Comparing the reliability and predictive power of child, teacher, and guardian reports of noncognitive skills

Shuaizhang Feng[a,b], Yujie Han[a], James J. Heckman[c,1], and Tim Kautz[d]

[a]Institute for Economic and Social Research, Jinan University, Guangzhou 510632, China; [b]School of Economics, Jinan University, Guangzhou 510632, China; [c]Department of Economics, The University of Chicago, Chicago, IL 60637; and [d]Mathematica, Inc., Princeton, NJ 08540

Contributed by James J. Heckman; received August 9, 2021; accepted December 21, 2021; reviewed by Bart Golsteyn and Christopher Soto

Children's noncognitive or socioemotional skills (e.g., persistence and self-control) are typically measured using surveys in which either children rate their own skills or adults rate the skills of children. For many purposes—including program evaluation and monitoring school systems—ratings are often collected from multiple perspectives about a single child (e.g., from both the child and an adult). Collecting data from multiple perspectives is costly, and there is limited evidence on the benefits of this approach. Using a longitudinal survey, this study compares children's noncognitive skills as reported by themselves, their guardians, and their teachers. Although reports from all three types of respondents are correlated with each other, teacher reports have the highest internal consistency and are the most predictive of children's later cognitive outcomes and behavior in school. The teacher reports add predictive power beyond baseline measures of Intelligence Quotient (IQ) for most outcomes in schools. Measures collected from children and guardians add minimal predictive power beyond the teacher reports.

predictive power | psychological assessment | personality traits | respondent types | Big Five

Noncognitive or socioemotional skills—such as conscientiousness and emotional stability—predict success in life and can be shaped through interventions.* Historically, cognitive skills—like Intelligence Quotient (IQ) and academic achievement—have been favored in evaluations of social interventions and educational systems (1). However, noncognitive skills rival cognitive skills in predicting many life outcomes (e.g., educational attainment, health, earnings, and employment) (2, 3). Noncognitive skills can also be fostered through interventions and education (4–6).

Due to this evidence, policy makers and researchers have increasingly measured noncognitive skills, typically relying on surveys in which children rate their own skills or adults rate the skills of children (7). Oftentimes, data from multiple respondents are collected to triangulate responses. For example, some school districts—such as the District of Columbia Public Schools—administer surveys that assess children's noncognitive skills from the perspectives of children, teachers, and guardians (8). Similarly, the Organisation for Economic Co-operation and Development (OECD)'s Study on Social and Emotional Skills—which was designed to inform policy on noncognitive skill development—has collected data from all three respondent types from cities in nine countries (9). In addition, evaluations of educational programs and other social interventions often collect measures of noncognitive skills from multiple respondent types.† To collect such data, many survey instruments include forms designed for children, guardians, and teachers.‡ However, collecting data from multiple perspectives adds expense and burden, and the benefits are unclear.

Our study examines the benefits of collecting data from multiple types of respondents, focusing on the predictive power of the measures. Psychological measures are often validated by examining the correlation between one measure and another measure designed to capture a similar construct. However, the extent to which the measure predicts meaningful future outcomes is more consequential (15). This study compares the properties of child-, guardian-, and teacher-reported measures of noncognitive skills, focusing on their predictive power for cognitive outcomes and behavior in school.

Our study builds on the previous literature in four important ways. First, the predictive power of child, guardian, and teacher reports of noncognitive skills has rarely been examined in the same study [notable exceptions are Barbaranelli et al. (14) and Kankaraš et al. (16)]. Much of the research on this topic is based on two important meta-analyses that examine the predictive power of child-reported and adult-reported noncognitive skills across different studies (17, 18). However, most of the

Significance

Recent evidence has shown that noncognitive or socioemotional skills (e.g., persistence and self-control) are predictive of success in life and can be shaped through interventions. Accordingly, policy makers and researchers have increasingly measured children's noncognitive skills, typically relying on surveys in which children rate their own skills or adults rate the skills of children. Such ratings are often collected from multiple respondent types. This study demonstrates that, compared with child and guardian reports of noncognitive skills, teacher reports are more predictive of children's cognitive and behavioral outcomes in school. Child and guardian reports add minimal predictive power beyond teacher reports. These findings suggest that policy makers and researchers should prioritize teacher reports above those of children and guardians.

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

Reviewers: B.G., Universiteit Maastricht; C.S., Colby College.

The authors declare no competing interest.

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

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2113992119/-/DCSupplemental.

Published February 7, 2022.

*We use noncognitive for the sake of brevity but do not intend to draw a sharp contrast with other terms used to describe them, such as socioemotional skills, nonacademic skills, character, and personality.

†For example, the Social and Character Development Research Consortium (10) collects data from students, parents, and teachers, and Grissmer et al. (11) collect data from both teachers and parents.

‡For example, several measurement systems include versions for multiple respondent types, including the Social Skill Improvement System (SSIS) Social-Emotional Learning Brief Scales (12), the Social Emotional Assets and Resilience Scales (13), and the Big Five Questionnaire for Children (14).

individual studies in these meta-analyses do not include multiple respondent types, which could lead to incomparable estimates since the samples and surveys differ for different respondent types. Our study uses the same protocols to collect data from all three respondent types for the same sample of children.

Second, past research has mostly focused on comparing how different reports of noncognitive skills predict academic achievement, as measured by grades and achievement test scores. However, noncognitive skills predict a broader array of outcomes (2, 3, 19). In addition to measures of academic achievement, our study includes behavioral measures of performance in school, such as the extent to which children avoid fights and exhibit good discipline.

Third, past studies have not examined the incremental predictive power of using reports from multiple respondent types—for example, the extent to which children's self-reports add predictive power beyond the reports of teachers.[§] Previous meta-analyses and other studies suggest that adult reports are generally more predictive than children's reports (17, 18, 20). However, child reports might be valuable to collect if they add predictive power beyond that of adult reports. Our study finds that teacher reports of noncognitive skills are the most predictive of children's cognitive outcomes and behaviors in school, and child and guardian reports add little predictive power beyond that of teacher reports. For nearly all outcomes in schools, we also find that teacher reports add predictive power beyond baseline measures of IQ.

Fourth, past studies have not accounted for biases in self-reporting that could affect child reports. Child self-reports might be especially susceptible to reference bias, which arises when survey respondents have different benchmarks from which they assess themselves. For example, children who have peers with relatively high noncognitive skills might rate themselves relatively lower because they compare themselves with their peers. As a sensitivity analysis, our study uses anchoring vignettes, a survey method that adjusts the child reports to account for reference bias (21–25).

Our study uses data from a longitudinal survey of children, guardians, and teachers collected from primary school children in the Mianzhu region of China. We collect data on five noncognitive skills from the Big Five taxonomy, which are sometimes described as the "latitude and longitude" of personality (26). The Big Five include openness to experience (openness), which captures curiosity and intellectual pursuits; conscientiousness, which captures whether people are organized and hardworking; extraversion, which captures whether people are outgoing and sociable; agreeableness, which captures unselfishness and friendliness; and emotional stability (or neuroticism), which captures consistency in emotional reactions (27).

We examine four properties of these measures of noncognitive skills, as reported by children, guardians, and teachers. We estimate 1) the correlation among reports from each respondent type; 2) the internal consistency reliability of each measure, which suggests the extent to which reports suffer from measurement error; 3) the predictive power of measures of noncognitive skills for future cognitive and behavioral outcomes in school; and 4) the extent to which reports from each type of respondent add to the predictive power of reports from other respondents.

Although reports from all three types of respondents are correlated with each other, teachers' reports have the highest internal consistency and are the most predictive of children's later behavior and performance in school. Measures collected from children and guardians add little predictive power beyond those from the teacher reports.

Results

Correlation between Different Respondent Types. The reports of noncognitive skills from children, guardians, and teachers are only moderately correlated with each other, suggesting that the various respondent types have different views on child noncognitive skills (Table 1). The average correlation coefficient is 0.18 between teachers and children, 0.16 between children and guardians, and 0.15 between teachers and guardians. Across all of the skills, extraversion has the highest average correlation, which is consistent with the findings by Laidra et al. (28). For most noncognitive skills (except for agreeableness and emotional stability), older children have higher agreement with guardians and teachers than younger children, potentially because younger children have more difficulty understanding the questions and their self-reports could be less accurate (*SI Appendix*, Table S1).

We found that the levels of interrater agreement are similar to those found in other studies that focus on children or adolescents (14, 28) but are lower than those for adult targets (29). The levels of interrater agreement for child targets may be lower for two possible reasons. First, children, guardians, and teachers likely rely on different information when describing child noncognitive skills. Children may base their evaluation on some experiences that are not observed by their guardians and teachers. Meanwhile, guardians and teachers interact with children in different situations (home vs. school) and may, therefore, evaluate children from different perspectives. Manifest traits depend on situations (30). Second, raters for adult studies are more likely to be adults, while raters for child or adolescent studies may include both children and adults. Children's assessments might be less accurate than those of adults. Although the magnitudes of the correlations are generally comparable with those found in past studies (14, 28), past studies have also found that the correlations between teacher and child reports are lower than those for other assessments, differing from what we find. The difference might arise because 29% of the guardians in our sample are not parents (e.g., grandparents, aunts, or other relatives) and may be less familiar with the children. Compared with reports from nonparental guardians, the reports from parental guardians are more correlated with child and teacher reports (*SI Appendix*, Table S2). The average inter-trait correlation coefficients are 0.39, 0.41, and 0.51 for guardians, children, and teachers, respectively (*SI Appendix*, Table S3). The results suggest that reports from guardians have a slightly higher level of differentiation among the noncognitive skills compared with reports from children. However, the teacher reports have more concordance across skills, possibly due to halo effects and less measurement error in teacher reports.

Internal Consistency for Each Respondent Type. Overall, teacher reports have the highest internal consistency reliability, suggesting that they are measured with the least amount of error (Table 2). Cronbach's alpha—a standard measure of internal consistency—is uniformly better for the teacher reports than the child and guardian reports.[¶] Apart from emotional stability,

[§] Some studies have examined the predictive power of measures that combine reports from multiple respondent types. For example, OECD's Study on Social and Emotional Skills examined the predictive power of a "triangulated" measure based on reports from children, teachers, and guardians (16).

[¶] The teacher reports tend to exhibit the highest test–retest reliability between successive years, especially when the same teacher gave reports in successive years (*SI Appendix*, Table S4). However, given the relatively long period between measurements, the test–retest reliability could differ because of both differences in measurement error and growth and development of skills. For that reason, this finding may not reflect a more favorable property of the teacher reports.

Table 1. Correlations between child, guardian, and teacher reports of noncognitive skills

Pair of respondents	Extraversion	Agreeableness	Conscientiousness	Emotional stability	Openness
Children and teachers	0.32	0.11	0.20	0.06	0.23
Children and guardians	0.30	0.09	0.16	0.07	0.16
Teachers and guardians	0.30	0.04	0.15	0.08	0.16

The number of observations is 4,292 for the whole sample. All coefficients are statistically significant at the 1% level.

the child reports are the lowest.[#] Joint tests of the equality of the estimates between each pair of respondent types indicate that the differences are statistically significant at the 1% level.[‖]

Predictive Power of Individual Measures. For each of the individual noncognitive skills, teacher reports are the most predictive of nearly all child cognitive outcomes measured 1 y later (Table 3). (The results are similar for test scores measured 1.5 y later.) Teacher reports—especially of openness—are the most correlated with future cognitive outcomes, followed by child reports and guardian reports.[**] Measures of noncognitive skills better predict test scores on knowledge of Chinese language compared with other subjects.

Teacher reports are also the most predictive of behavioral outcomes in school, followed by child self-reports and then guardian reports (Table 4).[††] In most cases, noncognitive skills and behavioral outcomes in school are positively related; a higher level of the skill is associated with a more favorable outcome. There are, however, a few exceptions. For example, child and guardian reports of extraversion are negatively associated with avoiding fights, which might arise because lower levels of extraversion could lead to fewer interactions with other people in general, including fights. These findings are consistent with other evidence that higher levels of noncognitive skills do not always lead to better outcomes in all domains (2).

Table 2. Cronbach's alpha of child, teacher, and guardian reports of noncognitive skills

	Cronbach's alpha		
	Child	Teacher	Guardian
Noncognitive skill			
Extraversion	0.57	**0.82**	0.65
Agreeableness	0.63	**0.88**	0.73
Conscientiousness	0.57	**0.83**	0.61
Emotional stability	0.62	**0.75**	0.51
Openness	0.62	**0.82**	0.67
p-value from joint test of equality between respondent types			
Child vs. guardian		<0.01	
Child vs. teacher		<0.01	
Teacher vs. guardian		<0.01	

Bold indicates the respondent type with the highest Cronbach's alpha for each noncognitive skill. The number of observations is 5,422 to 5,465 for child reports, 5,570 for teacher reports, and 4,846 to 4,870 for guardian reports. The *p*-values are based on *F* tests, with the null hypothesis that the Cronbach's alpha estimates from each pair of respondent types are jointly equal and are obtained by bootstrapping with 1,000 replications.

Statistical tests reveal that the correlations between outcomes in school and teacher reports of noncognitive skills differ from those correlations based on the child and guardian reports. For each outcome, joint tests of equality of the correlation coefficients indicate significant differences between child and teacher reports as well as between guardian and teacher reports. However, some differences in the correlation coefficients between child and guardian reports—such as for science and math achievement—are not significantly different from each other.[‡‡]

Predictive Power of Groups of Measures. Consistent with the predictive power for individual measures, using the group of teacher reports of all five noncognitive measures in a multivariate regression better predicts cognitive outcomes than using the analagous groups of reports by children or guardians (Fig. 1). Similarly, the group of child reports tends to be more predictive than that of guardian reports.[§§] Unlike the finding that most correlation coefficients are of similar size, the pattern of regression coefficients suggests that conscientiousness and openness are the most predictive of cognitive outcomes when analyzing the group of noncognitive skills (*SI Appendix*, Table S10). Adding the child and guardian reports to the teacher reports makes a negligible difference in the predictive power (compare the Teacher and All respondents bars in Fig. 1), suggesting that the teacher reports capture the relevant information for predicting cognitive outcomes.[¶¶] Except when predicting future IQ, the teacher reports rival the predictive power of baseline IQ and add predictive power beyond baseline IQ when both are included as predictors (*SI Appendix*, Table S11). In contrast, the child and guardian reports are substantially less predictive than baseline IQ and add minimally to the predictive power of baseline IQ.

The results for behavioral outcomes in school display a similar pattern (Fig. 2). The group of teacher reports is the most predictive for each behavioral outcome in school. The group of child reports is more predictive than the group of guardian reports for four of five outcomes. Conscientiousness and openness are the most important measures for predicting most behavioral outcomes in school (*SI Appendix*, Table S10). In addition, adding the child and guardian reports to the teacher reports increases the predictive power minimally (compare the Teacher and All respondents bars in Fig. 2). Using the separate reports from all three respondent types in a multivariate regression is more predictive of the behavioral outcomes in school than using baseline IQ (*SI Appendix*, Table S11). Compared with child and guardian reports, teacher reports add more predictive power beyond IQ.

[#] The reliability of parental guardian reports is higher than that of nonparental guardian reports, but it is still lower than that of teacher reports (*SI Appendix*, Table S5).

[‖] For each noncognitive skill, the difference in Cronbach's alphas between each pair of respondent types is also significant at the 1% level (*SI Appendix*, Table S6).

[**] For cognitive outcomes, the correlations are similar but tend to be slightly higher for parental compared with nonparental guardians (*SI Appendix*, Table S7).

[††] For behavioral outcomes in school, the correlations tend to be slightly higher for parental guardians compared with nonparental guardians (*SI Appendix*, Table S7).

[‡‡] Individual tests of the differences in pairwise correlations reveal a similar pattern. In most cases, the differences in correlations between child and teacher reports are statistically significant at the 1% level. The same is true for the differences in guardian and teacher reports. However, many of the differences in correlations between child and guardian reports are not statistically significant, including for several cognitive measures and mental health (*SI Appendix*, Table S8).

[§§] Parental and nonparental guardian reports are similarly predictive (*SI Appendix*, Table S9).

[¶¶] The results of adjusted R are similar when restricting the sample to cases where the teachers and guardian respondents were the same in both time periods (*SI Appendix*, Figs. S1 and S2).

Table 3. Correlations between individual noncognitive skills and child cognitive outcomes in school measured 1 y later

	IQ	Chinese test score	Math test score	English test score	Morality test score	Science test score	Overall academic performance
Child							
Extraversion	0.08***	0.20***	0.15***	0.15***	0.14***	0.05***	0.16***
Agreeableness	0.06***	0.18***	0.10***	0.16***	0.10***	0.04**	0.15***
Conscientiousness	0.02	0.18***	0.09***	0.17***	0.11***	0.04**	0.19***
Emotional stability	0.04**	0.11***	0.10***	0.10***	0.09***	0.04**	0.13***
Openness	0.09***	0.21***	0.15***	0.16***	0.12***	0.09***	0.17***
Teacher							
Extraversion	0.19***	0.34***	0.32***	0.25***	0.24***	0.19***	0.36***
Agreeableness	0.06***	0.24***	0.17***	0.15***	0.14***	0.08***	0.28***
Conscientiousness	0.17***	0.39***	0.29***	0.30***	0.25***	0.20***	0.47***
Emotional stability	0.15***	0.32***	0.27***	0.20***	0.19***	0.16***	0.34***
Openness	0.26***	0.43***	0.42***	0.32***	0.28***	0.28***	0.50***
Guardian							
Extraversion	0.08***	0.09***	0.09***	0.09***	0.08***	0.03	0.09***
Agreeableness	0.07***	0.07***	0.06***	0.05***	0.05**	0.00	0.05**
Conscientiousness	0.05**	0.09***	0.07***	0.12***	0.07***	0.07***	0.12***
Emotional stability	0.05**	0.08***	0.06***	0.05**	0.06***	0.05***	0.06***
Openness	0.13***	0.14***	0.13***	0.09***	0.10***	0.11***	0.14***
p-value from joint test of equality between respondent types							
Child vs. guardian	0.70	<0.01	0.06	<0.01	0.05	0.11	<0.01
Child vs. teacher	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Teacher vs. guardian	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

The number of observations is 2,857 for the whole sample. IQ is measured using a 60-item version of Raven's Standard Progressive Matrices. Academic test scores are normalized by grade level. The morality test assesses whether students can analyze objectively and make reasonable judgements about different issues, including caring about others, confidence, perseverance, respect for others, responsibility, and integrity. Overall academic performance is evaluated on a one- to five-point scale, with one corresponding to "very bad" and five corresponding to "very good." The *p*-values are based on *F* tests with the null hypothesis that the correlations from each pair of respondent types are jointly equal for each outcome and are obtained by bootstrapping with 1,000 replications. **Significant at the 5% level; ***significant at the 1% level.

Adjusting Child Self-Reports for Reference Bias. Our main conclusions do not change when we use anchoring vignettes to adjust child self-reports for reference bias. Reference bias could arise if children have different reference points when responding to survey questions because they interpret the survey response options differently. The survey questions ask children the extent to which they agree that they exhibit behaviors related to each noncognitive skill, with response options ranging from "totally disagree" to "totally agree." The extent to which children agree could depend on their own reference point for that behavior.

To account for reference bias in children's reports, we included anchoring vignettes, which provide a way to adjust self-reported measures by "anchoring" the items of interest using the respondents' assessment of the behavior of hypothetical children (21–25). The instructions ask children to rate the behaviors of three hypothetical children in terms of each of the Big Five using the same response categories that we used to assess the Big Five. For each skill in the Big Five, the three vignettes exemplify people with varying degrees (low, medium, and high) of that skill and serve as measures of children's reference points.

Adjusting child reports using vignettes tends to increase the predictive power of child reports, but the estimates are sensitive to the method for making the adjustment (*SI Appendix*, Tables S12 and S13). In particular, results depend heavily on how we handle "inconsistent" responses to the anchoring vignettes, which arise when children rate the vignettes in the incorrect order—for example, by rating the "low" vignette higher than the "medium" vignette. For some approaches to handling inconsistent responses, the predictive power goes down after adjustment. Other scholars have also found that using anchoring vignettes reduces predictive power (31). Supplemental analyses reveal that whether children rated the vignettes inconsistently is predictive in and of itself and that the anchoring adjustment has little direct effect per se (*SI Appendix*, Table S13). Adjusting for reference bias also increases the Cronbach's alpha by between 0.07 and 0.23 (*SI Appendix*, Table S14). However, as noted by von Davier et al. (32), using anchoring vignettes may artificially improve reliability, so the increase does not necessarily reflect a lower level of measurement error. The correlations between the unadjusted child self-reports and adult reports are always higher than those when we use adjusted reports, suggesting there is no clear evidence that the difference in measuring noncognitive skills between children and adults is due to how children interpret the anchoring scale (*SI Appendix*, Table S15).

Discussion

Our results demonstrate that teacher reports of children's noncognitive skills have more favorable psychometric properties compared with child and guardian reports. Importantly, the teacher reports are the most predictive of children's cognitive and behavioral outcomes in school, and adding the child and guardian reports yields little incremental predictive power. Teacher reports might be more predictive for at least three reasons. First, teachers have experienced a greater variety of children, so they are more empirically grounded and can more accurately compare skills across children. Second, it is also possible that halo effects come into play if teachers assign higher noncognitive skill ratings to children who they generally view more positively—for example, based on their observations of children's academic performance (33). Third, children may have noncognitive skills specific to schools that might better predict outcomes in school, and teachers' ratings of children's noncognitive skills are primarily based on their observation of children within the school context. However, we cannot distinguish among these reasons with our data, which provides important directions for future research. We suggest that future

Table 4. Correlations between individual noncognitive skills and child behavioral outcomes in school measured 1 y later

	Leader	Mental health	Avoids fights	Honesty	Good discipline
Child					
Extraversion	0.17***	0.12***	−0.01	0.03*	0.12***
Agreeableness	0.15***	0.12***	0.10***	0.12***	0.16***
Conscientiousness	0.17***	0.13***	0.09***	0.13***	0.19***
Emotional stability	0.13***	0.12***	0.02	0.04**	0.11***
Openness	0.19***	0.12***	−0.05***	0.03	0.11***
Teacher					
Extraversion	0.36***	0.27***	−0.04**	0.12***	0.27***
Agreeableness	0.18***	0.29***	0.24***	0.31***	0.29***
Conscientiousness	0.34***	0.38***	0.29***	0.37***	0.46***
Emotional stability	0.26***	0.25***	0.19***	0.24***	0.32***
Openness	0.36***	0.32***	0.11***	0.27***	0.37***
Guardian					
Extraversion	0.12***	0.08***	−0.08***	−0.01	0.04**
Agreeableness	0.06***	0.05***	0.06***	0.05***	0.05***
Conscientiousness	0.10***	0.10***	0.13***	0.13***	0.13***
Emotional stability	0.07***	0.04**	0.03	0.03	0.07***
Openness	0.12***	0.09***	−0.02	0.05**	0.08***
p-value from joint test of equality between respondent types					
Child vs. guardian	<0.01	0.05	<0.01	0.01	<0.01
Child vs. teacher	<0.01	<0.01	<0.01	<0.01	<0.01
Teacher vs. guardian	<0.01	<0.01	<0.01	<0.01	<0.01

The number of observations is 2,857 for the whole sample. Leader is a dummy for being elected as a class or school leader. Mental health and good discipline are evaluated on a one- to five-point scale, with one corresponding to "very bad" and five corresponding to "very good." Avoiding fights and honesty are evaluated on a one- to five-point scale, with one corresponding to "never" and five corresponding to "often." The *p*-values are based on *F* tests with the null hypothesis that the correlations from each pair of respondent types are jointly equal for each outcome and are obtained by bootstrapping with 1,000 replications. *Significant at the 10% level; **significant at the 5% level; ***significant at the 1% level.

studies address this issue by examining outcomes inside and outside of school.

Our analyses of anchoring vignettes suggest that child reports of noncognitive skills do not lack predictive power compared with teacher reports because they suffer from reference bias. However, the analyses do suggest that whether children respond to anchoring vignettes consistently is predictive of their later outcomes in school, especially those related to cognitive outcomes. This finding might arise because consistently responding to anchoring vignettes requires cognitive abilities—such as reading comprehension and reasoning—and higher levels of conscientiousness that are also predictive of cognitive outcomes.

One potential limitation of this study is that the behavioral outcomes in school are based on teacher reports, which are subjective in some cases. For this reason, the teacher reports of noncognitive skills might be more predictive in part because they capture a common rater effect that also applies to the outcomes in school. However, it is unlikely that such effects drive the results because the patterns are similar for more objective outcomes in school, including whether a child has been elected to be a leader within their school and their scores on cognitive tests, which come from other records. To further investigate this issue, we conduct analyses separately for samples of children for whom 1) the same teacher reported the noncognitive skills and outcomes in school and 2) different teachers reported the noncognitive skills and outcomes in school. The predictive power of teacher reports of noncognitive measures for outcomes in school tends to be somewhat higher when the same teacher reports both, providing some evidence for a teacher rater effect. However, the teacher reports are still more predictive than the child and guardian reports when two different teachers report on child noncognitive skills and the outcomes in school, suggesting that our main conclusions are robust to teacher rater effects (*SI Appendix*, Tables S16 and S17).

Overall, this study suggests that teacher reports are relatively more accurate when collecting measures of noncognitive skills, especially if the goal is to identify children who are at risk for poor outcomes in school. Adding elicitations from children and guardians does not improve predictive power. By focusing on teacher reports, researchers and policy makers may be able to reduce both costs and burden on respondents while maintaining the ability to predict children's outcomes in school. In addition, the findings suggest that previous studies, such as those of Noftle and Robins (34) and Connelly and Ones (29), that use reports of noncognitive skills from other respondent types to predict school-related outcomes may underestimate their predictive power.

Materials and Methods

Sample. This study uses longitudinal survey data on children from 16 primary schools in Mianzhu, a county located in Sichuan Province in China. Three of the schools are in urban areas, and 13 are in rural areas. The survey—which was designed to capture the development of children—followed children for 2 y and was administered in two waves. We conducted the first wave of the survey in October 2017, interviewing all children in grades 4 through 6 in 14 of the schools. Because the two other schools were larger, we randomly selected half of the classes in those schools to interview. We conducted a follow-up survey of the children in grades 4 and 5 in November 2018. By the follow-up survey, the children were in the first term of grades 5 and 6. In total, the first wave consists of 5,606 children in 138 classes (*SI Appendix*, Table S18 has a description of the variables used in the analysis, and *SI Appendix*, Table S19 has descriptive statistics on the children). In the second wave, we followed 3,707 of the children. The sample includes 4,292 children from the first wave who have valid information on the Big Five measures from all three respondent types.

Measures of Noncognitive Skills. Children's noncognitive skills were measured using the Big Five, as elicited through reports from children themselves, their head teachers, and guardians. In particular, children completed the 60-item Big Five Inventory-2 (BFI-2) (35), an updated version of the widely used Big Five Inventory (36). The items—originally written at a fifth-grade reading level—elicit the extent to which respondents agree to a series of statements. We translated the questionnaire to Mandarin and adapted it to be appropriate for children in grade 4, the lowest grade in our sample. To ensure that the survey items perform well in our sample, we pretested the survey with 469

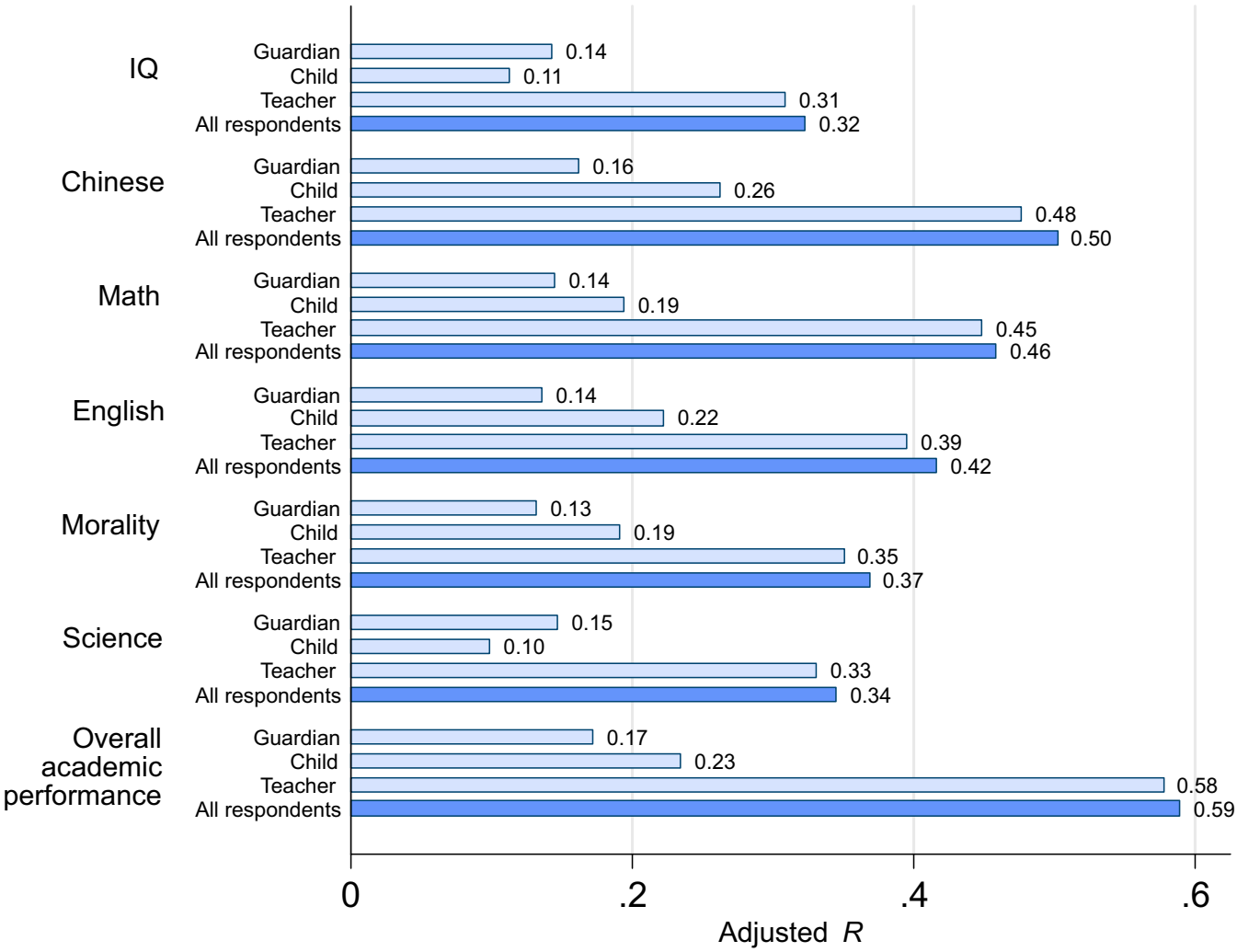

Fig. 1. Predictive power (adjusted *R*) of groups of noncognitive skills for cognitive outcomes 1 y later. Shown is the square root of the adjusted R^2 from OLS regressions of cognitive outcomes on the group of Big Five noncognitive skills for each respondent type. IQ is measured using a 60-item version of Raven's Standard Progressive Matrices. Academic test scores are normalized by grade level. The morality test assesses whether students can analyze objectively and make reasonable judgments about different issues, including caring about others, confidence, perseverance, respect for others, responsibility, and integrity. Overall academic performance is evaluated on a one- to five-point scale, with one corresponding to "very bad" and five corresponding to "very good." The full sample includes 2,857 observations.

students in grades 4 to 6 from Mianzhu and made some additional modifications based on feedback. To minimize the burden on teachers and guardians, we use an abbreviated survey that included four items from the BFI-2 for each of the five dimensions. For all respondents, the response options ranged from one to five, where one corresponds to totally disagree and five corresponds to totally agree. In this study, we focus on the 20 items common to the child, teacher, and guardian surveys (*SI Appendix*, Table S20).

To reduce measurement error in the noncognitive skill measures, we use factor scores based on factor models that were estimated separately for each group of Big Five items and respondent type. Using these models, we calculate Bartlett factor scores for each child, which are averages of the underlying items, weighted so that items that are measured with more error receive lower weights (37, 38) (*SI Appendix*, Table S20 shows the percentage of variance in each item explained by the noncognitive skill factors for each respondent type).

To account for potential reference bias in the child self-reports, we complement the standard measures of the Big Five with anchoring vignettes. Following previous studies (21, 24, 25), we adopt a nonparametric approach for the adjustment. When children rate the vignettes consistently in the defined order, the responses of the five-point Likert scales are extended to a seven-point Likert scale by relating the self-report response to the corresponding vignettes (*SI Appendix*, Table S21). To handle cases where children report the same response for two vignettes (tie) or rate the vignettes out of order (order violation), we follow previous studies and assign the lowest possible score to the related items (25), the highest possible value, or the average of the lowest and highest possible values.

Measures of Cognitive Skills. IQ was measured in each wave using a 60-item version of Raven's Standard Progressive Matrices test (39). The test was administered by paper and pencil at the same time as the survey, and children had 40 min to complete it. In each class, the test was proctored by the head teacher of the class and three study interviewers.

The academic test scores in Chinese, math, English, morality, and science were collected from school administrative records during each year of the study. The tests are designed by the Education Bureau of Mianzhu to incorporate the curriculum requirements of each subject and are identical for all students in the same grade in each year across schools. They were administrated at the same time in each school and were proctored by teachers. All tests in Mianzhu were graded by the same group of reviewers, ensuring that the scores are comparable across the schools in our sample.

Statistical Analyses. To estimate the predictive power of groups of noncognitive skills reported by different respondents, we used ordinary least squares to estimate the following equation:

$$Y_i = \alpha + \beta\hat{\theta}_i + \varepsilon_i,$$

where Y_i is the outcome in school for individual i, $\hat{\theta}_i$ is a vector of factor scores for the noncognitive skills, and ε_i is an error term. We reported the square

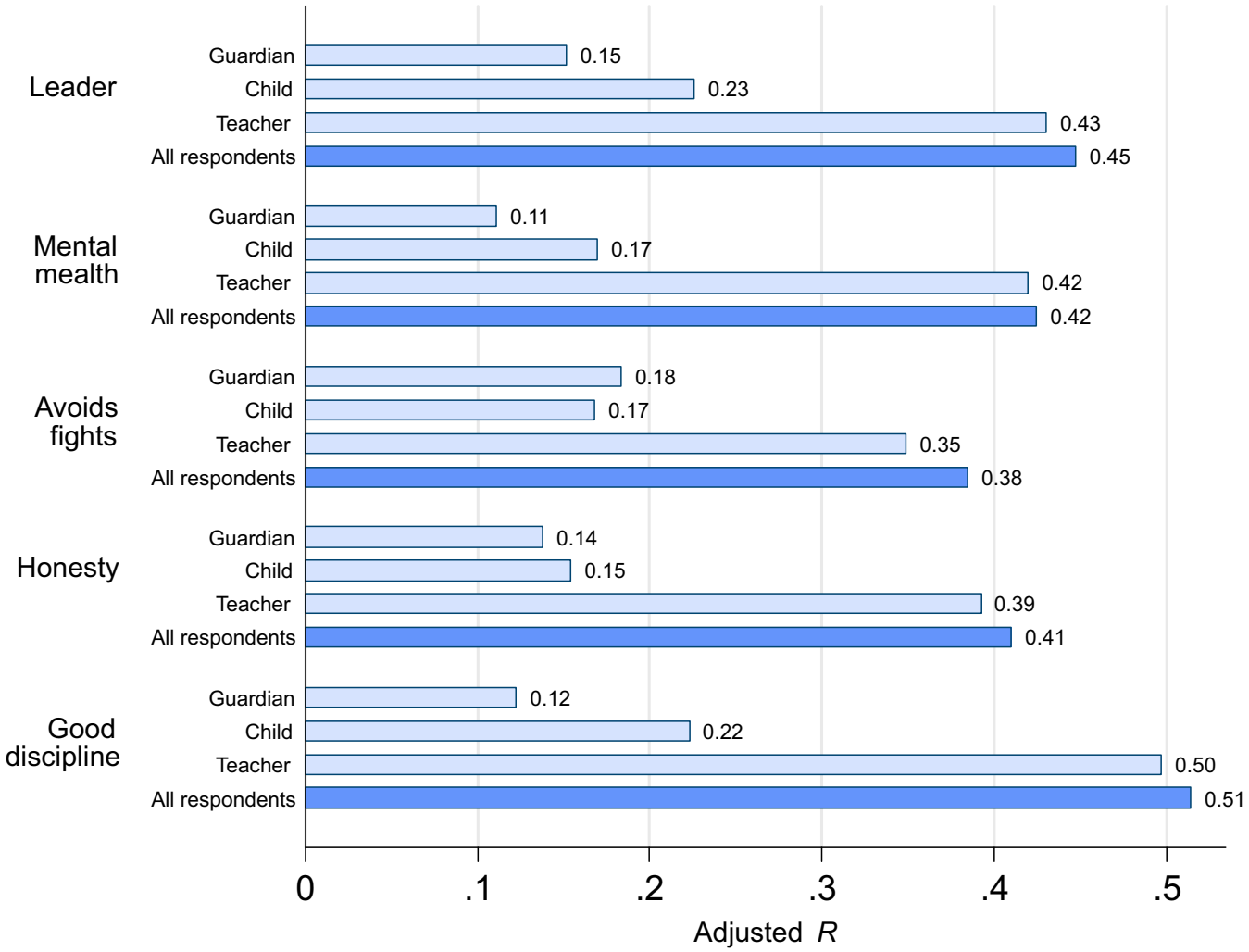

Fig. 2. Predictive power (adjusted *R*) of groups of noncognitive skills for behavioral outcomes in school measured 1 y later. Shown is the square root of the adjusted R^2 from the OLS regressions of behavioral measures in school on the group of Big Five noncognitive skills for each respondent type. Leader is a dummy for being elected as a class or school leader. Mental health and good discipline are evaluated on a one- to five-point scale, with one corresponding to "very bad" and five corresponding to "very good." Avoiding fights and honesty are evaluated on a one- to five-point scale, with one corresponding to "never" and five corresponding to "often." The full sample includes 2,857 observations.

root of the adjusted R^2 statistic because it is analogous to the univariate correlations.

Data Availability. The data and documentation for the paper are provided by the Survey Data Center of Jinan University as part of the Longitudinal Study of Children's Development in Mianzhu (https://sdc-iesr.jnu.edu.cn/2022/0107/c15992a676539/page.htm). The study was approved by the social science institutional review board (IRB) of Jinan University (JNU). All participants gave informed consent in accordance with the policies of the IRB of JNU.

ACKNOWLEDGMENTS. S.F. acknowledges funding support from National Natural Science Foundation of China Grants 72073052 and 71773037. J.J.H. acknowledges funding support from Eunice Kennedy Shriver National Institute of Child Health and Human Development of NIH Award R37HD065072.

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. L. 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, the Netherlands, 2011).
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* (OECD, Paris, France, 2014).
7. A. L. Duckworth, D. S. Yeager, Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educ. Res.* **44**, 237–251 (2015).
8. District of Columbia Public Schools, *DCPS 2019 Panorama Survey Results* (District of Columbia Public Schools, Washington, DC, 2018).
9. M. Kankaraš, J. Suarez-Alvarez, Assessment framework of the OECD Study on Social and Emotional Skills. 10.1787/5007adef-en. Accessed 24 January 2022.
10. Social and Character Development Research Consortium, *Efficacy of Schoolwide Programs to Promote Social and Character Development and Reduce Problem Behavior in Elementary School Children (NCER 2011-2001)* (National Center for Education Research, Institute of Education Sciences, US Department of Education, Washington, DC, 2010).

11. D. Grissmer *et al.*, *Final Report: The Evaluation of the WINGS After-School Social-Emotional Program for At-Risk Urban Children* (Social Innovation Fund, Corporation for National and Community Service, Washington, DC, 2019).
12. S. N. Elliott, C. A. Anthony, J. C. DiPerna, P. W. Lei, F. M. Gresham, *SSIS SEL Brief + Mental Health Scales: Expanded Guide & Technical Manual* (SAIL CoLab, Scottsdale, AZ, 2020).
13. R. N. T. Nese *et al.*, Social emotional assets and resilience scales: Development of a strength-based short-form behavior rating scale system. *J. Educ. Res. Online* **4**, 124–139 (2012).
14. C. Barbaranelli, G. V. Caprara, A. Rabasca, C. Pastorelli, A questionnaire for measuring the Big Five in late childhood. *Pers. Individ. Dif.* **34**, 645–664 (2003).
15. J. J. Heckman, T. Kautz, Hard evidence on soft skills. *Labour Econ.* **19**, 451–464 (2012).
16. M. Kankaraś, E. Feron, R. Renbarger, Assessing students' social and emotional skills through triangulation of assessment methods. 10.1787/717ad7f2-en. Accessed 24 January 2022.
17. A. E. Poropat, A meta-analysis of the five-factor model of personality and academic performance. *Psychol. Bull.* **135**, 322–338 (2009).
18. A. E. Poropat, A meta-analysis of adult-rated child personality and academic performance in primary education. *Br. J. Educ. Psychol.* **84**, 239–252 (2014).
19. M. R. West *et al.*, Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educ. Eval. Policy Anal.* **38**, 148–170 (2016).
20. B. M. Galla *et al.*, Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *Am. Educ. Res. J.* **56**, 2077–2115 (2019).
21. G. King, C. J. L. Murray, J. A. Salomon, A. Tandon, Enhancing the validity and cross-cultural comparability of measurement in survey research. *Am. Polit. Sci. Rev.* **98**, 191–207 (2004).
22. 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, 2014), pp. 277–285.
23. R. Primi, C. Zanon, D. Santos, F. De Fruyt, O. John, Anchoring vignettes: Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *Eur. J. Psychol. Assess.* **32**, 39–51 (2016).
24. J. Wand, Credible comparisons using interpersonally incomparable data: Nonparametric scales with anchoring vignettes. *Am. J. Pol. Sci.* **57**, 249–262 (2013).
25. S. Weiss, R. D. Roberts, Using anchoring vignettes to adjust self-reported personality: A comparison between countries. *Front. Psychol.* **9**, 325 (2018).
26. P. T. Costa, R. R. Mccrae, Four ways five factors are basic. *Pers. Individ. Dif.* **13**, 653–665 (1992).
27. American Psychological Association, *APA Dictionary of Psychology* (American Psychological Association, Washington, DC, ed. 1, 2007).
28. K. Laidra, J. Allik, M. Harro, L. Merenäkk, J. Harro, Agreement among adolescents, parents, and teachers on adolescent personality. *Assessment* **13**, 187–196 (2006).
29. B. S. Connelly, D. S. Ones, An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychol. Bull.* **136**, 1092–1122 (2010).
30. Y. Chen, S. Feng, J. J. Heckman, T. Kautz, Sensitivity of self-reported noncognitive skills to survey administration conditions. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 931–935 (2020).
31. J. Coenen, B. H. H. Golsteyn, T. Stolp, D. Tempelaar, Personality traits and academic performance: Correcting self-assessed traits with vignettes. *PLoS One* **16**, e0248629 (2021).
32. M. von Davier, H. J. Shin, L. Khorramdel, L. Stankov, The effects of vignette scoring on reliability and validity of self-reports. *Appl. Psychol. Meas.* **42**, 291–306 (2018).
33. R. E. Nisbett, T. D. Wilson, The halo effect: Evidence for unconscious alteration of judgments. *J. Pers. Soc. Psychol.* **35**, 250–256 (1977).
34. E. E. Noftle, R. W. Robins, Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *J. Pers. Soc. Psychol.* **93**, 116–130 (2007).
35. C. J. Soto, O. P. John, The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *J. Pers. Soc. Psychol.* **113**, 117–143 (2017).
36. 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).
37. M. S. Bartlett, The statistical conception of mental factors. *Br. J. Psychol.* **28**, 97–104 (1937).
38. G. H. Thomson, Methods of estimating mental factors. *Nature* **141**, 246–246 (1938).
39. J. Raven, J. C. Raven, J. H. Court, "The standard progressive matrices" *in Manual for Raven's Progressive Matrices and Vocabulary Scales* (Oxford Psychologists Press, Oxford, UK, 2000).