Sensitivity of self-reported noncognitive skills to survey administration conditions

Yuanyuan Chen[a], Shuaizhang Feng[b], James J. Heckman[c,d,1], and Tim Kautz[e]

[a]Institute for Advanced Research, Key Laboratory of Mathematical Economics of Ministry of Education, Shanghai University of Finance and Economics, Shanghai 200433, China; [b]Institute for Economic and Social Research, Jinan University, Guangzhou 510632, China; [c]Department of Economics, The University of Chicago, Chicago, IL 60637; [d]The American Bar Foundation, Chicago, IL 60611; and [e]Mathematica, Inc., Princeton, NJ 08540

Contributed by James J. Heckman, November 22, 2019 (sent for review July 2, 2019; reviewed by Armin Falk and Patrick C. Kyllonen)

Noncognitive skills (e.g., persistence and self-control) are typically measured using self-reported questionnaires in which respondents rate their own skills. In many applications—including program evaluation and school accountability systems—such reports are assumed to measure only the skill of interest. However, self-reports might also capture other dimensions aside from the skill, such as aspects of a respondent's situation, which could include incentives and the conditions in which they complete the questionnaire. To explore this possibility, this study conducted 2 experiments to estimate the extent to which survey administration conditions can affect student responses on noncognitive skill questionnaires. The first experiment tested whether providing information about the importance of noncognitive skills to students directly affects their responses, and the second experiment tested whether incentives tied to performance on another task indirectly affect responses. Both experiments suggest that self-reports of noncognitive skills are sensitive to survey conditions. The effects of the conditions are relatively large compared with those found in the program evaluation literature, ranging from 0.05 to 0.11 SDs. These findings suggest that the effects of interventions or other social policies on self-reported noncognitive skills should be interpreted with caution.

noncognitive skills | psychological assessment | personality traits | Big Five | incentives

Cognitive tests—like Intelligence Quotient (IQ) and achievement tests—do not capture important noncognitive skills that matter for success in life and can be shaped through interventions.* Until recently, these skills have largely been disregarded in evaluations of social interventions and educational systems (1). They include conscientiousness, emotional stability, persistence, self-control, social awareness, self-efficacy, and mindfulness. For many important life outcomes, such as educational attainment, health, earnings, and employment, the predictive power of noncognitive skills rivals that of cognitive skills (2, 3). Noncognitive skills are also malleable and can be shaped through interventions and education (4–7).

Because of evidence like this, policy makers and researchers have become increasingly interested in measuring noncognitive skills and typically rely on self-reported measures in which respondents rate their own skills. For example, a group of school districts in California has developed self-reported measures of noncognitive skills for use in their accountability and continuous improvement system (8). Other school districts, such as the District of Columbia Public Schools, are using self-reported measures of noncognitive skills to track progress toward achieving district-wide goals (9). In the international context, Chile has incorporated self-reported measures of noncognitive skills into its school accountability system (10), and the Organisation for Economic Co-operation and Development (OECD) has launched a study that will collect self-reported measures to inform policy (11). In addition, many evaluations of educational programs or other social interventions use self-reported measures of noncognitive skills as key outcomes (12, 13).

At the same time, researchers have warned about reliance on self-reported surveys of noncognitive skills in high-stakes settings or evaluations due to a range of potential biases (12). These biases include "social desirability bias," which arises when respondents consciously or subconsciously supply answers that might be viewed favorably by others (14), and "reference bias," which arises when respondents rate themselves relative to different reference points (15).

These types of biases could arise because any psychological measure is based on a behavior that could depend on incentives or other aspects of a person's situation broadly defined (16). For example, a series of experiments has demonstrated that incentives—part of a person's situation—can affect performance on cognitive tests, such as IQ tests (2, 17, 18). The same types of issues might apply to measures of noncognitive skills. For example, social desirability bias might be stronger in some situations if particular skills are perceived to be valued more in those situations. Similarly, reference bias might arise because respondents compare themselves with their peer group, one part of their situation. However, little is known about the degree to which situations or incentives can affect how people respond to self-reported measures of noncognitive skills.

This study provides experimental evidence on the extent to which incentives in situations related to survey administration can affect student responses on noncognitive skill questionnaires. We focus on one commonly used taxonomy of noncognitive skills known as the Big Five, which is sometimes referred to as the "latitude and longitude" of personality (19). The Big Five comprise 5

Significance

Recent evidence has shown that noncognitive skills matter for success in life and can be shaped through interventions. Because of this evidence, policy makers and researchers have increasingly become interested in measuring noncognitive skills and typically rely on self-reported measures in which respondents rate their own skills. Such self-reports have been applied in program evaluations, as well as school accountability and improvement systems. We demonstrate that self-reports are sensitive to survey administration conditions, including whether a survey administrator describes the skills being assessed and whether respondents receive incentives tied to performance on other tasks. These findings have implications for the interpretation of self-reported measures. Social policies or interventions might affect responses on self-reported noncognitive skills without affecting the skills themselves.

Author contributions: Y.C., S.F., J.J.H., and T.K. designed research; Y.C., S.F., J.J.H., and T.K. performed research; Y.C. and T.K. analyzed data; and Y.C., S.F., J.J.H., and T.K. wrote the paper.

Reviewers: A.F., University of Bonn; and P.C.K., Educational Testing Service.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: jjh@uchicago.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1910731117/-/DCSupplemental.

*In this study, we use the term "noncognitive skills" but do not intend to draw a sharp distinction between this term and similar terms used to describe these skills in the literature, such as socioemotional skills, character, and personality.

constructs, including Openness to Experience (Openness), which relates to curiosity and intellectual pursuits; Conscientiousness, which is the extent to which people are organized and hardworking; Extraversion, which is the extent to which people are outgoing and sociable; Agreeableness, which relates to unselfishness and friendliness; and Emotional Stability (or Neuroticism), which relates to consistency in emotional reactions (20). The Big Five are typically elicited through self-reports, such as the Big Five Inventory (BFI), a 44-item inventory that asks respondents the extent to which they agree to a series of statements (21). For example, one item is "I see myself as someone who tends to be lazy," which has 5 response categories that range from "disagree strongly" to "agree strongly." This study conducted 2 experiments that examined the extent to which aspects of survey administration affect students' responses to the BFI.

The first experiment was designed to test whether providing information about the Big Five affects how students report on their skills. Before completing the BFI, students were randomly assigned to either a treatment or a control group and were separated into different classrooms based on their assignment status. In treatment classrooms, the survey administrator provided instructions for completing the survey and read a description of the Big Five and their importance for life outcomes (*SI Appendix*, Table S1 provides the script used for each of the Big Five traits). In the control group, the survey administrator provided only the basic instructions for completing the survey.

The treatment condition was designed to mimic aspects of noncognitive skill development interventions that define and explain the importance of various skills. Such explanations could conceivably affect the way that students view themselves or perceive the value of certain responses. This experiment provides evidence on whether such programs affect how students report their skills in addition to affecting the skills themselves. Because students received the explanation immediately before administering the questionnaire, any effects of this type of explanation were unlikely to reflect changes in the underlying skill.

The second experiment was designed to test whether a less direct condition could affect students' responses on a noncognitive skill questionnaire. In this experiment, immediately before taking a math test and completing the BFI, students were randomly assigned to 1 of 3 groups: 1) a control group, 2) a treatment group that could receive a certificate of recognition if they performed well on the math test (honor incentive), or 3) a treatment group that could receive financial rewards if they performed well on the math test (financial incentive) (*SI Appendix*, Table S2 provides a detailed description of the experimental groups).

The BFI was administered directly after the math test to study the possibility that the incentives might also affect the way students reported their noncognitive skills, even though the incentive was not directly related to their reports. This experiment sheds light on the possibility that self-reports of noncognitive skills could be sensitive to aspects of survey administration that are less directly related to noncognitive skills. It is also relevant to the intervention literature because programs often consist of several components, such as providing incentives for positive outcomes and delivering instruction designed to boost skills (22).

Results

Conducting 2 randomized experiments, we estimated the sensitivity of self-reported noncognitive skills to survey administration conditions.

Experiment #1: The Effect of Providing Information about Noncognitive Skills. The results from the first experiment show that providing information to students about noncognitive skills can affect how students report their own skills. In the treatment group, the survey administrator read a description of the Big Five that included definitions of the constructs and an overview of how each construct can be beneficial for life outcomes (*SI Appendix*, Table S1). For example, the survey administrator explained Agreeableness as follows: "The third [trait] is Agreeableness, which is related to how friendly, helpful, and trusting people are. A person who is high in Agreeableness is less likely to get into arguments with others. People who are high in Agreeableness are also less likely to commit crimes." In the control group, the survey administrator provided only the standard instructions about completing the test.

The explanation condition affected self-reports of each of the Big Five domains by between 0.05 and 0.11 SDs (Fig. 1); 4 of 5 effects were statistically significant at the 5% level. In addition, the effects were all positive in the sense that they were in the directions that the description suggested were favorable.[†] For example, some dimensions of the Big Five, like Extraversion, are not inherently positive, but the description gave examples when Extraversion could be helpful, such as with jobs that require interactions with others (*SI Appendix*, Table S1).

We also considered whether the treatment condition affected other properties of the Big Five measures, including the reliability (internal consistency) as measured by Cronbach's alpha (*SI Appendix*, Table S3), the correlation with teacher- and peer-reported Big Five measures (*SI Appendix*, Table S4), and the dispersion as measured by the SD (*SI Appendix*, Fig. S1). We found little evidence that the treatment condition systematically affected any of these other properties of the self-reported measures. We also found similar patterns when estimating effects separately for males and females, although the effects were less precisely estimated for each group and somewhat larger for females (*SI Appendix*, Table S5).

The effects of this experiment are consistent with several possible interpretations. First, with a better understanding of the Big Five, students in the treatment group might have been better able to rate themselves. However, our supplementary analyses cast some doubt on this possibility. If students' reports were more accurate as a result of the explanation, then they might also exhibit higher levels of reliability and higher correspondence with reports from teachers and peers, which we did not find. Second, students in the treatment group might have rated themselves more highly in order to view themselves more positively and enhance their self-image. Consistent with this possibility, other evidence suggests that self-image is generally an important determinant of behavior (23). Third, in a similar vein, students might have consciously or subconsciously responded in a way that they thought their teachers or peers would have viewed more favorably, a form of social desirability bias. For example, while students were told that their responses would be confidential, they might have either misunderstood these instructions or thought there was a chance that their responses would have been revealed by accident. Future research might be able to distinguish between these possibilities.

Experiment #2: The Effect of Honor and Financial Incentives. The second experiment shows that incentivizing performance on a math test can affect how students report their noncognitive skills. In particular, students who were assigned to a treatment group that could receive a certificate of recognition (honor incentive) for performing well on a math test reported higher levels of noncognitive skills relative to a control group that received no incentives (Fig. 2). This honor incentive had positive effects (0.07 to 0.10 SDs) on all 5 self-reported measures of noncognitive skills; 4 of 5 effects were statistically significant at the 5% level. Students who were assigned to a treatment group that could receive a financial incentive for their performance on a math test reported similar levels of noncognitive skills relative to the control group. The differences in effects between the honor and financial conditions were statistically significant at the 5% level for Openness, Conscientiousness, and Extraversion (*SI Appendix*, Table S6).

Similar to Experiment #1, we found little evidence that either treatment condition affected other properties of the self-reported

[†]We coded Emotional Stability so that a higher value indicates more stability and lower levels of Neuroticism (the reverse of Emotional Stability).

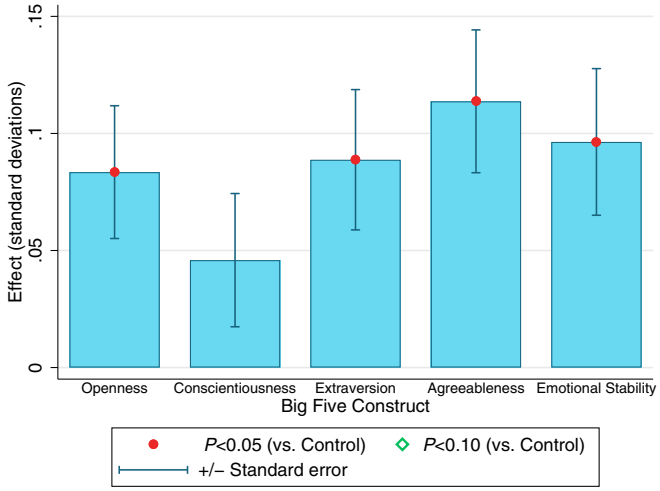

Fig. 1. Effect of the explanation condition on students' self-reported Big Five (Experiment #1). The figure shows estimated effects of the explanation condition on students' self-reports of the BFI (21). The light blue bars represent the effect estimates. The error bars indicate plus and minus 1 SE. The red dots indicate whether the effect is statistically different from 0 at the 0.05 level from a 2-tailed test. The presence of a red dot or green diamond indicates that the effect is statistically different from 0 at the 0.05 and 0.10 levels, respectively, from a 2-tailed test. Each outcome measure was constructed as a Bartlett factor score using the items in the relevant domain (31, 32). The scores were converted into SD units by dividing them by the SD in the control group. The effects were estimated using ordinary least squares. Reported SEs were corrected for heteroscedasticity with the Huber–White sandwich estimator. The model adjusted for gender; cohort; whether the student had a Shanghai hukou; migrant school status; teacher ratings of each student's overall performance, punctuality, and discipline; whether the student was elected to be a leader of the class; baseline measures of the Big Five based on student self-reports from the prior year; and school fixed effects. Observations with missing outcome or covariate data were excluded from the sample. The sample sizes for the analysis of each outcome (from left to right) are n = 3,202, n = 3,223, n = 3,226, n = 3,231, and n = 3,217.

measures, including the reliability (*SI Appendix*, Table S3), the correlation with peer- and teacher-reported measures (*SI Appendix*, Table S4), or the dispersion (*SI Appendix*, Fig. S2). We also found similar patterns when estimating effects separately for males and females, although the effects were somewhat less precisely estimated overall and tended to be larger for males (*SI Appendix*, Table S5).

The findings from Experiment #2 are consistent with several interpretations, each related to how students might consider image-related issues relative to monetary compensation. First, because the honor incentive rewarded students publicly for performance on the math test, it might have caused students to consider how the Big Five would be viewed publicly. For example, if students believed that there was a chance that their results on the Big Five might also be publicized, then the honor incentive could have caused them to enhance the description of their own traits. The monetary incentive would not have had the same effect because it did not involve public visibility. Second, in considering how others value the Big Five, students might have also considered more deeply how they view the Big Five and reported values that enhanced their self-image (23). These interpretations align with other research that has found important differences between motivation based on extrinsic factors (e.g., financial incentives) and motivation based on image concerns (e.g., the perceptions of others or one's self) (24).

Baseline Equivalency and Placebo Tests. To investigate the possibility that the effects we found arose due to chance, we explored whether baseline characteristics were balanced across the treatment and control groups. The data suggest that both Experiment #1 and Experiment #2 achieved baseline equivalence between the treatment and control groups (*SI Appendix*, Tables S7 and S8). For Experiment #1, there were only 2 baseline variables for which the treatment and control groups differed at a 5% level of statistical significance. Importantly, the sample was balanced on baseline measures of the BFI that students reported during the previous year. For Experiment #2, there were no statistically significant differences in baseline variables between the treatment groups and the control group. These results suggest that there were no systematic differences across experimental groups at baseline.

To further rule out the possibility that the effects arose by chance, we conducted a placebo test based on peer-reported measures of the Big Five. After each student completed the self-assessment of the Big Five, students returned to their homeroom classroom and were asked to assess his or her tablemate. Since the students were randomly assigned to treatment and control groups, students were equally likely to assess a peer from either of the two experimental groups. We then estimated the "effect" of a student's own treatment condition on their peer-reported Big Five measures. This analysis serves as a placebo test in that a student's own treatment status should be uncorrelated with that of the student who assesses them, and therefore, on average, we would expect no effect. Across the 2 experiments, only 1 of 15 effect estimates on peer-reported Big Five measures was statistically significant at the 5% level, approximately the number that would arise from chance (*SI Appendix*, Table S9). In addition, the distribution of peer-reported scores was very similar across the treatment and control groups (*SI Appendix*, Figs. S3 and S4). This evidence further suggests that our main findings did not arise due to chance.

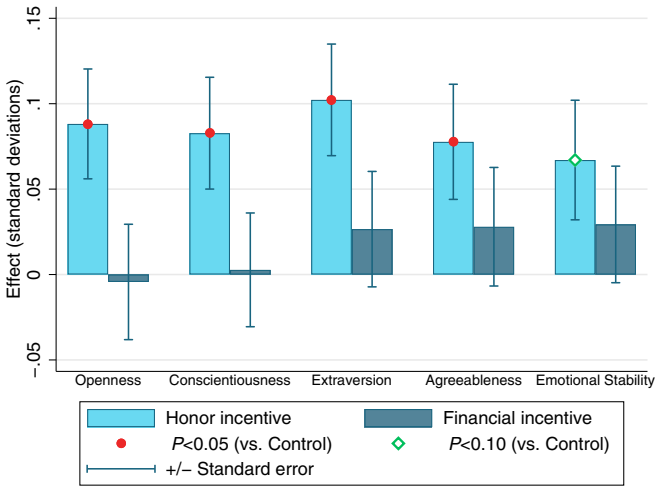

Fig. 2. Effects of the honor and financial incentives on students' self-reported Big Five (Experiment #2). The figure shows estimated effects of the honor and financial incentive conditions on students' self-reports of the BFI (21). The light and dark blue bars represent the effect estimates of the honor and financial incentives, respectively. The error bars indicate plus and minus 1 SE. The red dots and green diamond indicate whether the effect is statistically different from 0 at the 0.05 and 0.10 levels, respectively, from a 2-tailed test. The presence of a red dot or green diamond indicates that the effect is statistically different from 0 at the 0.05 and 0.10 levels, respectively, from a 2-tailed test. Each outcome measure was constructed as a Bartlett factor score using the items in the relevant domain (31, 32). The scores were converted into SD units by dividing them by the SD in the control group. The effects were estimated using ordinary least squares. Reported SEs were corrected for heteroscedasticity with the Huber–White sandwich estimator. The model adjusted for gender; cohort; whether the student had a Shanghai hukou; migrant school status; teacher ratings of each student's overall performance, punctuality, and discipline; whether the student was elected to be a leader of the class; baseline measures of the Big Five based on teacher reports; and school fixed effects. Observations with missing outcome or covariate data were excluded from the sample. The sample sizes for the analysis of each outcome (from left to right) are n = 4,733, n = 4,752, n = 4,796, n = 4,780, and n = 4,773.

Discussion

The effects from both experiments were relatively large compared with effects found in the evaluation literature. For example, a metaanalysis of mostly short-term interventions designed to boost noncognitive skills found effect sizes of 0.22 to 0.27 across 5 domains and 0.57 for another domain (25). The effects from the 2 experiments in this study ranged from 0.05 to 0.11 SDs, which could account for up to half of the magnitude of the effects found in the intervention literature. Given that many interventions share commonalities with the experiments in this study, this evidence suggests that part of the effect of interventions on self-reported measures might operate by affecting how students report their noncognitive skills and do not reflect any effect on true skills.

Similar issues arise in the context of school accountability systems. This possibility is especially relevant given that some school districts and countries are using self-reported measures of noncognitive skills, such as growth mindsets, self-efficacy, and self-esteem (8, 10, 26). One obvious concern with using self-reports for school accountability is that teachers or principals might coach students to give favorable responses on noncognitive skill surveys. However, the evidence in this study suggests that such surveys could also be susceptible to more subtle types of biases. For example, in an earnest effort to boost noncognitive skills, schools might introduce programs that explain the importance of noncognitive skills for life outcomes, similar to the treatment condition in Experiment #1. These types of programs might influence student responses to noncognitive skill questionnaires. Although our study does not reveal whether such influences lead to more or less accurate reports, it does reveal that the self-reports are sensitive to administration conditions in ways that could lead to biases when comparing schools.

Including other types of more objective measures in evaluations or accountability systems could help to mitigate the risks of some of these biases. Alternative measures include third-party reports; school administrative records, such as absences or disciplinary infractions that proxy noncognitive skills (5); innovative modifications designed to improve self-reports (15); task-based measures, such as the Academic Diligence Task (27); or incentivized measures—such as those used in economics—that provide explicit incentives to all respondents, potentially limiting the effect of other differences in survey administration conditions (28). For example, in an evaluation of an intervention, trained observers might provide a more objective view than students in the treatment and control conditions. In such evaluations, measures based on forms of objective behaviors can also suggest different conclusions than self-reported measures. For instance, an evaluation of charter schools found that students in the treatment group who were given access to charter schools had slightly lower levels of self-reported effort in school compared with the students in the control group, although the difference was not statistically significant (13). However, based on both student and parent reports, students with access to the charter schools spent significantly more time on homework each night, suggesting a different conclusion. Because such alternative measures might suffer from other forms of biases (12), including a combination of measures from various sources might help mitigate these types of risks.

Materials and Methods

Sample. The data come from 2 cohorts of students from 20 schools in Shanghai.[‡] The experiments were embedded in a larger data collection effort that aims to study the education of migrant children in Shanghai. The school sample consists of 11 public schools that enroll a significant number of migrant students in Shanghai and 9 migrant schools that enroll only migrant students.[§] Each school had between 2 and 6 classes in each grade and between 45 and 328 students per grade.

The first cohort consisted of 2,950 students who were in fourth grade during the spring of 2016. The second cohort consisted of 2,793 students who were in fourth grade during the spring of 2017. Experiment #1 was conducted during the spring when students were in fifth grade, and Experiment #2 was conducted during the spring when students were in fourth grade. A total of 3,995 students were included in Experiment #1, and 5,216 were included in Experiment #2[¶] (*SI Appendix*, Tables S7 and S8 show descriptive statistics on the students included in each experiment).

Experimental Design. Both experiments followed a similar procedure that was designed to maximize statistical power while minimizing the possibility of contamination. The experiments were embedded in a larger data collection effort that included a math test and a self-reported BFI skills questionnaire. Prior to administering the math test and questionnaire, individual students were randomly assigned to a treatment group or a control group. Assignment was conducted separately within each school. To minimize contamination effects, students were notified to go to the assigned classrooms just before the survey. Teachers were not aware of the plan to conduct randomization or the outcome of which students were assigned to which treatment group.

We also asked the head teachers of each homeroom class to rate the Big Five of each student in their classroom. Hence, each teacher assessed students from both treatment and control groups. Teachers were unlikely to know whether the students were assigned to the treatment group and did not have incentives to assess students differently based on student assignment. Teachers' assessments were submitted within 1 wk of when the students took the survey.

Measures of Noncognitive Skills. To measure noncognitive skills, we used the BFI, a commonly used 44-item inventory of the Big Five (21). The questionnaire was translated to Mandarin and was adapted to be appropriate for fourth- and fifth-grade students. To reduce measurement error in the Big Five constructs, we estimated a factor model separately for each construct. We used this model to estimate Bartlett factor scores for each student, which can be interpreted as averages of the underlying items, weighted by the extent to which the items contain information on the latent factors so that items with more measurement error receive lower weights (32, 33). We converted the scores into SD units by dividing the scores by the SD in the control group for each year.

Statistical Analyses. To estimate the effects of both experiments, we used ordinary least squares to estimate the following equation:

$$Y_{is} = \alpha + \beta T_{is} + \gamma X_{is} + \varepsilon_{is},$$

where Y_{is} is the outcome for individual i in school s, T_{is} is an indicator for treatment status (a vector in the case of Experiment #2), X_{is} is a vector of baseline covariates (including school dummy variables and baseline measures of the outcome for our main specification), and ε_{is} is an error term. Reported SEs are corrected for heteroscedasticity with the Huber–White sandwich estimator.

Additional Covariates. To increase the precision of the estimates, our main specification included school fixed effects and a parsimonious set of covariates that capture students' background and skill (34). These covariates included gender; cohort; whether the student had a Shanghai hukou[#]; migrant school status; teacher ratings of each student's overall performance, punctuality, and discipline; whether the student was elected to be a leader of the class; and baseline measures of the Big Five (*SI Appendix*, Table S10 shows definitions of these covariates). For Experiment #1, we used students' self-report of the Big Five collected during the prior year as a baseline measure of the Big Five. For Experiment #2, we did not have access to a baseline measure of students' self-report of the Big Five but did have teacher

[‡]The data and documentation for these experiments are available via http://iar.sufe.edu.cn/6717/list.htm and are provided by the Institute for Advanced Research (IAR) of Shanghai University of Finance and Economics (SUFE) as part of the Longitudinal Study of Shanghai Migrant Children. The study was approved by the institutional review board (IRB) of the IAR of SUFE. All participants gave informed consent in accordance with policies of the IRB at the IAR of SUFE.

[§]More information on migrant schools is in refs. 29–31.

[¶]The sample size decreased between the 2 experiments because many migrant students in our sample returned to their hometowns between grades 4 and 5.

[#]The hukou system has evolved over time in China, especially since the 1980s. A hukou has 2 dimensions. One is rural vs. urban, and the other is locality (e.g., Beijing vs. Shanghai). Over time, the rural vs. urban dimension has weakened. In many provinces, the government has abolished the rural or urban distinction and has given all permanent residents the same hukou. However, the locality dimension has strengthened due to increased regional inequality, and it resembles "citizenship" because it allows people to enjoy "fully" welfare rights, including children's access to public schools.

reports of each student's Big Five, which we used as an alternative baseline measure. We also had measures of family background, including parental education and household income. Because the family background measures were missing at higher rates, we did not include them in our main specification but did include them in sensitivity analyses. We conducted sensitivity analyses using various specifications, including ones with no covariates; basic demographics; and basic demographics, baseline Big Five measures, and school behaviors (with and without school fixed effects and measures of family background). The results were generally stable across specifications (*SI Appendix*, Tables S11 and S12).

Because the data on outcomes and covariates were missing infrequently (*SI Appendix*, Tables S13 and S14), we removed from our main analysis any observations that had missing data on outcomes or covariates. To help ensure that this approach did not bias our results by creating a selected sample, we conducted a sensitivity analysis in which we imputed missing covariates as the mean in the sample. The results were very similar with this imputation approach (*SI Appendix*, Tables S15 and S16).

ACKNOWLEDGMENTS. We received valuable comments from 2 anonymous referees and the editor; Carmit Segal; Charles Tilley; and seminar participants at the Institute on Behavior and Inequality at Bonn University (2019), the Institute for Economic and Social Research at Jinan University (2018), and the Measuring and Assessing Skills Conference at the University of Chicago (2017). We thank Jing Zhang for his excellent research assistance. Y.C. acknowledges funding support from National Natural Science Foundation of China Grant 71773074 and the Shanghai Soong Ching Ling Foundation (Lu Jiaxian and Gao Wenying Special Foundation). S.F. acknowledges funding support from National Natural Science Foundation of China for Distinguished Young Scholars Grant 71425005 and National Natural Science Foundation of China Grant 71773037. J.J.H. acknowledges support from The Chicago–Jinan Joint Initiative on Human Development at the Institute for Economic and Social Research, Jinan University; NIH Grants National Institute of Child Health and Human Development (NICHD) R37HD065072 and NICHD R01HD054702; and the American Bar Foundation. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders or the official views of the NIH.

1. J. J. Heckman, T. Kautz, "Achievement tests and the role of character in American life" in *The Myth of Achievement Tests: The GED and the Role of Character in American Life*, J. J. Heckman, J. E. Humphries, T. Kautz, Eds. (University of Chicago Press, Chicago, IL, 2014), pp. 3–56.
2. M. Almlund, A. Duckworth, J. J. Heckman, T. Kautz, "Personality psychology and economics" in *Handbook of the Economics of Education*, E. A. Hanushek, S. Machin, L. Woessmann, Eds. (Elsevier, Amsterdam, Netherlands, 2011), pp. 1–181.
3. B. W. Roberts, N. R. Kuncel, R. Shiner, A. Caspi, L. R. Goldberg, The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect. Psychol. Sci.* **2**, 313–345 (2007).
4. J. Heckman, R. Pinto, P. Savelyev, Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *Am. Econ. Rev.* **103**, 2052–2086 (2013).
5. C. K. Jackson, What do test scores miss? The importance of teacher effects on non–Test score outcomes. *J. Polit. Econ.* **126**, 2072–2107 (2018).
6. T. Kautz, J. J. Heckman, R. Diris, B. ter Weel, L. Borghans, *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success* (Organisation for Economic Co-operation and Development, Paris, France, 2014).
7. B. W. Roberts, D. Wood, A. Caspi, "Personality development" in *Handbook of Personality: Theory and Research*, O. P. John, R. W. Robins, L. A. Pervin, Eds. (Guilford Press, New York, NY, 2008), pp. 375–398.
8. M. R. West, K. Buckley, S. B. Krachman, N. Bookman, Development and implementation of student social-emotional surveys in the CORE districts. *J. Appl. Dev. Psychol.* **55**, 119–129 (2018).
9. District of Columbia Public Schools, A capital commitment: Year 1 update (District of Columbia Public Schools, Washington, DC, 2018). https://dcps.dc.gov. Accessed 16 April 2019.
10. F. Gajardo, N. Grau, Competition among schools and educational quality: Tension between various objectives of educational policy. *Int. J. Educ. Dev.* **65**, 123–133 (2019).
11. O. Chernyshenko, M. Kankaraš, F. Drasgow, *Social and Emotional Skills for Student Success and Well-Being: Conceptual Framework for the OECD Study On Social and Emotional Skills* (OECD Publishing, Paris, 2018).
12. A. L. Duckworth, D. S. Yeager, Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educ. Res.* **44**, 237–251 (2015).
13. C. C. Tuttle *et al*., "KIPP middle schools: Impacts on achievement and other outcomes: Final report" (Rep. 06441.910, Mathematica Policy Research, Princeton, NJ, 2013).
14. D. L. Paulhus, "Measurement and control of response bias" in *Measures of Personality and Social Psychological Attitudes*, J. P. Robinson, P. R. Shaver, L. S. Wrightsman, Eds. (Academic Press, San Diego, CA, 1991), pp. 17–59.
15. P. C. Kyllonen, J. P. Bertling, "Innovative questionnaire assessment methods to increase cross-country comparability" in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, L. Rutkowski, M. von Davier, D. Rutkowski, Eds. (CRC Press, Boca Raton, FL, 2013), pp. 277–286.
16. J. J. Heckman, T. Kautz, Hard evidence on soft skills. *Labour Econ.* **19**, 451–464 (2012).
17. L. Borghans, H. Meijers, B. ter Weel, The role of noncognitive skills in explaining cognitive test scores. *Econ. Inq.* **46**, 2–12 (2008).
18. C. Segal, Working when no one is watching: Motivation, test scores, and economic success. *Manage. Sci.* **58**, 1438–1457 (2012).
19. P. T. Costa, R. R. McCrae, Four ways five factors are basic. *Pers. Individ. Dif.* **13**, 653–665 (1992).
20. American Psychological Association, *APA Dictionary of Psychology* (American Psychological Association, Washington, DC, ed. 1, 2007).
21. O. P. John, E. M. Donahue, R. L. Kentle, *The Big Five Inventory—Versions 4a and 54* (University of California, Berkeley, Institute of Personality and Social Research, Berkeley, CA, 1991).
22. A. Schirm, E. Stuart, A. McKie, *The Quantum Opportunity Program Demonstration: Final Impacts* (Mathematica Policy Research, Princeton, NJ, 2013).
23. A. Falk, Facing Yourself: A Note on Self-Image. IZA Discussion Paper No. 10606. https://www.iza.org/publications/dp/10606/facing-yourself-a-note-on-self-image. Accessed 8 November 2019.
24. D. Ariely, A. Bracha, S. Meier, Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Am. Econ. Rev.* **99**, 544–555 (2009).
25. J. A. Durlak, R. P. Weissberg, A. B. Dymnicki, R. D. Taylor, K. B. Schellinger, The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Dev.* **82**, 405–432 (2011).
26. C. S. Dweck, *Mindset: The New Psychology of Success* (Ballantine Books, New York, NY, 2007).
27. B. M. Galla *et al*., The Academic Diligence Task (ADT): Assessing individual differences in effort on tedious but important schoolwork. *Contemp. Educ. Psychol.* **39**, 314–325 (2014).
28. G. Charness, U. Gneezy, A. Imas, Experimental methods: Eliciting risk preferences. *J. Econ. Behav. Organ.* **87**, 43–51 (2013).
29. Y. Chen, S. Feng, Access to public schools and the education of migrant children in China. *China Econ. Rev.* **26**, 75–88 (2013).
30. Y. Chen, S. Feng, Quality of migrant schools in China: Evidence from a longitudinal study in Shanghai. *J. Popul. Econ.* **30**, 1007–1034 (2017).
31. Y. Chen, S. Feng, Y. Han, Research on the education of migrant children in China: A review of the literature. *Front. Econ. China* **14**, 168–202 (2019).
32. M. S. Bartlett, Methods of estimating mental factors. *Nature* **141**, 609–610 (1938).
33. M. S. Bartlett, The statistical conception of mental factors. *Br. J. Psychol.* **28**, 97–104 (1937).
34. P. Z. Schochet, Statistical power for random assignment evaluations of education programs. *J. Educ. Behav. Stat.* **33**, 62–87 (2008).