

Compared to Whom? Identifying the Mechanism of Reference Bias

Benjamin Lira and Angela Duckworth
University of Pennsylvania

When answering the same questionnaire items, people may nevertheless judge themselves against different standards. This leads to systematic error and paradoxical results (e.g., more self-regulated individuals rating themselves lower in self-regulation because their standards are higher). Little is known about the mechanisms of reference bias. Where do standards of comparison come from? In this two-year longitudinal study, $N = 1,071$ students rated their own academic self-regulation, nominated peers who in their view exemplified academic self-regulation and, named up to two close friends. In tandem, we collected teacher ratings of academic self-regulation and report card grades. Using established methods for quantifying reference bias, we found evidence that students compared themselves to role models more so than to close friends.

Keywords: reference bias, social networks, personality traits

Comparison is the thief of joy.

– Proverb

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, we 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).

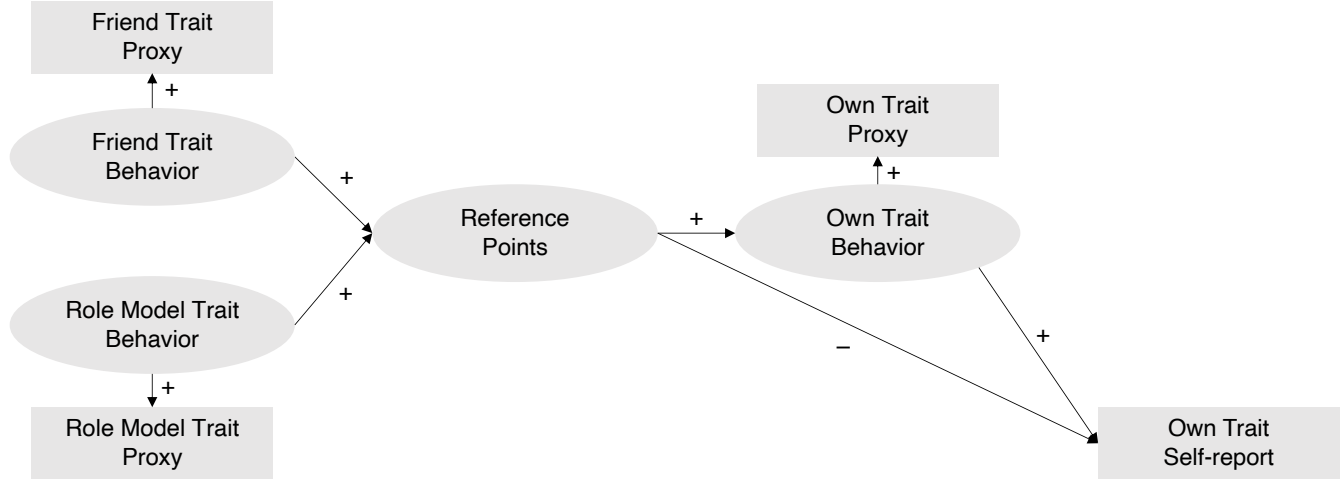
Specifically, we capitalized on a large sample of adolescents to examine evidence for the relative contribution of peer role models and close friends to reference bias in self-assessed self-control. Adolescents in the school setting provide us an ideal testing ground for this question for two reasons: they provide closed social networks from which it is easy to assess friendship and role model relationships, while also providing a context where there are more objective proxies to identify reference bias in self-control.

Reference Bias in Questionnaires

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 1**; 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 character-

 Benjamin Lira

Correspondence concerning this article should be addressed to Benjamin Lira, Psychology Department, University of Pennsylvania, 425 S. University Avenue, Philadelphia, PA, 19104. E-mail: blira@upenn.edu

Figure 1*Theoretical Model for Reference Bias*

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

istics 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).

As shown in **Figure 1**, we 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 close 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).

Reference bias explains otherwise puzzling findings 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 (Möttus 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 effi-

ciency 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.

This evidence on reference bias requires certain untested 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 where do the reference points explaining reference bias come from?

Where Do Our Reference Points Come From?

Differing reference points leading to reference bias might 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 1**, 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, we examine the mechanism of reference bias. First, we establish reference bias in a sample of students transitioning from middle school to high-school. Second, we leverage social network data on friendship and role models to identify the comparative relevance of close friends and role models for reference bias. Third, we 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.

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 Self-Control

Students completed items the Domain-Specific Impulsivity Scale for Children (Tsukayama et al., 2013), which includes items assessing self-control in the academic setting ("I came to class prepared.", "I followed directions.", "I got to work right away instead of waiting around until the last minute.", "I paid attention, even when there were distractions.", "I stayed focused when doing independent work.") 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 $\alpha = .74$ to $.76$.

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

Teacher-Reported Self-Control

At every wave of data collection, four teachers (i.e., English language arts, math, science, social studies teachers) rated each student on self-control. Teachers saw the same

items as the students in the third person, 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.

Role Model and Peer Nominations

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

Grade Point Average (GPA)

Using 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 each of the time points in which students took surveys during the 2019-2020 school year.

Analytic Strategy

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

$$\bar{y}_i = \alpha \bar{a}_i + \gamma \bar{b}_{-i} + \delta x_i + \epsilon_i$$

where \bar{y}_i is the average of student i self-reported self-control across the four waves of data collection. Term \bar{a}_i is that students' averaged core GPA from school records. Term \bar{b}_{-i} is the average core GPA of the student's close 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, we 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, we used cluster-corrected standard errors. We estimated the following regression equation:

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

where y_{it} is student i self-reported self-control 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 close 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 self-control was high (range of $r_s = .60$ and $.63$).

As shown in **Table 1**, self-control correlated with GPA, but teacher-reports were more highly correlated with academic performance ($r = .82, p < .001$) than were self-ratings ($r = .39, p < .001$).

Nominations for friendship were more likely to be reciprocated than nominations for role models of self-control. On average, 39.3% of students nominated close 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: 8% of ties within time points and 13% of ties collapsing across time points were reciprocated, respectively. Students' best friend networks and role model peer networks were distinct: In only 20% of cases, students nominated as close friends were also nominated as a role model by the same student. **Figure 2** shows an example of a 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 for Self-Control is Driven by Role Models

Consistent with prior research, students who earned higher GPAs, self-reported greater self-control ($\beta = .42, p < .001$). See estimates on **Table F1**.

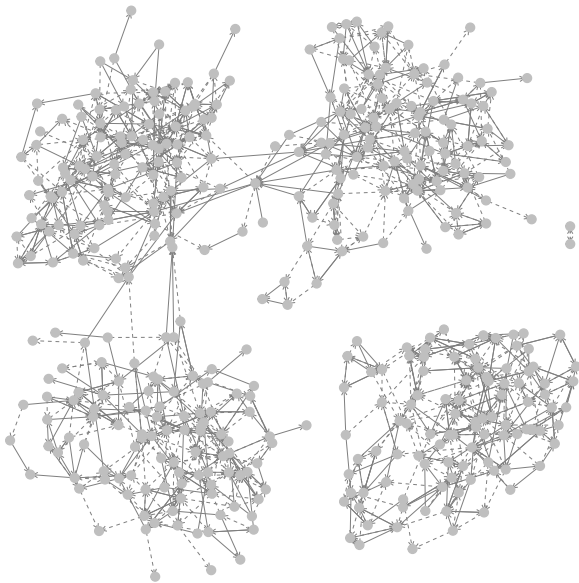
In contrast, when self-control role models had higher achievement, students rated themselves *lower* in self-control ($\beta = -.15, p < .001$). One standard deviation in the average GPA of the peers whom students nominated as role models of self-control was associated with a respective 0.15 *SD* decrease in how students rated their own self-control (see **Figure 3**). Results were equivalent when controls for demographics are not included (see **Table F1**). See **Appendix F** for complete models, models not controlling for demographic characteristics, and robustness checks.

As shown in **Table F1**, self-control was associated with a student's own core GPA ($\beta = .42, p < .001$). Self-control was also related to GPA, both in terms of bivariate correlations ($r_{GPA} = .39$ and unique variance explained in a model predicting GPA from self-control and student characteristics

Table 1*Bivariate Correlations and Descriptive Statistics*

	1	2	3	4	5	6	7
Self Composites							
1. GPA		.37***	.76***	.30***	.27***	.42***	.36***
2. Self-Reported Self-Control	.39***		.41***	.03	.03	.12***	.11**
3. Teacher-Reported Self-Control	.82***	.45***		.21***	.27***	.31***	.39***
Role Model Composites							
4. Role Models' GPA	.45***	.10**	.38***		.70***	.47***	.35***
5. Role Models' Teacher-Reported Self-Control	.40***	.10**	.41***	.73***		.33***	.41***
Close Friend Composites							
6. Close Friends' GPA	.58***	.23***	.49***	.56***	.43***		.71***
7. Close Friends' Teacher-Reported Self-Control	.51***	.21***	.52***	.45***	.48***	.79***	
<i>M</i>	79.49	3.80	3.41	85.34	3.92	80.40	3.46
<i>SD</i>	10.03	0.56	0.96	7.13	0.71	7.19	0.69

Note. *ns* ranged from 912 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). *** $p < .001$, ** $p < .01$, * $p < .05$

Figure 2*Illustrative Friendship and 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 self-control role model nominations.

($\beta_{GPA} = .30$). Teachers agreed with each other when rating their students self-control (ICC = .84, $p < .001$, [95% CI] = [.84 - .85]), and these ratings correlated moderately with self-ratings ($r = .45$, $p < .001$), evidencing consensual validity.

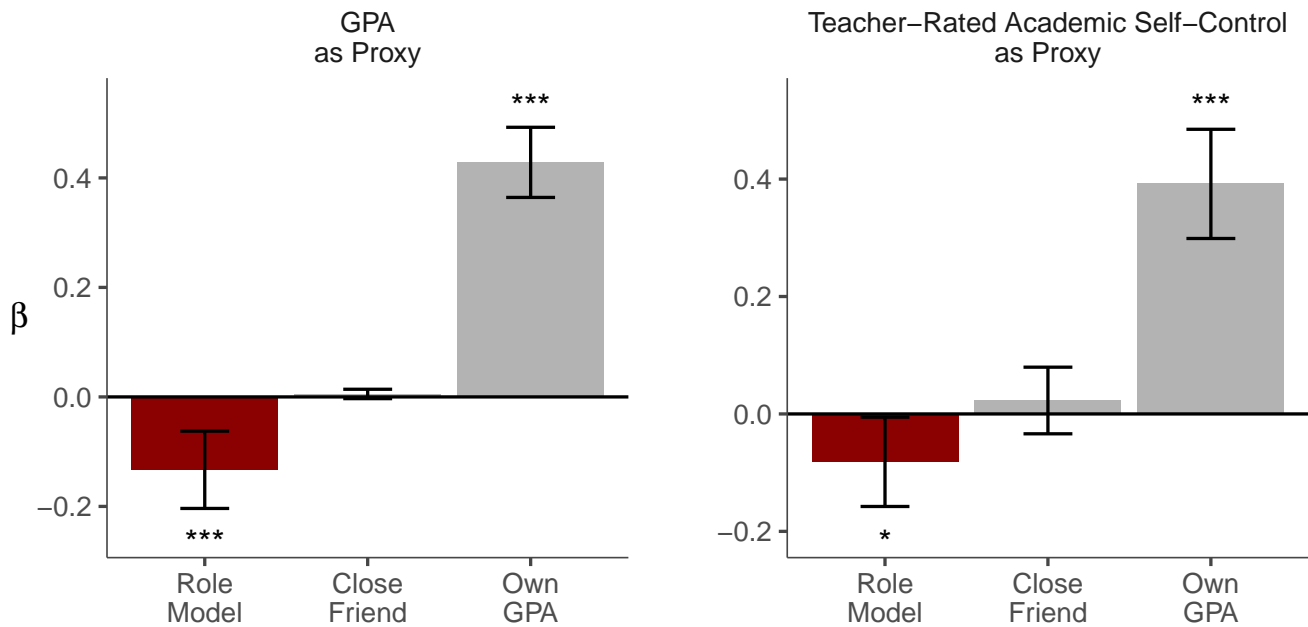
No Evidence for Reference Bias From Close Friends

As shown in **Figure 3**, reference bias effects on self-control was diminished when defining peers as close friends rather than as role models of a trait. The size of reference bias for self-control shrinks to a third of its size ($\beta = .$, $p = .$). Results were equivalent when controls for demographics were not included (see **Table F1** for details). 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 **Supplementary Materials** reference bias effects shrink when estimated within-people. Relative to our main specification, reference bias for self-control shrinks to 40% of its size ($\beta = -.05$, $p = .014$).

The lack of within-person effects was not explainable because of lack of variation in the outcome. Intraclass correlation coefficients for self-control across time points was ICC = X, and ranged from .18 to .27 for each operationalization of role-model peer GPA. 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.

Figure 3*Effects of Role Model and Peer GPA on Students Self-Reported Self-Control*

Note. Each facet corresponds to an OLS regression model. Between-person models include controls for student characteristics, Within-person models include fixed effects for students. Error bars represent 95% confidence intervals. *** $p < .001$, ** $p < .01$, * $p < .05$

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 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 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 close 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 close friends than role models, we would expect that they would have more availability of information regarding close 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.

Finally, the smaller effect sizes when measuring refer-

ence bias within students suggest that reference points are not ephemeral. 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 students' 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 relied on core GPA as a proxy for a peer's self-control, and did not have adequate behavioral proxies that would have allowed to test for reference bias in interpersonal traits such as gratitude or interpersonal self-control. However, proxies to behaviorally index interpersonal traits have a set of limitations. For example, tasks measuring these traits are often more unreliable (Enkavi et al., 2019) and are also not strongly correlated with self-reports (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). Future studies might test for reference bias in other kinds of traits, distinct from self-control.

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 self-control. A direct measure of the reference points for comparison would allow us to disambiguate if the smaller effects observed for close 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 self-control 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, sug-

gesting 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 (Möttus 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>
- 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.
- 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.
- 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.
- Möttus, 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.
- 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.
- 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>
- 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., & Frieze, 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-Reported Self-Control Items

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.

Appendix B

Peer Nomination Prompts

T1 Prompt

Close 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.

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.

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.

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

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...

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.

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
Self-Control	.820	.756	.758	.761	.737	.761

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
Self-Control Role Model	1.34	0.70	0.58	0.58	3.20

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

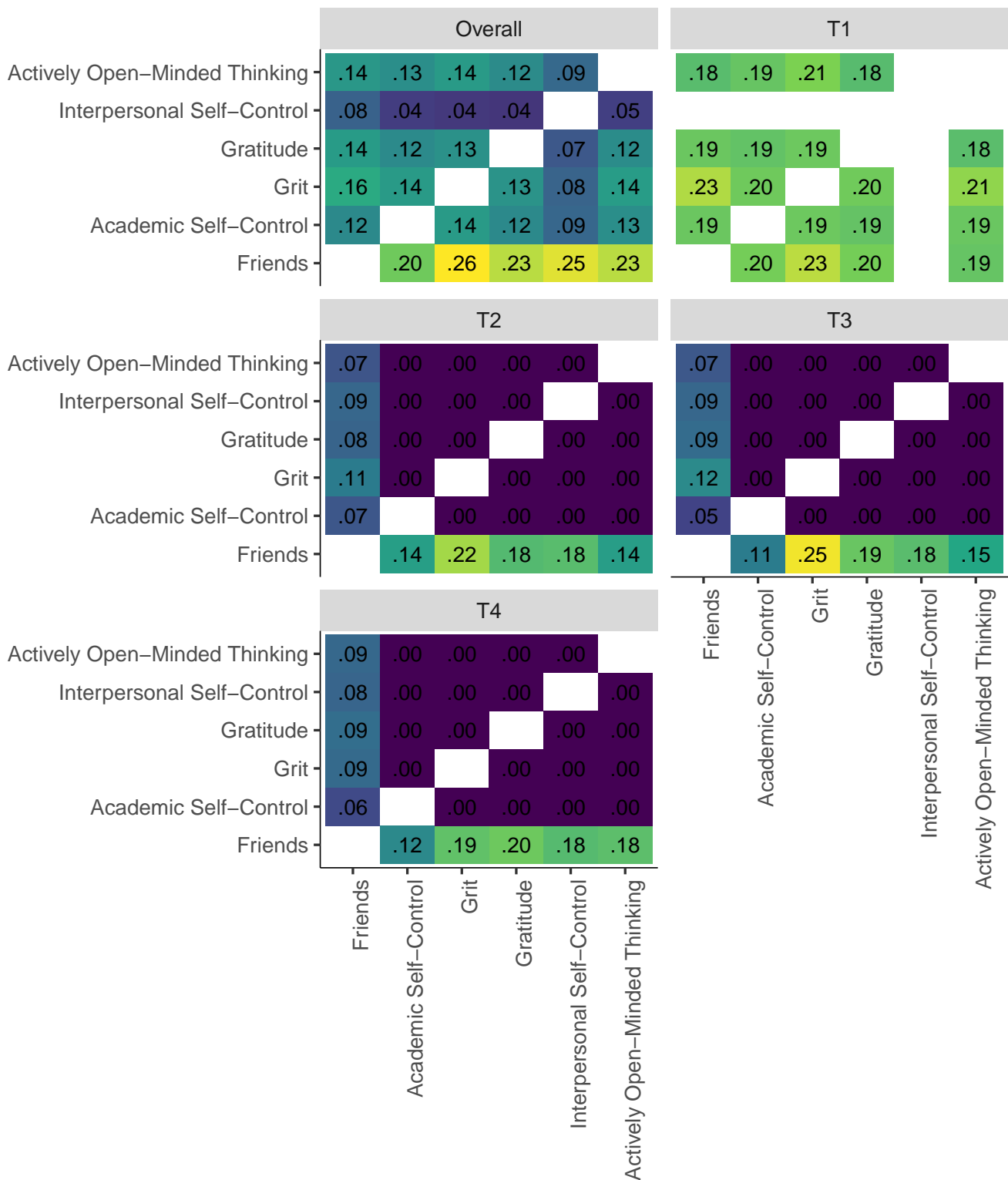
Variable	T1	T2	T3	T4	Total
Friends	.39	.41	.36	.40	.48
Self-Control Role Model	.17	.05	.06	.05	.13

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. [Remaking this plot to exclude outcomes]

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. Self-Control Role Model	.25***				
Self-Control					
6. Self-Reported	-.08***	.20***	.13***	.07***	.05**
10. Teacher-Reported	.02	.32***	.26***	.18***	.12***
<i>M</i>	1.27	0.76	0.76	0.77	0.45
<i>SD</i>	1.45	1.39	1.49	1.23	0.87

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

Appendix F
Robustness Checks

Table F1 shows OLS models predicting self-control 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. [Rerunning this to exclude other outcomes]

Figure F2 shows a graphical representation of the different model specifications and how reference bias estimates react to different analytical decisions. [Remaking to exclude other outcomes]

Table F2 shows multilevel models predicting self-control from own and peer GPA, with peers operationalized as role models or friends, with and without controls for demographics. For each time-varying predictor, the within person average (L1) and the deviations from that average (L2) are included as predictors.

Figure F1

Subgroup Analysis for Each Time Period

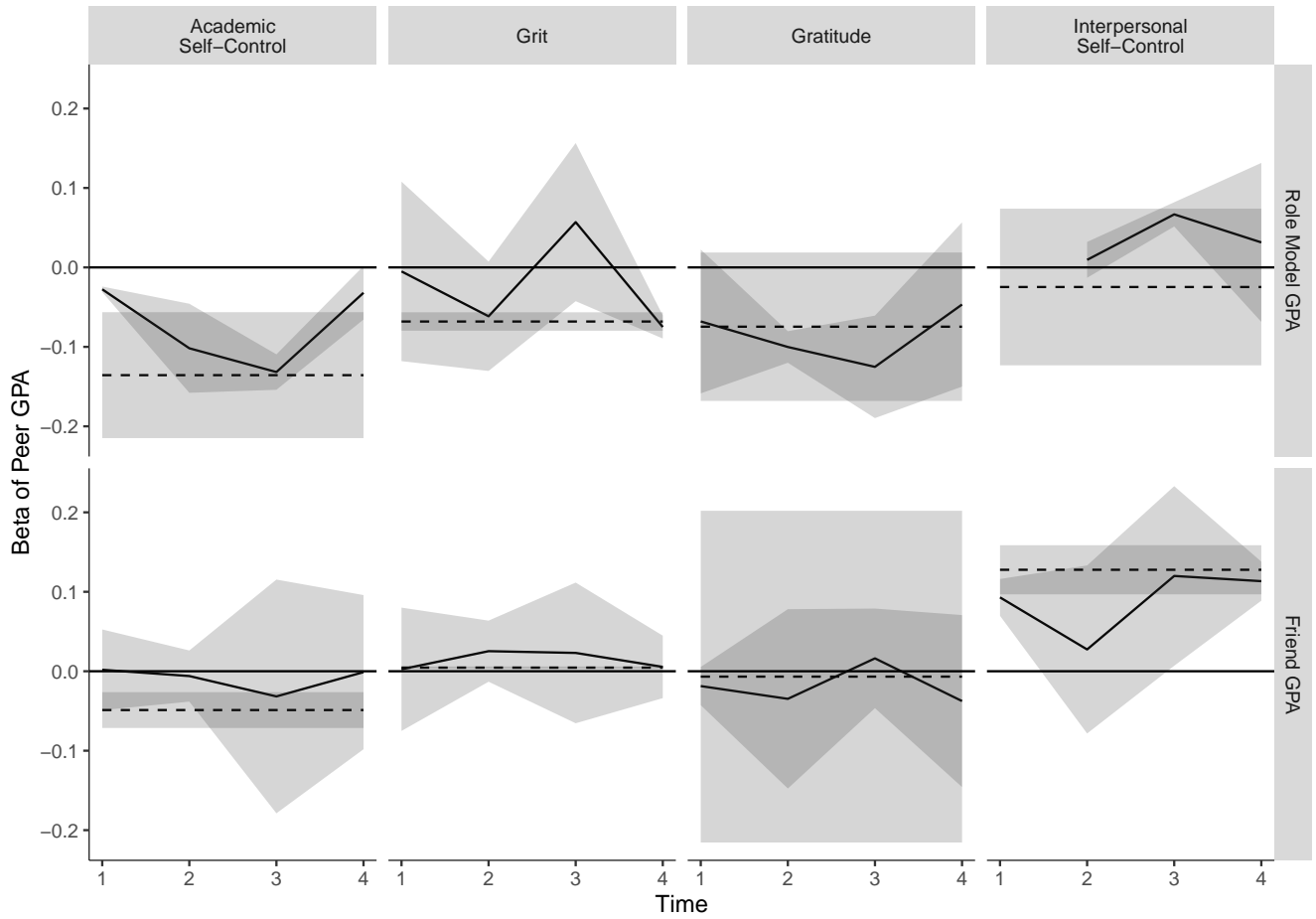


Table F1*OLS Models Estimating Self-Reported Academic Self-Control From Peer GPA, Own GPA, and Demographics*

	Academic Self-Control									
	Role Models					Friends				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Role Model Core GPA	-.145*** (.030)	-.146*** (.036)	-.107*** (.022)	-.105*** (.022)	-.034 (.023)					
Friend Core GPA						-.043* (.021)	-.054** (.020)	-.048** (.015)	-.036*** (.007)	-.041 (.023)
Own Core GPA	.415*** (.035)	.447*** (.007)	.370*** (.028)	.392*** (.025)	.373*** (.026)	.153*** (.033)	.142*** (.034)	.393*** (.021)	.405*** (0.000)	.361*** (.028)
School	.030*** (.006)	-.029*** (.001)	.034 (.064)	-.003 (.053)	-.003 (.051)			.008 (.009)	-.051*** (0.000)	.035 (.061)
Race/Ethnicity										
Caucasian	-.088*** (.010)		-.055 (.075)					-.096* (.039)		-.059 (.072)
Asian	.076 (.084)		.069 (.082)					.036 (.055)		.057 (.082)
Hispanic	-.119*** (.021)		-.117 (.136)					-.040 (.070)		-.080 (.131)
American Indian	-.837*** (.014)		-.856*** (.270)					-.937*** (.008)		-.879*** (.251)
Multi-Racial	-.318** (.102)		-.231 (.247)					-.273*** (.038)		-.199 (.220)
Demographics										
Female	.001 (.015)		.001 (.027)					-.024* (.011)		-.010 (.026)
English Language Learner	.087*** (.013)		.094*** (.029)					.080*** (.010)		.078*** (.029)
Special Education Student	.017 (.014)		.037 (.026)					.032 (.019)		.043 (.026)
Eligible for Free or Reduced-Priced Meals	-.044 (.035)		-.028 (.026)					-.035 (.037)		-.027 (.026)
Constant	-.001 (.018)	-.006*** (.001)	.003 (.041)	0.000 (.034)	0.000 (.033)			.027 (.019)	.020*** (0.000)	.003 (.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*Multi-Level Models Estimating Self-Reported Academic Self-Control From Peer GPA, Own GPA, and Demographics*

	Academic Self-Control					
	(1)	(2)	(3)	(4)	(5)	(6)
Role Models' Core GPA (Mean)	-.169*** (.043)		-.160*** (.039)	-.177*** (.042)		-.158*** (.038)
Role Models' Core GPA (Residual)	-.050** (.019)		-.056** (.019)	-.052** (.019)		-.058** (.019)
Friends' Core GPA (Mean)	.017 (.045)	-.042 (.040)		.039 (.043)	-.030 (.038)	
Friends' Core GPA (Residual)	-.028 (.023)	-.046* (.021)		-.028 (.023)	-.046* (.021)	
Own Core GPA (Mean)	.452*** (.039)	.418*** (.037)	.452*** (.036)	.469*** (.038)	.425*** (.036)	.480*** (.034)
Own Core GPA (Residual)	.179*** (.034)	.172*** (.032)	.169*** (.032)	.182*** (.034)	.174*** (.032)	.172*** (.032)
School	.068 (.066)	.044 (.063)	.067 (.066)	.002 (.053)	-.015 (.052)	.003 (.053)
Female	-.002 (.027)	-.019 (.026)	.003 (.027)			
American Indian	-.847 (.454)	-.919* (.451)	-.844 (.455)			
Asian	.039 (.099)	.010 (.097)	.052 (.098)			
Caucasian	-.103 (.074)	-.094 (.071)	-.093 (.073)			
Hispanic	-.130 (.141)	-.074 (.136)	-.131 (.141)			
Multi-Racial	-.333 (.244)	-.287 (.222)	-.341 (.244)			
English Language Learner	.093** (.033)	.087** (.032)	.094** (.033)			
Special Education Student	.025 (.027)	.041 (.026)	.026 (.027)			
Eligible for Free or Reduced Priced Meals	-.035 (.029)	-.027 (.028)	-.034 (.029)			
Constant	-.036 (.043)	-.012 (.042)	-.038 (.043)	-.051 (.035)	-.029 (.034)	-.050 (.035)
Role Models and Friends Modelled Simultaneously	Yes	No	No	Yes	No	No
Controls for Demographics?	Yes	Yes	Yes	No	No	No
Observations	2,594	2,885	2,624	2,601	2,894	2,631
Log Likelihood	-3,200.279	-3,544.214	-3,229.509	-3,212.663	-3,559.747	-3,242.214
Akaike Inf. Crit.	6,438.557	7,122.427	6,493.019	6,445.326	7,135.495	6,500.427
Bayesian Inf. Crit.	6,540.015	7,222.871	6,502.951	6,502.062	7,182.258	6,547.428

Figure F2*Graphical Representation of Multiple Model Specifications*