THE IMPACT OF CLASS SIZE ON NONCOGNITIVE ABILITY: RE-ANALYSIS OF DEE AND WEST (2008)

A THESIS

SUBMITTED BY

JAEHEE CHOI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

IN

ECONOMICS

TUFTS UNIVERSITY

FEBRUARY 2011

ADVISER: DAVID GARMAN

ABSTRACT

Noncognitive ability is difficult to delineate because it encompasses virtually all skills and personality traits that cannot be measured in standardized tests. In this paper, I focus on a specific construct, student engagement to study the impact of class size on this noncognitive variable. I employ the identification method that Dee and West (2008) developed to analyze data sets from NELS:88/90 (National Education Longitudinal Study of 1988). I examine whether a lower student-teacher ratio leads to positive changes in 10th graders' noncognitive skills. Comparing the eighth grade estimation from replication of Dee and West, I find that the effects of class size are minimal in Grade 10. Moreover, I do not observe any heterogeneous class size effect across various demographic groups.

ACKNOWLEDGMENTS

Special thanks go to my adviser David Garman. I cannot even begin to describe how deeply I appreciate his wholehearted support throughout my thesis. I am grateful to the members of my thesis committee Professor Randall Akee and Professor Thomas Downes for their comments, time, and attention. I also offer my sincerest gratitude to Professor Bianconi who supported me in many ways during my training at Tufts. Finally, I thank my family and friends – especially M.K. Bae, Karen Witts, and Yu Zhao – who have always believed in me and encouraged me to keep going.

Contents

List of Tables

1 Introduction

The concept of "noncognitive ability" has garnered increased and renewed attention among economists in recent years. It is an axiom of modern life that certain personality traits such as self-discipline, perseverance, motivation, and dependability – the attributes that are commonly categorized into noncognitive skills – are relevant for success in the economic sphere. Though they are familiar and cliched, the importance of these traits has been overlooked in both economic literature and policy discussions.

Bowles and Gintis (1976) were the first to challenge the then-standard practice of relying solely on IQ as a determinant of academic and economic success. The basic tenet of their reasoning rested on the "unimportance of heritability of IQ" (2001) in explaining intergenerational economic immobility; they claimed that overestimation of the IQ effect was prevalent (1976) and that other components of ability, hence "noncognitive," mattered more. In a cross-sectional study, they showed that IQ was not a predictor of high school GPA, but that standardized scores and noncognitive variables – consistency, external motivation, punctuality, dependability, and persistence – made significant contributions.

Two decades later, recent research provides strong empirical evidence that corroborates their accounts, possibly due to the increased availability of rich and varied data sets. In particular, Heckman and his colleagues have been leading the research on incorporating noncognitive ability into the model of skill formation. In an insightful article (2001) on GED recipients, Heckman, Hsee and Rubinstein elucidated the connection of noncognitive ability to academic achievement and various other social behaviors. Numerous studies (Jacob 2002, Coleman and DeLeire 2003, Heckman, Stixrud and Urzua 2006, Segal, 2006, Eren 2008, Piatek and Pinger 2009) that followed established similar positions in demonstrating the importance of noncognitive ability.

Heckman and Masterov (2007) reported that noncognitive and cognitive abilities are determined early by parental environment and socioeconomic status. As such, skills diverge in children as early as kindergarten, and the gap widens as schooling continues and persists throughout lifetime. Unlike cognitive skills, however, noncognitive skills can be improved through early intervention. This plasticity is supported by neurodevelopment literature in which claims that the human prefrontal cortex, the region of the brain that deals with executive functions or self-control, does not fully develop until mid-20's whereas IQ stabilizes in early childhood (Heckman 2006).

One commonly cited success story of early childhood interventions that targeted disadvantaged students is that of Perry Preschool. This randomized experiment did not have lasting effects on IQ, but resulted in positive changes in motivation and other noncognitive skills. The treatment group was followed until their 40's demonstrating far better outcomes in achievement tests, schooling, employment and home ownership and was less likely to be imprisoned (Heckman and Masterov 2007). Other successful programs include the Chicago Child-Parent Center (CPC) and Abecedarian Project.

CPC was a half-day pre-school program for three- or four-year olds. It was designed as a non-experimental trial, but subsequent studies showed that it had similar benefits as the Perry program. The Abecedarian Project was another randomized experiment and that tracked infants from only a few months of age. In terms of the participant-to-teacher ratio and the duration per year, it was more intensive than the Perry and CPC. The program improved the children's IQ, a possible consequence of intervening during infancy.

These early interventions offer useful insights for shaping education policies: a comparable emphasis has to be place on fostering noncognitive skills rather than narrowly focusing on improving IQ and test scores. Because pupil-teacher ratio is the most basic measure in investment of human capital in schools, I study the effect of smaller classes on development of noncognitive ability.

Class size reduction (CSR) is perhaps the most popular educational policy initiative and is desired by many parents and teachers. The supporters maintain that smaller classes enable teachers to provide more attention and individualized instruction to students. A large body of evidence of positive effects of reduced pupil-teacher ratios comes from the celebrated STAR experiment (Krueger and Whitmore 2001, Schanzenbach 2006). Critics argue that cutting class sizes is expensive and past CSR programs have produced mixed results. As the Stanford education specialist Michael Kirst notes in a New York Times article on August 18, 1997, a 1989 Johns Hopkins study of 14 separate cases concluded that the effects were minimal. Other researchers also have questioned the external validity and reliability of the project STAR. Hanushek (1999) claims that more randomized trials are needed to fully determine the efficacy.

While the effectiveness measured in test scores remain inconclusive, very little has been done to understand the impact of class size on the development of noncognitive skills. Dee and West (2008) provided the first and only study on gains in noncognitive skills for 8th graders. In the first part of their study, they re-analyzed the follow-up data from the STAR experiment and obtain evidence that smaller class assignment in K-3 improved student effort and decreased non-participatory behavior in 4th grade. But the improvements in student initiative did not persist through the 8th grade.

In the second part of the paper, the authors examine another set of data – NELS: 88 (National Education Longitudinal Study of 1988) – by exploiting the matched pairs feature of the teacher survey: a pair teachers evaluated each student's engagement behaviors in the classroom. Dee (2005, 2007) took advantage of this aspect in the data structure and produced a series of research explicating whether sharing the same demographic characteristic such as gender, ethnicity, and race with teachers influence student engagement and achievement. Dee and West (2008) isolate the "noncognitive returns" to class size utilizing the same empirical strategy. Their analysis indicates that smaller classes in 8th grade boost student engagement.

In this paper, I replicate Dee and West (2008) and re-analyze their method using a different wave – the 10th grade survey from NELS: 88/90. I examine whether a lower student-teacher ratio leads to positive changes in 10th graders' noncognitive skills. It would be beneficial to revisit their analysis to test the external validity of their model considering that the problem is typically called into question in class size studies. Furthermore, this paper can help understand the link between smaller classes in the middle and upper grades and noncognitive ability. Relatively few studies have been devoted to examining of benefits of class size reduction in the later stages of K-12 education. Comparing the evaluation of the 8th and 10th grade would allow us to identify any systematic changes that take place in the class size effects when students transition from junior high to high schools.

The second scope of this paper is to explore whether smaller classes can produce more pronounced effects for various demographic groups. It has been well-documented that girls tend to be more resilient to family stress to the extent that they are more likely to attain positive outcomes when confronted with considerable risks to development (Skinner et al. 2009). Diagnosis rates of ADHD are higher among males from families with income below the poverty line, but the rate remain stable for girls coming from below and above the threshold (Pastor and Reuben 2008). These separate findings suggest that noncognitive development may be selectively diminished for a certain subset of population. It would be helpful to identify who are at most risk and provide them with targeted support at a developmentally critical period.

The paper proceeds in the following manner. The next section reviews previous research on the importance of noncognitive abilities and how to measure them. Section III describes how my samples are constructed from the NELS:88/90 and representativeness of the samples is addressed. Section IV presents the identification strategy. Section V provides findings on the effects of reduced class size and sensitivity analysis. Finally, Section V concludes by considering policy implications.

2 Background information

How do we translate the notions of self-discipline, perseverance, motivation, and dependability into a good measure that is accessible? According to the late psychologist Samuel Messick (1979), the notion of noncognitive ability captures the belief, attitudes, temperament, coping skills, locus of control, information-processing consistency, social sensitivity, etc. that estimate the chance of success in a given learning environment. In most psychology circles, however, the terminology is a foreign idea. The label is also misleading because cognitive and noncognitive abilities are not mutually exclusive concepts. The economists' account of noncognitive ability relies on the use of cognition or the exercise of perception, thought, and reason. It is best understood as a "catch-all" construct that consists of all the components of ability that cannot be measured by standardized test scores (Farkas 2003).

Because a variety of components fall into the noncognitive category, there is no single representation (See Appendix 2 in Duckworth et. al (2009). The authors provide a list of diverse terms and concepts that are generally categorized as noncognitive in economics, psychology and sociology literature.). Instead, a lot of methodological ambiguities surround the concept. There is a fair overlap with the psychological concept of executive function (EF) – cognitive processes involving self-control, regulation, and working memory required for decision makings and goal-directed behaviors. EF is comprised of skills such as "inhibitory control, planning, attentional flexibility, error correction and detection, and resistance to interference (Carlson 2005)."

The effects of executive functions on student performance are well-documented. In Wolfe and Johnson (1995), the authors examined personality as a predictor of the college performance of psychology students. The "Big 3" introduced in Tellegen's 1985 paper emerged as the second highest predictor after high school GPA.

A similar interepreation is made in Duckworth and Seligman (2005); self-discipline outdoes IQ in predicting high school students' academic achievement. In a year-long study, a student's self-control was evaluated at the beginning of the school year with psychology tests such as the Junior Impulsiveness Subscale (Eysenck, Easting, and Pearson 1984), the Brief Self-control Scale (Tangney et al. 2004), and the Kirby Monetary Choice Questionnaire. Not only was executive function a much stronger predictor of the grades students earned at the end of the school year than IQ, it was also the only predictor of academic improvement over the school year.

Waber et. al (2006) showed neurobehaviral characteristics, especially executive functioning has a strong predictive power on fifth-graders' performance on standardized English and math tests. They administered neuropsychological tests and teacher-evaluated behavioral questionnnaires to 91 fifth-graders from Boston public schools in low-income neighborhoods. The neurocognitive tests were selected to measure information processing skills and executive functions. Among these, Delis-Kaplan Executive Function System (D-KEFS) Color-Word Interference was designed to quantify executive functions.

The teacher-questionnaires – Behavioral Rating Inventory of Executive Function (BRIEF) and Behavioral Assesment System for Children (BASC) – asked teachers to evaluate the student behaviors in the classroom contex. In particular they were developed to capture behavioral regulation, metacognition and locus of control. The results show that executive functions explain more than 30% variations in both state-mandated English and math scores. Moreover one of the two teacher reports (BRIEF metacognition index) was especially sensitive to the test performance.

Obviously many psychological assessments on executive functions mentioned earlier are not available in large-scale data sets that are oftentimes studied in economics literature. EF is often assessed using neuropsychology tasks and it would be expensive to administer those tests when sample size is substantially larger than that of psychology studies. Another issue has been raised concerning the EF measures. Some doubt has been cast upon the nature of the laboratory setting because the examiner puts external control to the participants (Gioia and Isquith 2004).

In economics, commonly used noncognitive variables are Locus of Control derived from Rotter (1966) and the Self-Esteem scale from Rosenberg (1979).

In many case, noncognitive ability and locus of control and/or self-esteem are applied interchangeably. These two scales are available in longitudinal data sets such as NLSY (National Longitudinal Survey of Youth), NELS, Panel Study of Income Dynamics (PSID), and HS&B (High School and Beyond).

Heckman and his collaborators have produced copious evidence that cognitive ability is not the single dominant predictor of educational attainment and earnings. In particular, Heckman, Hsee, and Rubinestein (2002) identified the presence of noncognitive effect without specifying noncognitive variables. They examined why GED (General Educational Development) recipients drop out of high school when they have similar levels of cognitive ability as high school graduates who do not go to college. When cognitive ability measured by the AFQT (Armed Forces Qualification Test) is conditioned, the certificate holders actually attain lower schooling years and earn less than other high school dropouts without the GED certificates. These recipients lack noncognitive skills and hence the GED sends a "mixed signal" to the labor market.

Coleman and DeLeire (2003) present a model of how locus of control affects education attainment. Their analysis of the NELS data confirms that locus of control influences a student's education investment decision through outlook on future earnings. Cebi (2007) replicates their study using a different data set, NLSY, but her findings do not support their predictions. However, a number of other studies find positive effects of locus of control.

Identifying noncognitive ability by locus of control and self-esteem, Heckman, Stixrud and Urzua (2006) estimate the effects of both types of abilities applying the MCMC routine to NLSY. They correct for the problem of reverse causality in schooling and test scores by combining the noncognitive ability with schooling levels of individuals from an older cohort; the final schooling years estimated by this method are independent of the latent ability. The authors simulate the distribution of schooling years for different levels of locus of control and self-esteem. The noncognitive skills are as important in the sense that when the cognitive element controlled for, increasing the decile across the noncognitive distribution brings positive and substantial changes in income and schooling decisions and lowers the likelihood of engaging in risky behaviors such as illicit activities and teenage pregnancy.

Piatek and Pinger (2009) adopt this empirical strategy in examining the data from German Socioeconomic Panel. They find that locus of control measured at the age of 17 is an important predictor of college enrollment; increasing the decile across the noncognitive distribution substantially increases an individual's schooling years. However, locus of control has no impact on wages when schooling is fixed which suggests that it affects earnings only through schooling.

Utilizing the NELS data, Eren (2008) employs instrumental quantile regression in studying the heterogeneous noncognitive effect on earnings. He addresses measurement error that arises from using self-reported measures of noncognitive ability. He finds that higher levels of 8th grade locus of control and self-esteem scales are especially effective for lower quantiles of the wage distribution. By contrast, having higher values of cognitive ability, measured by 8th grade standardized math scores, is most advantageous to workers in the upper quantiles. This reversed pattern can be related to the early research of Bowles and Gintis (1976). In their explanation of how schooling reproduces class inequalities, they claim that lowest levels of division of labor stress docility and dependability while higher levels emphasize more independent activity and creativity. And schools socialize students to accept personality and behavior traits that are required for the different levels within the hierarchical occupational structure. Eren provides evidence that different sets of skills are rewarded to low- and high-wage earners in the labor market.

Rotter's locus of control have had been criticized for low reliability and multi-dimensionability. It falls short of being a robust psychometric instrument. Duttweiler (1984) developed the Internal Control Index (ICI), but this version of locus of control was not administered in any of the NELS surveys. A principal component analysis on the six locus of control items from the NELS base year data shows that the first component accounted for only 39% of the combined variance. Cronbach's alpha method also reported only modest statistical consistency (0.67) which is slightly below the of a lenient cutoff. These findings suggest that those six items do not capture a uniform construct.

Many researchers found little relevance between self-esteem and academic achievement among low-achieving high school students (Ekstrom et al. 1986, Rutter 1986). A plausible explanation is that, as Dee and West (2008) pointed out, self-esteem includes narcissism and defensiveness.

Because noncognitive ability is an elusive concept and the boundaries remain largely unclear, finding its measures is a difficult undertaking. There is no single definite representation. Each measure that was discussed earlier has its own strength and weakness. I therefore chose to focus on more clearly defined noncognitive subject: student engagement. In educational psychology circles, there has been growing consensus on treating noncognitive ability in light of "student engagement" in the classroom. Drawing from an extensive review of literature, Lee and Shute (2010) identified four major noncognitive domains: student engagement, learning strategies, school climate, and social-familial influences. The following variables fall into the category of student engagement: following school rules and arriving at school on time ("behavioral engagement"), self-efficacy and self-discipline ("cognitive-motivational engagement"), feelings toward learning, school and classmates ("emotional engagement").

Finn and Rock (1997) offer solid empirical evidence of a relationship between student engagement and student achievement in high school. Their analysis of the NELS base year sample shows that between two groups of disadvantaged students that successfully completed high school and those who dropped out, there is a substantial gap in either teacher-reported or self-reported classroom behaviors: trying hard in school work, being more attentive and cooperative in class, and coming to class more prepared. The engagement variables that were used provide a useful guide in selecting noncognitive variables for this study – variables which I list in the following section. From among the many engagement variables that they drew from the data, I selected noncognitive indicators that could be utilized for the empirical strategy adopted in this paper.

Furthermore, Finn, Pannozzo, and Achilles (2003) explain the mechanism of academic gains from cutting class sizes in the elementary grades. Improvement in students' academic performance is mediated by changes in student engagement. They borrow the principle of visibility from psychology to claim that students feel more responsible and motivated in smaller classes because they cannot "hide in the crowd (Finn, Pannozzo, and Achilles 2003)." Another change is that students feel a stronger sense of belonging in a smaller group which leads students to follow the group norms. And thus it is likely that students become more academically and socially engaged. Concerning the class size effect in later grades, however, the authors comment that further research needs to be conducted to understand the mechanism in the later grades.

3 NELS:88/90 Data

The data is taken from National Educational Longitudinal Study (NELS) which was conducted by the National Center for Education Statistics (NCES) in 1988. Approximately 24 eighth graders were randomly selected from a sample of 1052 public and private junior high or high school. In the base-year of 1988, a total of 24,599 eighth graders were surveyed and many of the students were followed over the course of subsequent twelve years.

The follow-ups were administered in 1990, 1992, 1994, and 2000 garnering five waves of 12,144 cases. In the first and second resurveys, 338 and 63 fresh students, respectively, were added to maintain the representativeness of the 10th and 12th grade cohorts. The data contains extensive information; about 11,000 pieces of information are available not only from surveys of students, but also those from teachers, parents, and school administrators. Furthermore, scores from standardized tests in English, math, social studies, and science are available for 8th, 10th, and 12th graders, corresponding to the base year and the first and second follow ups.

The most relevant information comes from the teacher surveys. One or two of a student's teacher(s), from the four subjects areas of English, math, social studies/history or science, were required to evaluate the student's characteristics and performance in the classroom. For the base year and the first follow-up, a set of two teachers were assigned to participate and thus four possible subject teacher combinations are available: English-math or English-science and history-math or history-science. The student and teacher responses were collected in the middle of the spring semester in 1988 and 1990, so the teachers had at least two or three months after the initial assignment to observe changes students' engagement. Theoretically it is the same for students given students' self-reports are accurate and honest.

Dee and West (2008) cleverly exploit this unique matched pairs design to mimic a natural experiment with this non-experimental data. Common identification problems addressed in class size studies of non-experimental data such as selection bias and teacher quality can be alleviated with their strategy. Since only one teacher report per student was collected for the 12th grade cohorts, I exclude the 1992 wave from the scope of this paper and utilize the base year and the first follow-up data.

The noncognitive variables are taken from either the teacher evaluations or students' self-reports on classroom engagement behaviors. The teacher responses encompass the following attributes: disruptiveness (DISRUPT), inattentiveness (INATT), homework completion (NOHWK), passiveness (WITHDRAWN), absenteeism/tardiness (ABSTDY) and diligence (WORKHARD). On the other hand, most of the student-reported variables relate to the emotional engagement in the subject matter: how much the student looks forward to the subject (NOTLF), how often she feels challenged in class (FEELCHALLENGED), whether she is afraid to ask questions in subject class (AFASK), etc. These outcome variables are detailed in Table 1 and Table 2. The set of classroom behaviors available in the first follow-up is somewhat different from that of the base year. Most of the items in the base year were measured in binary format whereas the measures in the first follow-up were scored on a five point Likert scale with values "never," "rarely," "some of the time," "most of the time," and "all of the time". The values of some items had to be reversed so that all the variables would capture negative behaviors or feelings;

a higher score indicated lower levels of engagement exhibited in classroom by the student.

Tables 1 and 2 also report other variables that reflect the classroom environment: class size (CLSSIZE), subject-pairing distributions, teacher qualifications (NOVICE), the percentage of limited English proficient classmates (PERLEP), and whether the teacher share similar demographic traits with the student (OTHRACE, OTHSEX). Dee (2005) provided evidence that whether a teacher is of similar gender or race to a student has a large effect on student achievement especially for low SES students. In the original study, only students in public schools were included, but I do not exclude private schools in this paper. The rationale is to make the base year students comparable to my 1990 tenth graders in the sample as there is considerable data loss in constructing the first follow-up sample. Still, the summary statistics detailed in Table 1 are comparable, although they tend to be slightly lower, to those in the corresponding table that Dee and West presented. For example, the average class size in 1988 was 23.8 with a standard deviation of 6.3. If I rule out private schools, the average class size somewhat increases to 24.5 and the dispersion reduces to 5.7 which is very close to the values reported in the original paper. A similar pattern can be observed with other variables; the means and standard deviations of the variables match those in Dee and West more precisely when non-public schools are dropped.

In the following subsections, I describe how my samples for the base year and the first follow-up were constructed. I chose to create two separate samples instead of creating a balanced panel fore two reasons: (1) About a half of the base year sample would be lost when it is merged with the first follow-up sample. (2) Many of base year students transitioned from junior high to high school when the first follow-up was conducted. School fixed effects, in particular, would not be eliminated. Therefore, differences-in-differences was not considered.

3.1 Representativeness of the base year sample

The teacher component is at the level of student-teacher pairs and each teacher student pairing will be the unit of analysis throughout this paper. The raw data contains information on 23,188 students – or 44,512 teacher-student pairs – out of the 24,599 total sampled students in 1988. That is to say 1411 (5.7%) students did not have any teacher reports.

Nationally representing 1,052 schools were randomly assigned with a subject combination from four possible choices and thus all of the sampled students in each school were allocated the same subject pairing. If there was not any data loss in constructing the sample, each pairing would have had equal proportion of about 25%. The unequal distribution in Table 1 indicates that this is not the case.

Usually in the case of small schools, both subjects were taught by the same teacher. This accounts for 776 (3.3%) students. There were also 1864 (8.0%) students with only one report. These two cases were dropped. After merging the student surveys my base year sample contains 20,548 students or 41,096 teacher-student pairs. In sum, there were a total of 24,599 participating students and 4051 (16.4%) students were excluded in the process of constructing the student-teacher sample.

How is this sample different from the entire 8th grade cohort? Table 3 provides a comparison between students in the original student component and those in the new sample. Noticeably, white students make up three percent more of the former. This implies the data loss came mainly from minority students; the students that had only one teacher report or were taught by the same teacher were likely be minority students This is confirmed by the fact that the proportions of Hispanic and black students in my sample is slightly lower. The difference in the racial composition may be the reason behind the slight increase in the SES quartiles and in all four subject standardized test scores. Also the percentage of urban schools decreased perhaps due to the same reason. The proportion of public schools slightly increased indicating non-responses from teachers were likely to come from private schools. Assuming small classes benefit minority students the most, I would expect larger differences in means and standard deviations of classroom behaviors between the 8th grade cohort and my sample. But the students in the new sample actually showed lower levels of engagement for NOTLF and NOTUSE and higher levels for AFASK, only to a very small degree for all three variables. Hence I do not suspect that the effect of excluding some students would results in large differences in outcomes.

3.2 Representativeness of the first follow-up sample

The total participating students in the first follow-up are 20,706 students attending 1,296 schools. Of the total, 15,908 students (76.8%) were linked to 9,987 teachers. But 3,822 students (24.0%) did not carry a set of two teacher reports required for identification and were eliminated. Additionally 96 (0.6%) students had subject combinations such as English-history and math-science and 46 students (0.2%) were evaluated by the same teacher. These two cases were also removed. In sum, the student-teacher sample contains 11,944 students or 23,888 student-teacher pairs. Considerable attrition of students –

24.5% from the raw teacher data and 42.3% from the entire 10th grade cohorts – should be addressed in the analysis of estimations. For now, Table 4 provides a basic comparison between the student survey and the constructed sample.

As in the base year comparisons, the proportion of white students is higher in the student-teacher sample. This implies that the students without either one or two teacher reports were likely to be minority students. Differences in both means and standard deviations of the race indicators show that the racial composition in my sample does not represent the 10th graders in 1990. Students in my sample scored slightly higher in standardized English and history tests and performed slightly worse in standardized math and science tests. In terms of the noncognitive abilities, I observe mixed implications; the students in the new sample showed lower levels engagement in some areas and higher in others. They also reported that they worked harder except in history. Interestingly, both groups reported that they put less effort in history. The differences in FEELCHALLENGED are much smaller than NOTTRYHARD. Students in the new sample were considerably more likely to attend public schools.

Unlike the base year in which the school sample constituted a representative sample for all 8th grade schools in 1988, the first follow-up school sample does not constitute a national probability of schools. This is because students made transitions from middle/junior high schools to high schools. The subject combinations listed in Table 2 show that the English and math teachers completed teacher evaluations significantly more than history and science teachers.

4 Empirical strategy

The matched pair structure of the teacher survey allows Dee and West (2008) to adopt a variation of the model used by Ashenfelter and Krueger (1994) in their study of economic returns to schooling using a sample of twins. For each student i, a set of two student-teacher pairs is given by:

$$y_{1it} = \boldsymbol{\alpha}\mathbf{X_i} + \beta(SIZE_{1it}) + \boldsymbol{\lambda}\mathbf{Z_{1t}} + \theta_{1t} + \mu_i + \epsilon_{1it} \tag{1}$$

$$y_{2it} = \boldsymbol{\alpha}\mathbf{X_i} + \beta(SIZE_{2it}) + \boldsymbol{\lambda}\mathbf{Z_{2t}} + \theta_{2t} + \mu_i + \epsilon_{2it} \tag{2}$$

The indices t, 1 and 2 refer to a teacher, English or history and math or science class, respectively. An important assumptions are that the class size effect on noncognitive abilities and students' unobserved intrinsic characteristics do not vary across subject pairs. And thus, these orderings actually do not matter in theory.

For each student i, denote y_{1it} and y_{2it} the student engagement behaviors evaluated by the first and second teachers. More specifically, these outcomes are misbehaviors in the sense that the higher the the value, the lower the engagement in classroom. The term $\mathbf{X_i}$ are observed student traits and $\mathbf{Z_{1t}}$ and $\mathbf{Z_{2t}}$ are other teacher and classroom traits that capture the dynamics in classroom. An unobserved teacher and student fixed effects are given by θ_t and μ_i respectively.

Note the unobserved student fixed effect or μ_i is assumed to be same across the subjects. In other words, we ignore potential presence of subject-specific fixed effect or μ_{1i} and μ_{2i}. In addition, assume that the error terms, ϵ_{1it} and ϵ_{2it} have zero conditional means with respect to all the covariates. Put it differently, the number of students in a math class has no bearing on a student's engagement in English after controlling for the size of her English class, her own fixed effect, both teachers' fixed effects, and other observed characteristics from both classes. An OLS estimation based on stacking these two equations produce a meaningless estimator due to presence of "lurking variables."

First-differencing can control for those unobserved confounding factors. The difference between these two equations is:

$$y_{1it} - y_{2it} = \beta(SIZE_{1it} - SIZE_{2it}) + \boldsymbol{\lambda}(\mathbf{Z_{1t}} - \mathbf{Z_{2t}}) + (\theta_{1t} - \theta_{2t}) + (\epsilon_{1it} - \epsilon_{2it}) \tag{3}$$

Individual fixed effects, both observed and unobserved are effectively eliminated.

The assumption that the individual fixed effect is subject invariant may pose as a threat to internal validity if students have a particular subject-specific preference. For instance, if students who like math show positive behaviors in classroom are likely to be assigned to smaller classes because of their preference, then I expect the class size impact to be inflated. To account for potential presence of subject-variant unobserved variables, gender and subject interactions are included in the vector of $\mathbf{Z_t}$ along with other classroom characteristics that may affect assignment into smaller classes. Subject test scores will be included in some specifications as well.

Another critical threat to internal validity, possibly more serious than the other, is unobserved teacher characteristics. Although this methodological framework strives to mimic a natural experiment, teachers were not randomly assigned. The NELS administrators randomly assigned subject combinations to each participating schools in the base year. This implied that the teachers who had to evaluate their students were randomly chosen, but students had either direct and indirect choices to select classes. Therefore, one would expect a teacher's characteristics to be related to the size of class she taught and the selection effect can potentially weaken the impact of class size.

5 Class size and student engagement

5.1 Class size effects in the base year and the first follow-up

Table 5 presents estimation of class size impacts on noncognitive and cognitive (measured by standardized subject test scores) variables for the base year sample. It is a direct replication of Dee and West except that my sample is larger by about 4000 students because I do not exclude students attending non-public schools. Table 6 also provides estimates of class size effect for the first follow-up data.

In Column (1) of Table 5, I estimate the equation (1) and (2) together. This specification includes observed student characteristics (OTHSEX, OTHRACE, SCERTIFD, PERLEP), school fixed effects, and observed student characteristics such as race, gender, SES, English proficiency. As discussed in the earlier section, the estimates obtained from stacking the two equations are not to be trusted. These estimates show that a smaller class lowers classroom engagement (DISRUPT, INATT, NOHWK, ABSTDY) and reduces test scores. The only intuitive result is that a smaller class leads students to look more forward to the subject. The standard errors for all six outcomes (NOTLF, NOTUSE, AFASK, DISRUPT, INATT, STEST) that are studied in both papers are almost exactly the same although the class size estimates are varied. In fact, across all five specifications of Table 5 the standard errors match closely to those in the original paper. This demonstrates that the replication was carried out correctly.

Column (1) of Table 6 tells a similar story. The results show that a smaller class leads to decreased levels of engagement for some variables (FEELCHALLENGED, DISRUPT, ABSTDY) and to lower test scores. Column (2) of Table 5 estimates the first-differenced equation (3). The signs of many outcomes change in the direction that makes more common sense. Accounting for student fixed effects, a smaller class increases all of the response variables except for DISRUPT, ABSTDY, and STEST. Still, the magnitude of the impacts are very small. One standard deviation (6.31) decrease in class size would result in about 2% (6.31×0.0027) increase in students' attentiveness in classroom. If the estimate on test scores were statistically significant, it would translate to a scant 0.3% (6.31×0.0005) increase in test scores. Interestingly, the same estimate in Dee and West are about four times bigger and can be converted to only 1.3% increase.

The related column in Table 6 indicates that a smaller class makes students to work harder in the subject (NOTRYHARD). The effect size is quite small; because this variable was evaluated on a five-point Likert Scale, it would be hard to make a precise interpretation. In addition, a smaller class leads students to display more disruptive behaviors. This anomaly is due to unobserved teacher quality and other classroom traits that drives the class size impacts. For instance, disruptive students were more likely to be assigned to smaller classes taught by more experienced teachers. Column (3) detects presence of negative selection. when teacher combination fixed effects are introduced, the coefficients become positive, though the magnitudes still very small; one standard deviation (7.06) decrease in class size would result in about 0.1% (7.06×0.0002) decrease in students' disruptiveness in classroom.

Including teacher pairing fixed effects in Column (3) of Table 5 inflate all of the R^2's to the range of 0.30 and 0.40. Many of the estimators become insignificant (NOTUSE, INATT, WITHDRAWN). Also the class size effect on AFASK increases considerably, suggesting confounding teacher characteristics were not accounted for in previous specifications. Introducing observable classroom characteristics in Column (4) increases the effect size on NOTLF, NOT USE, AFASK indicating that there was nonrandom sorting into smaller classes. Students' self reported measures (NOTLF and AFASK) and one teacher-reported variable (NOHWK) are shown to be significant estimators. Column (5) introduces subject test scores that may be endogenous to subject-specific student fixed effects, if there are any. The values of the estimates show robustness to addition of subject test scores and thus the assumption made earlier in the previous section that student fixed effects remain same across subjects is not unreasonable.

Introducing teacher combination fixed effects in Column (3) of Table 6 increases R^2's much more dramatically to above 0.70. This because there are more teacher pairings relative to the total number of students in the 1988 data. Most of the noncognitive coefficients increase considerably except for NOTTRYHARD and NOTWORKHARD and the opposite pattern shows for the cognitive coefficient. This demonstrates that teacher quality is correlated with the treatment. Addition of classroom characteristics do not influence the estimates much in Column (4). This shows that there was less nonrandom sorting related to classroom traits than to teacher quality in the first follow-up. As in Table 5, Column (5) shows that many of the outcomes are robust to subject-specific test scores, corroborating the assumption that individual fixed effects are subject-invariant. However, none of the estimators from Column

(3) to (5) carry any statistical significance. It is not surprising to find that the standard errors are about two to six times larger in Table 6 since the sample size is almost half as small and we expect a smaller degree of variation in class size.

5.2 Selection in to a smaller class

As mentioned in Section IV, The non-randomness in teacher assignment may bias Grade 10 estimates. The matched-pair method cleverly eliminates student fixed effects, but may not fully contained unobserved teacher characteristics that are correlated with class size. Techniques to identify and control the presence of the omitted variable bias needs to be devised.

Table 7 and Table 8 test for presence of non-random treatment to smaller classes based on classroom characteristics. In Column (1) to Column (5) of both Tables are estimates produced from auxiliary regressions. The specifications are the same as before and that dependent variables are now PERLEP, SCERTIFD, and NOVICE. Column (1) and (2) in Table 5 show that there is negative correlation between PERLEP and class size. Student body with higher fraction of limited English skills were more likely to be assigned to smaller classes. State-certified teachers in the subject were more likely to be assigned to larger classes. Teachers with 1-3 years of experience did not show any pattern of selection. The correlations weaken as I control for more endogenous explanatory variables.

Similar trends in selections are observed in the first follow-up. Table 8 indicates that there seem to be a negative selection with PERLEP and a positive one with teacher certifications. Also the correlations weaken and become insignificant as we include more control for confounding variables. It was implied in Table 6 that classroom traits influenced to a lesser degree than in the base year. This is confirmed by the smaller coefficients in all columns.

5.3 Sample size reduction in the first follow-up

An interpretation from the first follow-up estimates need to be made with some caution. Problems arise because there was a considerable sample size reduction in construction of the student-teacher data. The new sample contains 24.5% less observations from the original NELS teacher component and 42.3% less from the student component. Could we have obtained different results if we had a larger sample size?

Virtually none of the class size effect is significant in the first follow-up. There are two possible explanations. Firstly, the effect size is already very small in the base year. We can suspect that class size effect diminish by the time students reach the 10th grade. It is not counterintuitive that 10th grade outcomes are less dependent on class size. Secondly it can be that the effect size exists but statistical significance cannot be achieved due to a small sample size.

In the base year, only two self-reported measures (NOTLF and AFASK) and one teacher-reported noncognitive variable had statistical significance at a 1% level. Unfortunately a side-by-side comparison against 1990 results can not be made since the first two variables are not available in the first follow-up and NOHWK was measured on different scale in the second wave. Even if some of the first follow-up estimates in Column (4) of Table 6 were significant, the estimated impact would be not at all substantial. For instance, reducing a class size by one standard deviation or 7.06 students would results in only 0.04 (7.06×0.0061) points increase in standardized test scores. Furthermore, the sample contains information of 10,000 students. If this sample size was not enough to give significance to small class size effect, then it would be hard to expect different outcomes with a larger data set.

As an ad-hoc check, I replicated the base year estimation with students who participated in both wave in Table 9. These are 9663 students out of a total of 20548 base year students. Column (4) shows significant and somewhat larger estimates on NOTLF and AFASK compared to those in the first follow-up. My reading from these evidence is that statistical insignificance of the estimates is due to small variation in the data rather than small sample size. The impacts of class size on developing noncognitive development is minimal.

Additionally, I carried out a complementary analysis. I replicated the first follow-up estimation with the same 9,663 students and the results are reported in Table 10. These students account for about 81% of the total 11,944 first follow-up participants in my sample. As expected, the estimates are similar to those in Table 6. In Column (4) in particular, standard errors are generally slightly larger and magnitudes of some estimates differ moderately (NOHWK, NOTTRYHARD).

To test the effect of reduced sample size more formally, we can carry out a multiple imputation analysis on the raw teacher component and fill in engagement variables for the students with one teacher responses. We could expect an increase of 3,822 students or 7644 student pairs. I do not expect that this method would give a much differing results.

5.4 Class size effect across demographic groups

Table 10 and 11 show whether the class size effect is heterogeneous across demographic groups. For the base year results show that there is more meaningful effect on boys than girls. The effect size on NOTLF (0.0102) is larger than the Column (4) of Table 5 estimate of 0.0094. Similarly, effect on WITHDRAWN (0.0031) is three times bigger than the sample average. One standard deviation reduction in class size translates to decrease students' passiveness by 2.0%. This is about 23 percent increase from the average (0.0852) for boys (These numbers are to provide a benchmark. It would be hard to make a real meaning.). This is not surprising because there are numerous articles that establish a connection between gender and outcomes that can be associated as noncognitive (Skinner et al. 2009, Pastor and Reuben 2008).

Furthermore, the effect is more pronounced for students in urban schools than in other areas. The impacts are not significantly different for Black or Hispanic students. Since minority students were more likely to be eliminated in the process of creating the data (See Table 3), these estimates may be imprecise. On the other hand, students attending schools where the fraction of minority students is above 40 percent are shown to benefit more than other 8th graders. Students attending schools in which more than half of the student body receive reduced-price lunch respond more to smaller classes with respect to classroom behaviors. The effect size is not significantly different for students from a single mother family and for students with high school dropout or for college educated parents. The estimates for students attending public schools are significant, but not necessarily higher than the average. Interestingly for students attending catholic schools, smaller classes induce disruptive and inattentive behaviors.

In Table 11, any significant treatment heterogeneity cannot be observed across all groups. The estimates tend to be larger for boys, Hispanics, the low SES and the urban group, but they are not significant. This is not surprising since there is good evidence to believe that the 10th grade effect is almost minimal.

6 Conclusions

Noncognitive ability is difficult to delineate because it encompasses virtually all skills and personality traits that cannot be measured in standardized tests. In this paper, I focus on a very specific construct, student engagement. Utilizing the two cohorts from NELS:88/90, this study finds that the impacts of class size are minimal for 10th graders. Also we do not observe heterogeneous class size effect across various demographic groups. We may suspect that these results are due to a large sample size reduction from constructing the sample, but we have good evidence to believe that these impacts are small; the replication of Dee and West (2008) suggests that the class size effect in Grade 8 is significant, but still weak. Another factor that may have led to insignificant estimates is the presence of a lurking variable. The matched-pair method cleverly eliminates student fixed effects, but may not have fully contained unobserved teacher characteristics that are correlated with class size. The non-randomness in teacher assignment may have weakened the 10th grade impact and consequently subgroup variations.

In economics, research on noncognitive effect of class size has been nonexistent, but a very recent study by Chetty et al. (2010) sheds light on this issue using the data from the STAR experiment. They conclude that a good quality small class in the early grades has a better long-term consequences due to improvement in noncognitive abilities the children gained in the early education classrooms. Their findings are consistent with the lessons from the Perry Program; class size reduction has far better noncognitive consequences in early education.

While empirical evidence in the later grades remain inconclusive and more research should be conducted, numerous studies that were cited earlier support the importance of noncognitive ability in achieving better outcomes in schooling and labor market. These studies suggest that fostering noncognitive ability should be incorporated into any education policies that aim to improve student achievement. Diamond et. al (2007) demonstrate that noncognitive skills – executive functioning in particular – can be improved as a result of effective training in class. Building curriculum that promote the development noncognitive ability may be a cost-effective policy.

Table 1: Outcome and independent variables in the base year sample[1]

Variable	(Variable name)	Mean	Std. dev.	Min-Max	Sample size
Student-reported engagement behaviors in subject:					
Do not look forward to subject	NOTLF	1.3667	0.8902	0-3	39,401
Subject not useful for my future	NOTUSE	1.0019	0.8631	0-3	39,302
Afraid to ask questions in subject class	AFASK	1.1730	0.8968	0-3	39,351
Teacher-reported engagement behaviors in subject:					
Student is frequently disruptive	DISRUPT	0.1279	0.3339	0-1	40,084
Student is consistently inattentive	INATT	0.2068	0.4050	0-1	40,042
Student rarely completes homework	NOHWK	0.1993	0.3995	0-1	40,119
Student is exceptionally withdrawn	WITHDRAWN	0.0823	0.2748	0-1	40,021
Student is frequently absent/tardy	ABSTDY	0.0828	0.2204	0-1	39,606
Standardized test score in subject	STEST	0.0298	0.9932		39,815
Subject combinations:					
English-math		0.2837	0.4508	0-1	41,096
English-science		0.2375	0.4256	0-1	41,096
History-math		0.2284	0.4198	0-1	41,096
History-science		0.2504	0.4332	0-1	41,096
Class size	CLSSIZE	23.7626	6.3133	1-49	40,378
Teacher is of opposite race/ethncity	OTHRACE	0.2985	0.4576	0-1	41,096
Teacher is of opposite gender	OTHSEX	0.4901	0.4999	0-1	41,096
Teacher is certified by state in subject	SCERTIFD	0.8197	0.3844	0-1	41,096
Teacher has 1-3 years of experience	NOVICE	0.1079	0.3102	0-1	41,096
% Classmates with limited English proficiency	PERLEP	0.0120	0.0632	0-1	39,687

[1] This table corresponds to *Table 5* in Dee and West (2008) except my sample includes non-pubic schools.

Table 2: Outcome and independent variables in the first follow-up sample

Variable	(Variable name)	Mean	Std. dev.	Min-Max	Sample size
Student-reported engagement behaviors in subject:					
Often do not work hard in subject class	NOTTRYHARD	0.9194	1.2327	0-5	23,388
Often feel challenged in subject class	FEELCHALLENGED	2.7875	1.2814	0-5	22,930
Teacher-reported engagement behaviors in subject:					
Student is frequently disruptive	DISRUPT	0.6522	0.8663	0-4	23,742
Student is consistently inattentive	INATT	1.1195	0.8603	0-4	23,672
Student rarely completes homework	NOHWK	1.0447	0.9873	0-4	23,609
Student is exceptionally withdrawn	WITHDRAWN	0.0881	0.2834	0-1	23,226
Student does not usually work hard	NOTWORKHARD	0.3617	0.3617	0-1	23,056
Student is frequently absent/tardy	ABSTDY	0.9924	0.6019	0-4	23,192
Standardized test score in subject	STEST	0.0789	0.9795		23,064
Subject combinations:					
English-math		0.3736	0.4838	0-1	23,888
English-science		0.2895	0.4535	0-1	23,888
History-math		0.1718	0.3772	0-1	23,888
History-science		0.1651	0.3713	0-1	23,888
Class size	CLSSIZE	23.3287	7.0631	1-95	22,602
Teacher is of opposite race/ethncity	OTHRACE	0.29000	0.4543	0-1	23,888
Teacher is of opposite gender	OTHSEX	0.4995	0.5000	0-1	23,888
Teacher is certified by state in subject	SCERTIFD	0.9077	0.2895	0-1	23,888
Teacher has 1-3 years of experience	NOVICE	0.1120	0.4370	0-1	23,888
% Classmates with limited English proficiency	PERLEP	0.0201	0.0930	0-1	22,188

Table 3: Characteristics of the base year sample

Variable	Means (standard deviations in parentheses) Population	Student-teacher sample
Not look forward to English	1.3948	1.4030
	(0.8600)	(0.8586)
Not look forward to math	1.4060	1.4122
	(0.8983)	(0.8960)
Not look forward to history	1.3523	1.3534
	(0.9063)	(0.9066)
Not look forward to science	1.3096	1.3096
	(0.9153)	(0.9143)
English not useful	0.8455	0.8510
	(0.7786)	(0.7780)
Math not useful	0.7136	0.7172
	(0.7697)	(0.7686)
History not useful	1.3420	1.3460
	(0.8767)	(0.8735)
Science not useful	1.1380	1.1416
	(0.9040)	(0.9003)
Afraid to ask in English	0.9031	0.8993
	(0.7329)	(0.7281)
Afraid to ask in math	2.0130	2.0213
	(0.7884)	(0.7816)
Afraid to ask in history	0.8793	0.8749
	(0.7436)	(0.0.7399)
Afraid to ask in science	0.8751	0.8682
	(0.7408)	(0.7350)
Female	0.5024	0.5015
	(0.5000)	(0.5000)
Black	0.1223	0.1168
	(0.3277)	(0.3211)
Hispanic	0.1289	0.1163
	(0.3351)	(0.3205)
White	0.6633	0.6903
	(0.4726)	(0.4624)
SES quartiles (1=Low to 4=High)	2.5678	2.5975
	(1.1402)	(1.1369)
Std. English test	50.5002	50.8176
	(10.0812)	(9.9824)
Std. math test	50.6425	50.9126
	(10.2182)	(10.1584)
Std. science test	50.4006	50.7879
	(10.0573)	(10.0382)
Std. history	50.5064	50.8275
	(10.0573)	(9.9474)
Public school	0.7884	0.8033
	(0.4084)	(0.3975)
Urban school	0.3098	0.2876
	(0.4624)	(0.4526)
Sample students (N)	24,566	20,548

Table 4: Characteristics of the first follow-up sample

	Means (standard deviations in parentheses)	
Variable	Population	Student-teacher sample
Often do not work hard in English	0.9095	0.8844
	(0.1747)	(1.1457)
Often do not work hard in math	0.8864	0.8497
	(1.2897)	(1.2366)
Often do not work hard in history	2.1982	2.2385
	(2.1048)	(2.1184)
Often do not work hard in science	1.2096	1.1726
	(0.9153)	(1.5476)
Often feel challenged in English	2.5536	2.5483
	(1.2974)	(1.2922)
Often feel challenged in math	3.0729	3.0789
	(1.2277)	(1.2163)
Often feel challenged in history	2.5742	2.5610
	(1.3252)	(1.3229)
Often feel challenged in science	2.9459	2.5610
	(1.2267)	(1.2140)
Female	0.4947	0.4987
	(0.5000)	(0.5000)
Black	0.1071	0.0970
	(0.3093)	(0.2959)
Hispanic	0.1329	0.1074
	(0.3394)	(0.3097)
White	0.6683	0.7206
	(0.4708)	(0.4487)
SES quartiles (1=Low to 4=High)	2.5542	2.6441
	(1.1397)	(1.1174)
Std. English test	50.4668	51.1688
	(10.0830)	(9.9194)
Std. math test	50.7082	51.5017
	(10.1832)	(9.9380)
Std. history	50.3792	50.5064
	(10.1302)	(9.8523)
Std. science test	50.5146	51.2243
	(10.2290)	(10.0732)
Public school	0.8120	0.8670
	(0.3907)	(0.3395)
Urban school	0.2758	0.2876
	(0.4469)	(0.4471)
Sample students	18,221	11,944

Table 5: Base year: Class size effects on noncognitive and cognitive outcomes[1]

			First-difference (FD) estimates							
	(1)		(2)		(3)		(4)		(5)	
Dependent variable	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2
Do not look forward to subject	0.0056‡	0.0938	0.0090‡	0.0097	0.0088‡	0.3472	0.0094‡	0.3471	0.0081‡	0.3595
	(0.0016)		(0.0021)		(0.0028)		(0.0029)		(0.0030)	
Subject not useful for my future	-0.0021	0.0645	0.0039†	0.0077	0.0048	0.2765	0.0055*	0.2748	0.0044	0.2788
	(0.0015)		(0.0016)		(0.0030)		(0.0031)		(0.0031)	
Afraid to ask questions in subject class	-0.0004	0.0636	0.0089‡	0.0032	0.0117‡	0.2978	0.0120‡	0.2978	0.0110‡	0.3030
	(0.0017)		(0.0016)		(0.0028)		(0.0029)		(0.0029)	
Student is frequently disruptive	-0.0034‡	0.1040	-0.0009	0.0003	-0.0005	0.3440	-0.0006	0.3448	-0.0010	0.3485
	(0.0005)		(0.0006)		(0.0011)		(0.0011)		(0.0011)	
Student is consistently inattentive	-0.0027‡	0.1139	0.0027‡	0.0018	0.0013	0.3600	0.0011	0.3632	0.0008	0.3671
	(0.0007)		(0.0008)		(0.0012)		(0.0013)		(0.0013)	
Student rarely completes homework	-0.0023‡	0.1486	0.0028‡	0.0018	0.0023†	0.3638	0.0023†	0.3662	0.0019	0.3689
	(0.0007)		(0.0007)		(0.0011)		(0.0011)		(0.0011)	
Student is exceptionally withdrawn	-0.0002	0.0643	0.0016‡	0.0009	0.0012	0.3574	0.0013	0.3562	0.0012	0.3607
	(0.0004)		(0.0006)		(0.0009)		(0.0009)		(0.0009)	
Student is frequently absent/tardy	-0.0023‡	0.0998	-0.0005	0.0004	-0.0009	0.3779	-0.0010	0.3737	-0.0012	0.3746
	(0.0003)		(0.0004)		(0.0008)		(0.0008)		(0.0008)	
Standardized test score in subject	0.0212‡	0.3100	-0.0005	0.0081	0.0007	0.3081	0.0008	0.3071	n/a	
	(0.0020)		(0.0014)		(0.0020)		(0.0019)			
Sample sizes (range)	36,232-37,709		18,943-19,373		18,943-19,373		18,293-18,772		17,724-18,089	
Control variables										
Student observables	x									
School fixed effects	x									
Student fixed effects			x		x		x		x	
Teacher-pairing fixed effects					x		x		x	
Teacher/Classroom observables							x		x	
Subject test score									x	

[1] This table corresponds to *Table 6* in Dee and West (2008) except my sample includes non-pubic schools.
[2] All models include gender and subject interaction terms.
[3] Standard errors are adjusted for school-level clustering.
[4] * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 6: First follow-up: Class size effects on noncognitive and cognitive outcomes[1]

			First-difference (FD) estimates							
	(1)		(2)		(3)		(4)		(5)	
Dependent variable	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2
Often do not work hard in subject class	0.0043	0.1244	0.0050‡	0.0080	0.0024	0.7453	0.0024	0.7453	0.0035	0.7390
	(0.0018)		(0.0018)		(0.0079)		(0.0079)		(0.0081)	
Often feel challenged in subject class	-0.0052‡	0.1095	-0.0048*	0.0011	-0.0024	0.7173	-0.0024	0.7176	-0.0040	0.7152
	(0.0020)		(0.0021)		(0.0088)		(0.0088)		(0.0089)	
Student is frequently disruptive	-0.0052‡	0.1476	-0.0041‡	0.0032	0.0002	0.7661	0.0003	0.7666	0.0002	0.7639
	(0.0014)		(0.0014)		(0.0058)		(0.0058)		(0.0061)	
Student is consistently inattentive	-0.0022*	0.1516	0.0005	0.0007	0.0046	0.7581	0.0045	0.7581	0.0033	0.7549
	(0.0012)		(0.0013)		(0.0057)		(0.0057)		(0.0059)	
Student rarely completes homework	-0.0057	0.1931	-0.0016	0.0028	0.0010	0.7551	0.0010	0.7556	0.0009	0.7547
	(0.0014)		(0.0004)		(0.0019)		(0.0063)		(0.0067)	
Student is exceptionally withdrawn	-0.0000	0.0947	0.0005	0.0008	0.0028	0.7473	0.0028	0.7476	0.0030	0.7462
	(0.0004)		(0.0004)		(0.0019)		(0.0019)		(0.0020)	
Student does not usually work hard	-0.0001	0.1396	0.0014	0.0016	-0.0025	0.7392	-0.0030	0.7356	-0.0030	0.7356
	(0.0007)		(0.0008)		(0.0038)		(0.0039)		(0.0039)	
Student is frequently absent/tardy	-0.0034‡	0.1659	-0.0011	0.0016	-0.0001	0.7918	-0.0002	0.7922	-0.0008	0.7907
	(0.0009)		(0.0009)		(0.0036)		(0.0036)		(0.0038)	
Standardized test score in subject	0.0087‡	0.3345	-0.0011	0.0313	-0.0061	0.7012	-0.0061	0.7014	n/a	
	(0.0013)		(0.0010)		(0.0046)		(0.0046)			
Sample sizes (range)	20,749-21,984		10,011-10,600		10,011-10,600		10,011-10,600		9,438-9,866	
Control variables										
Student observables	x									
School fixed effects	x									
Student fixed effects			x		x		x		x	
Teacher-pairing fixed effects					x		x		x	
Teacher/Classroom observables							x		x	
Subject test score									x	

1 All models include gender and subject interaction terms.
2 Standard errors are adjusted for school-level clustering.
3 * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 7: Base year: Selection on classroom and teacher observables[1]

		First-difference (FD) estimates			
	(1)	(2)	(3)	(4)	(5)
Dependent variable	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)
% Students with limited English proficiency	-0.0010‡	-0.0009‡	-0.0004	-0.0004	-0.0004
	(0.0002)	(0.0003)	(0.0004)	(0.0004)	(0.0004)
Teacher certified in subject	0.0060‡	0.0039†	n/a	n/a	n/a
	(0.0011)	(0.0019)	n/a	n/a	n/a
Teacher with 1-3 years of experience	-0.0003	0.0000	n/a	n/a	n/a
	(0.0007)	(0.0012)	n/a	n/a	n/a
Control variables					
School fixed effects	x				
Student fixed effects		x	x	x	x
Teacher-pairing fixed effects			x	x	x
Teacher/Classroom observables	x			x	x
Subject test score					x

1 This table corresponds to *Table 7* in Dee and West (2008) except my sample includes non-pubic schools.
2 All models include gender and subject interaction terms.
3 Standard errors are adjusted for school-level clustering.
4 * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 8: First Follow-up: Selection on classroom and teacher observables[1]

		First-difference (FD) estimates			
	(1)	(2)	(3)	(4)	(5)
Dependent variable	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)	$\hat{\beta}$ (SE)
% Classmates with limited English proficiency	-0.0007‡	-0.0004‡	-0.0003	-0.0003	-0.0003
	(0.0002)	(0.0002)	(0.0004)	(0.0004)	(0.0004)
Teacher certified in the subject	0.0031‡	0.0018*	n/a	n/a	n/a
	(0.0007)	(0.0008)	n/a	n/a	n/a
Teacher has 1-3 years of experience	0.0002	-0.0004	n/a	n/a	n/a
	(0.0006)	(0.0009)	n/a	n/a	n/a
Control variables					
School fixed effects	x				
Student fixed effects		x	x	x	x
Teacher-pairing fixed effects			x	x	x
Teacher/Classroom observables	x			x	x
Subject test score					x

[1] All models include gender and subject interaction terms.
[2] Standard errors are adjusted for school-level clustering.
[3] * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 9: Base year: Class size effect for students in first follow-up

			First-difference (FD) estimates							
	(1)		(2)		(3)		(4)		(5)	
Dependent variable	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2
Do not look forward to subject	0.0043†	0.1205	0.0068‡	0.0107	0.0104†	0.4338	0.0104†	0.4362	0.0080*	0.4466
	(0.0021)		(0.0025)		(0.0045)		(0.0046)		(0.0047)	
Afraid to ask questions in subject class	0.0000	0.0952	0.0070‡	0.0014	0.0100†	0.3830	0.0098†	0.3824	0.0077*	0.3888
	(0.0022)		(0.0022)		(0.0042)		(0.0044)		(0.0045)	
Student is consistently inattentive	-0.0020†	0.1210	0.0023†	0.0016	-0.0011	0.4530	-0.0012	0.4504	-0.0014	0.4560
	(0.0008)		(0.0010)		(0.0016)		(0.0017)		(0.0017)	
Student rarely completes homework	-0.0023‡	0.1504	0.0016*	0.2433	-0.0007	0.4618	0.0007	0.4597	0.0004	0.4657
	(0.0008)		(0.0009)		(0.0014)		(0.0015)		(0.0015)	
Student is exceptionally withdrawn	-0.0003	0.0936	0.0010	0.0004	0.0000	0.4583	0.0004	0.4581	0.0006	0.4612
	(0.0005)		(0.0007)		(0.0013)		(0.0014)		(0.0014)	
Sample sizes (range)	17,395-17,974		9,064-9,188		9,064-9,188		8,756-8,885		8,499-8,595	
Control variables										
Student observables	x									
School fixed effects	x									
Student fixed effects			x		x		x		x	
Teacher-pairing fixed effects					x		x		x	
Teacher/Classroom observables							x		x	
Subject test score									x	

[1] All models include gender and subject interaction terms.
[2] Standard errors are adjusted for school-level clustering.
[3] * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 10: First follow-up: Class size effect for students in base year

	(1)		First-difference (FD) estimates (2)		(3)		(4)		(5)	
Dependent variable	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2	$\hat{\beta}$ (SE)	R^2
Often do not work hard in subject class	0.0042†	0.1356	0.0050†	0.0078	0.0013	0.7382	0.0013	0.7382	0.0021	0.7315
	(0.0019)		(0.0020)		(0.0089)		(0.0089)		(0.0092)	
Often feel challenged in subject class	-0.0049†	0.1158	-0.0050†	0.0017	-0.0012	0.8482	-0.0013	0.8970	-0.0032	0.7157
	(0.0021)		(0.0022)		(0.0065)		(0.0095)		(0.0096)	
Student is frequently disruptive	-0.0044‡	0.1490	-0.0042‡	0.0039	-0.0020	0.7562	-0.0019	0.7666	-0.0012	0.7630
	(0.0015)		(0.0015)		(0.0065)		(0.0064)		(0.0067)	
Student is consistently inattentive	-0.0016	0.1574	0.0001	0.0011	0.0034	0.7559	0.0034	0.7560	0.0024	0.7537
	(0.0013)		(0.0014)		(0.0064)		(0.0064)		(0.0067)	
Student rarely completes homework	-0.0043‡	0.1931	-0.0012	0.0033	0.0029	0.7487	0.0030	0.7490	0.0022	0.7481
	(0.0015)		(0.0015)		(0.0072)		(0.0072)		(0.0076)	
Student is exceptionally withdrawn	0.0000	0.0943	0.0008	0.0010	0.0029	0.7490	0.0029	0.7492	0.0028	0.7482
	(0.0004)		(0.0005)		(0.0020)		(0.0020)		(0.0020)	
Sample sizes (range)	17,299-17,888		8,228-8,629		8,228-8,624		8,228-8,624		7,780-8,087	
Control variables										
Student observables	x									
School fixed effects	x									
Student fixed effects			x		x		x		x	
Teacher-pairing fixed effects					x		x		x	
Teacher/Classroom observables							x		x	
Subject test score									x	

[1] All models include gender and subject interaction terms.
[2] Standard errors are adjusted for school-level clustering.
[3] * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 11: Base year: Class size effects across demographic groups[1]

Dependent variable (# Sampled students)	Boys (20,488)	Girls (20,608)	Black (4,798)	Hispanic (4,778)	Low SES (9,504)	High SES (12,106)	Urban (11,818)	Suburban (17,238)	Rural (12,040)
Do not look forward to subject	0.0102†	0.0056	0.0067	0.0145	0.0065	0.0092	0.0107†	0.0078*	0.0078*
Subject not useful for my future	0.0080	0.0064	0.0213	0.0072	0.0087	0.0023	0.0100†	0.0062	0.0055
Afraid to ask questions in subject class	-0.0004	0.0033	-0.0052	0.0218	-0.0004	0.0025	-0.0078	0.0085*	0.0051
Student is frequently disruptive	0.0007	-0.0018	-0.0039	0.0078	-0.0011	-0.0022	-0.0011	-0.0005	0.0000
Student is consistently inattentive	0.0022	-0.0016	0.0093	0.0091*	0.0016	-0.0012	0.0003	0.0003	0.0029
Student rarely completes homework	0.0044†	-0.0012	0.0091	0.0001	0.0088	-0.0008	0.0028	0.0024	0.0018
Student is exceptionally withdrawn	0.0031*	-0.0012	0.0045	0.0013	0.0034	-0.0001	0.0036†	0.0003	0.0012
Student is frequently absent/tardy	-0.0005	-0.0021	-0.0030	0.0001	0.0008	-0.0003	-0.0029*	0.0003	-0.0011
Standardized test score in subject	0.0036	-0.0022	-0.0038	-0.0043	-0.0012	-0.0005	-0.0031	0.0048	-0.0019

[1] This table corresponds to *Table 10* in Dee and West (2008).
[2] N = 20,548 students (41,096 student-teacher pairs)
[3] All models include gender and subject interaction terms.Standard errors are adjusted for school-level clustering.
[4] * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 12: Base year: Class size effects across demographic groups[1] *(cont'd)*

	School types			Parents characteristics			School environment	
Dependent variable (# Sampled students)	Public (33,014)	Private (4,366)	Catholic (3,716)	Single mother (6,424)	HSD (3,962)	College graduates (12,568)	%Reduced lunch (5,242)	% Minority (9,654)
Do not look forward to subject	0.0079‡	0.0235‡	0.0045	0.0045	-0.0063	0.0121*	0.0041	-0.0021
Subject not useful for my future	0.0080‡	0.0138*	-0.0080	0.0017	0.0105	0.0039	0.0140†	0.0125†
Afraid to ask questions in subject class	0.0043	-0.0054	-0.0023	-0.0250*	-0.0065	0.0082	-0.0028	0.0026
Student is frequently disruptive	0.0002	0.0002	-0.0081‡	-0.0045	-0.0041	-0.0032	0.0029	-0.0002
Student is consistently inattentive	0.0017	0.0043	-0.0072†	0.0007	-0.0092	-0.0003	0.0034	0.0032
Student rarely completes homework	0.0028†	0.0038	-0.0030	0.0028	0.0094	0.0004	0.0067	0.0028
Student is exceptionally withdrawn	0.0011	0.0029	0.0012	0.0039	0.0012	0.0001	0.0010	0.0001
Student is frequently absent/tardy	-0.0008	-0.0013	-0.0012	-0.0007	0.0048	-0.0035	0.0025	-0.0002
Standardized test score in subject	0.0002	0.0023	0.0033	-0.0008	-0.0157	0.0015	-0.0064	-0.0009

[1] N = 20,548 students (41,096 student-teacher pairs)
[2] All models include gender and subject interaction terms.Standard errors are adjusted for school-level clustering.
[3] * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 13: First follow-up: Class size effects across demographic groups[1]

Dependent variable (# Sampled students)	Boys (11,976)	Girls (11,912)	Black (2,316)	Hispanic (2,566)	Low SES (4,850)	High SES (7,026)	Urban (6,596)	Suburban 13,328)	Rural (3,620)
Often do not work hard in subject class	0.0063	0.0024	-0.0065	0.0029	0.0056	-0.0156	0.0034	0.0011	0.0057
Often feel challenged in subject class	0.0039	-0.0025	0.0015	0.0069	0.0053	-0.0001	0.0179	-0.0097	0.0024
Student is frequently disruptive	-0.0059	0.0003	-0.0232	-0.0023	0.0070	-0.0017	-0.0084	-0.0011	0.0101
Student is consistently inattentive	0.0206	0.0046	-0.0026	0.0082	0.0179	0.0100	0.0075	0.0015	0.0098
Student rarely completes homework	0.0067	0.0010	-0.0017	0.0222	0.0253	0.0003	0.0097	-0.0060	0.0115
Student is exceptionally withdrawn	0.0067	0.0010	-0.0017	0.0222	0.0253	0.0003	0.0097	-0.0060	0.0115
Student does not usually work hard	-0.0072	-0.0030	-0.0097	0.0137	0.0166	-0.0095	0.0078	-0.0069	-0.0001
Student is frequently absent/tardy	0.0040	-0.0002	0.0015	0.0058	-0.0039	-0.0053	-0.0009	0.0005	-0.0012
Standardized test score in subject	-0.0038	-0.0061	-0.0071	-0.0108	0.0009	0.0011	0.0014	-0.0083	-0.0066

[1] N = 11,944 students (23,888 student-teacher pairs)
[2] All models include gender and subject interaction terms.Standard errors are adjusted for school-level clustering.
[3] * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

Table 14: First follow-up: Class size effects across demographic groups *(cont'd)*[1]

	School types			Parents characteristics			School environment
Dependent variable (# Sampled students)	Public (20,712)	Private (1,872)	Catholic (1,304)	Single mother (3,008)	HSD (1,954)	College graduate (7,382)	%Reduced lunch (1,794)
Often do not work hard in subject class	0.0012	0.0421	-0.0070	-0.0721	-0.0564	-0.0039	-0.0001
Often feel challenged in subject class	-0.0025	-0.0372	0.0247	0.0154	-0.0232	-0.0004	-0.0145
Student is frequently disruptive	0.0004	-0.0009	-0.0019	0.0024	-0.0042	0.0046	0.0300
Student is consistently inattentive	0.0037	0.0239	0.0036	0.0170	0.0185	0.0141	0.0179
Student rarely completes homework	-0.0006	0.0089	0.0141	0.0429	-0.0061	0.0080	0.0215
Student is exceptionally withdrawn	0.0028	0.0093	-0.0019	0.0013	0.0025	0.0029	-0.0003
Student does not usually work hard	-0.0035	0.0041	0.0048	0.0178	0.0106	-0.0070	0.0139
Student is frequently absent/tardy	-0.0003	0.0012	0.0006	0.0193	0.0081	0.0016	-0.0009
Standardized test score in subject	-0.0061	0.0064	-0.0173	-0.0029	0.0168	-0.0037	-0.0139

[1] N = 11,944 students (23,888 student-teacher pairs)
[2] All models include gender and subject interaction terms.Standard errors are adjusted for school-level clustering.
[3] * Statistically significant at the 10-percent level; †Statistically significant at the 5-percent level; ‡Statistically significant at the 1-percent level

References

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. ter Weel (2008). The economics and psychology of personality traits. The Journal of Human Resources 43(4), 972–1059.

Bowles, S. and H. Gintis (1976). Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life. Basic Books.

Bowles, S. and H. Gintis (2002). Schooling in capitalist america revisited. Sociology Education 75, 1–18.

Carson, S. M. (2005). Developmentally sensitive measures of executive function in preschool child. Developmental Neuropsychology 28(2), 595–616.

Cebi, M. (2007). Locus of contorl and human capital investment revisited. The Journal of Human Resources 42, 919–932.

Coleman, M. and T. DeLeire (2003). An economic model of locus of control and the human capital investment decision. The Journal of Human Resources 38, 701–721.

Cunha, F. and J. Heckman (2007). The economics of human development: The technology of skill formation. American Economic Review 97, 31–47.

Dee, T. and M. West (2008). The non-cognitive returns to class size. NBER Working Paper No. 13994, May 2008.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity or gender matter? American Economic Review 95(2), 158–165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. Journal of Human Resources 43(2), 528–554.

Diamond, A., W. Barnett, J. Thomas, and S. Munroe (2007). Preschool program improves cognitive control. Science 318, 1387–1388.

Duckworth, A. L. and M. E. Seligman (2005). Self-discipline outdoes iq in predicting academic performance of adolescents. Psychological Science 16, 939–943.

Duckworth, K., R. Akerman, A. MacGregor, E. Salter, and J. Vorhaus (2009). Self-regulated learning: a literature review. The Wider Benefits of Learning Research Report Series 33, 1–98.

Eren, O. (2008). Ability, schooling inputs and earnings: Evidence from nels. 2008.

Eysench, S. B., G. Easting, and P. Pearson (1984). Age norms for impulsiveness, venturesomeness and empathy in children. Personality and Individual Differences 5, 315–321.

Farkas, G. (2003). Cognitive skills and noncognitive traits and behaviors in stratification processes. Annual Review of Sociology 29, 541–562.

Finn, J. D., G. M. Pannozzo, and C. M. Achilles (2003). The "why's of class size: Student behavior in small classes. Review of Educational Research 73(3), 321–368.

Finn, J. D. and D. A. Rock (1997). Academic success among students at risk for school failure. Journal of Applied Psychology 82, 221–234.

Gioia, G. A. and P. K. Isquith (2004). Ecological assessment of executive function in traumatic brain injury. Developmental Neuropsychology 25, 135–158.

Hanushek, E. A. (1999). Some findings from an independent investigation of the tennessee star experiment and from other investigations fo class size effects. Educational Evaluation and Policy Anlaysis 21(2), 143–163.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. Science 312, 900–902.

Heckman, J. J., J. Hsee, and Y. Rubinstein (2002). The ged is a "mixed signal": The effect of cognitive and non-cognitive skills on human capital and labor market outcomes. 2002.

Heckman, J. J. and D. V. Masterov (2007). The productivity argument for investing in young children. In T.W. Schultz Award Lecture at the Allied Social Sciences Association annual meeting, Chicago, January 5-7, 2007.

Heckman, J. J. and Y. Rubinstein (2001). The importance of noncognitive skills: Lessons from the ged testing program. The American Economic Review 91, 145–149.

Heckman, J. J., J. Stixrud, and S. Urzua (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor Economics 24, 411–482.

Jacob, B. A. (2002). Where the boys aren't: non-cognitive skills, returns to school and the gender gap in higher education. Economics of Education Review 21, 589–598.

Kirst, M. (1997). Smaller classes aren't a cure-all. New York Times August 18, 1997.

Krueger, A. B. and D. M. Whitmore (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. Economic Journal 111, 1–28.

Lee, J. and V. J. Shute (2010). Personal and social-contextual factors in k-12 academic performance: An integrative perspective on student learning. Educational Psychologist 45, 185–202.

Messick, S. (1979). Potential uses of noncognitive measurement in education. Journal of Educational Psychology 71, 281–292.

Pastor, P. N. and C. A. Reuben (2008). Hyperactivity disorder and learning disability: United states, 2004 - 2006. National Center for Health Statistics. Vital Health Statistics Series No. 10 237, 1–14.

Piatek, R. and P. Pinger (2008). Maintaining (locus of) control?: Assessing the stability and impact of noncognitive skills over the life-cycle. 2009.

Schanzenbach, D. W. What have researchers learned from project star? Harris School Working Paper Series 06.06.

Segal, C. (2008). Classroom behavior. The Journal of Human Resources 43, 783–814.

Skinner, M. L., K. P. Haggerty, and a. R. F. C. Charles B. Flemng (2009). Predicting functional resilience among young-adult children of opiate-dependent parents. Journal of Adolescent Health 44, 283–290.

Tangney, J. P., R. F. Baumeister, and A. L. Boone (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. Journal of Personality 72, 271–322.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. Anxiety and the anxiety disorders.

Waber, D. P., E. B. Gerber, V. Y. Turcios, E. R. Wagner, and P. W. Forbes (2006). Executive functions and performance on high-stakes testing in children from urban schools. Developmental Neuropsychology 29(3), 459–477.

Wolfe, R. N. and S. D. Johnson (1995). Personality as a predictor of college performance. Educational and Psychological Measurement 55, 177–185.

Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. The MIT Press.