016)	(.024)
Own Core GPA	.366***	.327***	.325***	.285***	.266***	.160***	.173***	.325***	.288***	.300***
	(.001)	(.010)	(.028)	(.027)	(.027)	(.039)	(.040)	(.002)	(.006)	(.028)
School	.022	.007***	.021	0.000	0.000			−.019	−.021***	.007
	(.028)	(0.000)	(.065)	(.054)	(.053)			(.023)	(0.000)	(.064)
Race/Ethnicity										
Caucasian	−.116		−.121					−.102**		−.102
	(.060)		(.075)					(.035)		(.074)
Asian	−.251***		−.269**					−.284***		−.282**
	(.042)		(.092)					(.015)		(.093)
Hispanic	−.326**		−.292*					−.322**		−.322*
	(.102)		(.140)					(.109)		(.133)
American Indian	−.914***		−.816					−.794***		−.921
	(.006)		(.548)					(.009)		(.576)
Multi-Racial	−.204*		−.201					−.289***		−.260
	(.088)		(.199)					(.049)		(.174)
Demographics										
Female	−.105***		−.099***					−.112***		−.096***
	(.004)		(.027)					(.017)		(.027)
English Language Learner	.013		.006					.019		.021
	(.009)		(.031)					(.010)		(.030)
Special Education Student	−.041***		−.034					−.038***		−.034
	(.004)		(.027)					(.001)		(.026)
Eligibile for Free or	−.007		−.006					−.011		−.011
Reduced-Priced Meals	(.008)		(.026)					(.016)		(.026)
Constant	.078***	−.008***	.086*	−.001	−.001			.092***	−0.000	.090*
	(.009)	(.001)	(.043)	(.037)	(.035)			(.006)	(.001)	(.041)
Fixed effects for Student	No	No	No	No	Yes	No	No	No	No	Yes
Composites across time?	Yes	Yes	No	No	No	Yes	Yes	No	No	No
Observations	936	941	2,621	2,627	2,894	2,287	1,970	959	965	2,885
R^2	.129	.102	.091	.072	.066	.644	.656	.122	.092	.086
Adjusted R^2	.118	.099	.087	.071	.065	.514	.525	.111	.090	.082

Note. Values in parenthesis are standard errors. Reference group for race/ethnicity is Black. *** $p < .001$, ** $p < .01$, * $p < .05$

Table F3

OLS Models Estimating Self-Reported Gratitude From Peer GPA, Own GPA, and Demographics

	Gratitude									
	Role Models					Friends				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Role Model Core GPA	−.082	−.094*	−.085***	−.091***	−.037					
	(.042)	(.039)	(.021)	(.021)	(.026)					
Friend Core GPA						−.039	−.063***	−.005	−.016	−.029
						(.024)	(.018)	(.100)	(.089)	(.026)
Own Core GPA	.183***	.168***	.167***	.146***	.134***	.072*	.051	.157*	.134	.154***
	(.041)	(.047)	(.031)	(.029)	(.028)	(.035)	(.036)	(.078)	(.089)	(.030)
School	.019	−.050***	.084	−.001	−0.000			−.024**	−.072***	.063
	(.010)	(0.000)	(.075)	(.058)	(.056)			(.009)	(.002)	(.074)
Race/Ethnicity										
Caucasian	−.195***		−.226**					−.171***		−.193*
	(.011)		(.083)					(.042)		(.082)
Asian	−.210***		−.204*					−.242**		−.217*
	(.005)		(.100)					(.074)		(.101)
Hispanic	−.277*		−.316					−.293		−.367*
	(.131)		(.176)					(.161)		(.171)
American Indian	−.649***		−.651***					−.774***		−.719***
	(.008)		(.103)					(.008)		(.069)
Multi-Racial	−.273		−.469					−.397***		−.481*
	(.181)		(.274)					(.064)		(.242)
Demographics										
Female	−.039**		−.025					−.052**		−.030
	(.013)		(.029)					(.020)		(.029)
English Language Learner	.054***		.035					.055***		.048
	(.011)		(.033)					(.010)		(.032)
Special Education Student	−.052***		−.029					−.037***		−.027
	(.006)		(.029)					(.009)		(.029)
Eligibile for Free or	−.008		−.003					−0.000		.006
	(.008)		(.103)					(.008)		(.069)
Reduced-Priced Meals	(.027)		(.029)					(.011)		(.029)
Constant	.100***	.019***	.085	−0.000	−.001			.117***	.030***	.087*
	(.006)	(.002)	(.044)	(.039)	(.038)			(.025)	(.004)	(.043)
Fixed effects for Student	No	No	No	No	Yes	No	No	No	No	Yes
Composites across time?	Yes	Yes	No	No	No	Yes	Yes	No	No	No
Observations	937	942	2,640	2,647	2,893	2,286	1,991	958	964	2,884
R^2	.044	.025	.034	.019	.014	.681	.695	.040	.020	.029
Adjusted R^2	.032	.022	.029	.018	.013	.565	.580	.028	.017	.025

Note. Values in parenthesis are standard errors. Reference group for race/ethnicity is Black. *** $p < .001$, ** $p < .01$, * $p < .05$

Table F4

OLS Models Estimating Self-Reported Interpersonal Self-Control From Peer GPA, Own GPA, and Demographics

	Interpersonal Self-Control									
	Role Models					Friends				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Role Model Core GPA	−.036	−.013	.005	.017	.075**					
	(.044)	(.043)	(.026)	(.026)	(.025)					
Friend Core GPA						−.050*	−.001	.116***	.148***	.053*
						(.023)	(.023)	(.013)	(.021)	(.025)
Own Core GPA	.232***	.261***	.194***	.221***	.215***	.056	−.045	.192***	.210***	.189***
	(.002)	(.011)	(.033)	(.031)	(.028)	(.031)	(.040)	(.042)	(.024)	(.030)
School	−.002	.020***	−.028	.003	0.000			−.023***	−.023***	.001
	(.024)	(.001)	(.077)	(.060)	(.055)			(.003)	(.001)	(.071)
Race/Ethnicity										
Caucasian	.116*		.121					.068***		.064
	(.055)		(.087)					(.011)		(.082)
Asian	.219***		.213*					.156***		.217*
	(.007)		(.097)					(.009)		(.090)
Hispanic	.075		.094					.273		.221
	(.222)		(.173)					(.148)		(.150)
American Indian	−.769***		−.774***					−.847***		−.788***
	(.056)		(.167)					(.036)		(.184)
Multi-Racial	−.407**		−.243					−.332**		−.230
	(.156)		(.222)					(.108)		(.198)
Demographics										
Female	−.066		−.064*					−.073***		−.045
	(.038)		(.031)					(.020)		(.028)
English Language Learner	.007		.016					.024		.018
	(.021)		(.035)					(.017)		(.032)
Special Education Student	−.033		−.035					−.043**		−.042
	(.044)		(.030)					(.013)		(.028)
Eligibile for Free or	−.046		−.046					−.055		−.049
Reduced-Priced Meals	(.040)		(.030)					(.049)		(.027)
Constant	−.107***	−.055***	−.055	−0.000	.001			−.061***	−.014***	−.054
	(.015)	(0.000)	(.050)	(.039)	(.036)			(.004)	(.001)	(.045)
Fixed effects for Student	No	No	No	No	Yes	No	No	No	No	Yes
Composites across time?	Yes	Yes	No	No	No	Yes	Yes	No	No	No
Observations	892	897	1,930	1,935	2,894	2,287	1,232	959	965	2,885
R^2	.085	.065	.070	.052	.068	.722	.757	.125	.103	.084
Adjusted R^2	.072	.062	.064	.051	.067	.621	.634	.114	.100	.080

Note. Values in parenthesis are standard errors. Reference group for race/ethnicity is Black. *** $p < .001$, ** $p < .01$, * $p < .05$

Figure F2

Graphical Representation of Multiple Model Specifications