The Non-Cognitive Returns to Class Size

Thomas S. Dee
University of Virginia

Martin R. West
Harvard Graduate School of Education

The authors use nationally representative survey data and a research design that relies on contemporaneous within-student and within-teacher comparisons across two academic subjects to estimate how class size affects certain non-cognitive skills in middle school. Their results indicate that smaller eighth-grade classes are associated with improvements in several measures of school engagement, with effect sizes ranging from .05 to .09 and smaller effects persisting 2 years later. Patterns of selection on observed traits and falsification exercises suggest that these results accurately identify (or possibly understate) the causal effects of smaller classes. Given the estimated earnings impact of these non-cognitive skills, the implied internal rate of return from an eighth-grade class-size reduction is 4.6% overall, but 7.9% in urban schools.

Keywords: *class size, non-cognitive skills, school engagement*

BOTH policymakers and the broader public have an enduring interest in identifying school reforms that will contribute to positive long-term social and economic outcomes. One of the most popular strategies in recent decades has been to reduce the size of classes, in particular for children in early grades. Over the past 30 years, 24 states have implemented measures encouraging or mandating class-size reductions (Education Commission of the States, 2005). The presumed benefits of smaller classes have figured prominently in recent legal battles over the equity and adequacy of state school finance systems (West & Peterson, 2007). And Howell, West, and Peterson (2007) reported that 77% of American adults would prefer to see new educational dollars spent on reducing class sizes rather than on increasing teacher salaries.

Although class-size reduction has strong intuitive appeal among parents and policymakers, its effectiveness (and cost-effectiveness) continues to be debated among researchers. Krueger (2003b), for example, asserted that a "consensus is emerging that smaller classes raise student achievement, both on average and in particular for children from low-income and minority communities" (p. 36). Hanushek (2003), on the other hand, argued that class-size reductions are an "expensive and generally unproductive policy" (p. F92).

The growing recognition of the importance of "non-cognitive" skills for later life outcomes

may have important implications for this debate. The term *non-cognitive skills* refers to a broad and multidimensional range of work habits (e.g., effort and self-control) and behavioral traits (e.g., confidence and emotional stability) that are not measured by conventional tests of cognitive ability (ter Weel, 2008). Several recent empirical studies by labor economists have drawn attention to the importance of such non-cognitive skills for long-term educational and labor-market outcomes (e.g., Heckman, Stixrud, & Urzua, 2006). Moreover, unlike IQ, which largely stabilizes while students are in elementary school, non-cognitive skills appear to be malleable at later ages. As Carneiro and Heckman (2003) noted, this evidence suggests that evaluations of educational interventions should incorporate analyses of their effects on both cognitive and non-cognitive skills. Yet, although numerous researchers have hypothesized that smaller classes could improve non-cognitive skills, there exists little reliable evidence on their effects on these types of outcomes.[1]

This article uses nationally representative survey data from the National Education Longitudinal Study of 1988 (NELS:88) to estimate the effect of eighth-grade class size on certain non-cognitive skills (i.e., measures of school engagement based on teacher and student surveys) as well as on traditional measures of cognitive skills. To identify the causal effect of class size on these outcomes with observational data, we rely on contemporaneous within-student, within-teacher comparisons across two academic subjects. This identification strategy, which to our knowledge is new to the literature on class size, closely parallels the approach used to evaluate data from identical twin pairs (e.g., Ashenfelter & Krueger, 1994; Ashenfelter & Rouse, 1998; Rouse, 1999). It has also been used in recent studies of the effects associated with teacher traits (e.g., Dee, 2005, 2007; Ouazad, 2008).

The results based on this strategy indicate that smaller class sizes in eighth grade are associated with improvements in several indicators of school engagement, with effect sizes ranging from .05 to .09, and persistent but smaller improvements observed 2 years later. The patterns of selection on observed student traits in these specifications suggest that these results accurately identify (or possibly understate) the true

causal effects of smaller classes. Similarly, falsification exercises based on assessing the effects of class size on other-subject outcomes strengthen confidence in their internal validity. Using data from the 2000 follow-up interview of adult NELS:88 participants, we constructed rough cost-benefit analyses of general and targeted class-size reductions in the eighth grade in light of their effects on both cognitive and non-cognitive skills.

The remainder of the article is organized as follows. Section 2 summarizes recent evidence on the effects of class-size reductions; the relationship between non-cognitive skills, academic achievement, and labor-market success; and the role of school engagement in educational production. Sections 3 and 4 present our analysis of the effects of eighth-grade class size on indicators of student engagement in the NELS:88 database and compare the costs and benefits of general and targeted class-size reductions in the eighth grade in light of the observed effects of class size on both cognitive and non-cognitive skills. The final section discusses the implications of our results for policy and research.

Class Size and Non-Cognitive Skills

The scholarly debate over the effectiveness of class-size reductions has a long history (Glass & Smith, 1978).[2] A central challenge in assessing the true effects of smaller classes is that students with a propensity for poor achievement may be systematically assigned to smaller classes (Lazear, 2001). Similarly, the effects of a teacher's unobserved quality on the size of the teacher's assigned class could also undermine conventional inferences about the effects of class size on student outcomes.

Because of these identification challenges, the results from the one large-scale experiment in class-size reduction have played a prominent role in research and policy debates. In the 1980s, the state of Tennessee carried out Project STAR (Student/Teacher Achievement Ratio), a 4-year study during which the students in 79 participating elementary schools were randomly assigned to small and regular-sized classes. Evaluations based on these experimental data indicated that assignment to a small class improved student achievement (e.g., Finn & Achilles, 1990; Finn, Fulton, Zaharias, & Nye, 1989). In an influential

24

reanalysis intended to address concerns about non-random attrition from the experiment as well as treatment crossover, Krueger (1999) found that students randomly assigned to classes with eight fewer students in kindergarten performed 0.2 standard deviations better on math and reading tests.[3] Krueger (2003a) compared the cost of an early class-size intervention like Project STAR with the estimated present discounted value of the adult earnings gains implied by improved test scores and concluded that the internal rate of return to class-size reductions is roughly 6%. However, enthusiasm for class-size reduction has been tempered by other research, using quasi-experimental methods, that shows no evidence of class-size effects on student achievement (e.g., Hoxby, 2000).[4]

The potential contributions of class-size reductions to the development of economically relevant non-cognitive skills have been largely missing from this debate. Several commentators have suggested that such effects are likely to exist and be empirically meaningful. For example, Krueger (2003a) stated that existing cost-benefit analyses of class-size reductions based on Project STAR probably understate its benefits because it is "likely that school resources influence non-cognitive abilities, which in turn influence earnings" (p. F58). There are multiple, plausible mechanisms by which smaller classes may promote educationally relevant non-cognitive skills. For example, by making it easier for teachers to limit disruptive behavior, smaller class sizes may facilitate the development of attentiveness and self-control. Furthermore, smaller class sizes may increase the capacity of teachers to shape student motivation, to elicit effort, and to develop their resiliency in the face of educational challenges.[5] However, there is little direct evidence on whether smaller classes actually improve non-cognitive skills.

It is arguable that the best extant evidence is based on follow-up studies that analyzed teacher reports of the traits of some Project STAR participants. Finn et al. (1989) found that the fourth graders who had been taught in small classes during the Project STAR experiment demonstrated significantly higher levels of effort and initiative and lower levels of non-participatory behavior than students who had been taught in regular classes. However, these effects could not

be detected among a group of Project STAR participants observed in eighth grade (Finn, Pannozzo, & Achilles, 2003; Voelkl, 1995). Furthermore, a recent re-analysis of these data (Dee & West, 2008) found that these treatment effects were limited to the fourth-grade initiative measure in specifications that addressed several shortcomings of the original evaluations.[6]

The lack of evidence on whether smaller classes improve non-cognitive skills is an important gap in the literature because of the growing recognition that such skills play a vital but underappreciated role in long-term academic and economic success. In particular, a recent and growing literature among labor economists has underscored the empirical relevance of such skills.[7] The contemporary interest in non-cognitive skills appears to have been motivated by the observation that high school dropouts who successfully complete a General Education Development (GED) test have lower wages and schooling levels than other high school dropouts after controlling for measured cognitive ability. Heckman and Rubenstein (2001) argued that the GED is a "mixed signal" that attracts high school dropouts with relatively high cognitive skills but lower levels of unspecified non-cognitive skills that are relevant for educational attainment and valued in the labor market.

More recent studies have examined the long-term consequences of *measured* non-cognitive traits. For example, several studies have linked the Rosenberg Self-Esteem Scale and the Rotter Locus of Control Scale to subsequent labor-market outcomes (e.g., Deke & Haimson, 2006; Heckman et al., 2006; Waddell, 2006).[8] Kuhn and Weinberger (2005) presented evidence that, conditional on cognitive ability, youthful indicators of leadership are associated with adult labor-market success. Another study by Borghans, ter Weel, and Weinberg (2008) found that an interpersonal trait characterized as directness has a positive effect on wages. Two other studies in this literature have explored the role of non-cognitive skills in explaining gender and racial gaps in labor-market performance (Fortin, 2008; Urzua, 2008).

Several other recent studies have focused on the work habits and behavioral traits of students. For example, Segal (2008b) found that a measure of test-taking motivation among young,

male NLSY participants strongly predicts labor-market earnings 23 years later. Similarly, Deke and Haimson (2006) found that a composite measure of student work habits (based on student and teacher reports) has an effect on subsequent educational attainment similar in magnitude to that of a test-score measure. Using British data, Blanden, Gregg, and Macmillan (2007) found that non-cognitive measures such as teacher reports of student effort appear to influence labor market outcomes but do so largely through their effects on educational outcomes.

In addition to providing evidence on the policy relevance of non-cognitive skills, these studies also illustrate that non-cognitive skills, as currently conceptualized, encompass an exceptionally broad array of constructs (e.g., discipline, motivation, locus of control, self-esteem, etc.). In fact, Heckman et al. (2006) discussed a principal-components factor analysis that suggests that non-cognitive skills reflect multiple latent factors. Our study's measures, which are described in more detail below, reflect both teacher observations of student behavior (e.g., disruptiveness, inattentiveness) and student reports of motivation and self-confidence similar to those used in other recent studies. We present evidence that these diverse indicators appear to influence long-term student outcomes. However, it should be noted that these indicators clearly do not encompass all the constructs that may constitute non-cognitive skills. Rather, our focus on these particular indicators is motivated both by their practical relevance (as demonstrated by their association with long-term outcomes) and by the fact that they are amenable to a research design that allows us to credibly identify class-size effects (i.e., they have contemporaneous, within-student variation across different academic subjects).

It should also be noted that the indicators of non-cognitive skills used in this and other recent studies correspond closely to those used by educational psychologists in an independent literature on "school engagement." Like the recent studies of non-cognitive skills, this literature has similarly described engagement as a broad multidimensional construct with important implications for academic success. For example, in a recent review of this literature, Fredricks, Blumenfeld, and Paris (2004) characterized school engagement

as a "fusion of behavior, emotion, and cognition" that implies an active commitment to education. They also noted that concerns about the level of school engagement among American students have grown more salient in recent years because of societal declines in respect for authority figures, institutions, and their attendant academic expectations.

This literature in educational psychology has identified specific dimensions of engagement that correspond closely with the developing conceptualization of non-cognitive skills and the indicators used in this study. *Behavioral* engagement focuses on forms of academic participation such as attendance, not being disruptive, effort, assignment completion, attentiveness in class, and asking questions (Fredricks et al., 2004; Glanville & Wildhagen, 2007). In contrast, *psychological* or *emotional* engagement (Fredricks et al., 2004; Glanville & Wildhagen, 2007) consists of students' affective reactions to teachers, peers, and academics in general (e.g., interest, boredom, motivation, anxiety, and sense of belonging).[9] We find that most of the non-cognitive variables used in this study cluster plausibly into these behavioral and psychological constructs. Furthermore, we find that aggregating the variables we use into these two constructs leads to results quite similar to those reported below.

Evidence From NELS:88

Some of the key economic and educational benefits that accrue from investments in smaller classes may be due to their effects on important, non-cognitive student skills. However, relatively little is known about whether smaller classes actually improve such skills. This section presents new evidence on this question by exploiting the unique features of a large, nationally representative longitudinal survey of students and teachers. In particular, this study identifies the effects of smaller classes on non-cognitive skills in specifications that control for unobserved traits specific to both students and teachers. As the class-size literature has generally recognized, student assignment to a class of a particular size is likely to reflect in part their unobserved propensity for achievement. In fact, both theory (Lazear, 2001) and previous evidence (West & Woessmann, 2006), which is confirmed here,

indicate that there is negative selection into smaller classes (i.e., students with a propensity for low achievement are more likely to be assigned to small classes). Another identification challenge that has been less widely acknowledged is that unobserved teacher quality is also likely to be related to class size. For example, an attentive principal might support struggling teachers (or placate an effective teacher) by allowing them to teach smaller classes. This study addresses these issues by exploiting linked student and teacher surveys across multiple subjects, which makes it possible to examine class-size effects conditional on both student and teacher fixed effects.

A second, policy-relevant contribution of the evidence based on these data is their comparative external validity. This is due in part to the national representativeness of these survey data. However, another key dimension to these results is that they provide information on the effects of class-size reductions in later grades. This is an important issue because state class-size reduction initiatives have been criticized for targeting multiple grade levels even though Project STAR provided evidence only on the effects of class-size reductions in grades K through 3 (Kim, 2007). Another relevant dimension to the grade-level issue is the claim that important non-cognitive skills are more malleable than cognitive skills for older students (Carneiro & Heckman, 2003; Heckman, 2000). A third benefit of the survey data analyzed here is that their longer term longitudinal components on educational attainment and labor market experiences make it possible to assess the cost-effectiveness of class-size reductions that improve non-cognitive student skills. The next three sections introduce the relevant data, specifications, and results and also discuss issues related to the possible remaining threats to the internal validity of these inferences.

National Education Longitudinal Study of 1988

NELS:88 is a nationally representative, longitudinal survey that began in 1988 with a sample of 24,599 eighth-grade students from more than 1,000 schools. The two-stage sampling design selected schools first and then approximately 26 students within each participating school (Ingels, Abraham, Karr, Spencer, & Frankel, 1990). This study is based on students from the 815 public schools that participated in the base-year sample. In addition to student surveys, NELS:88 also fielded surveys of teachers, administrators, and parents. The unique design of the teacher surveys is of particular relevance to this study's research design. For every participating student, two academic-subject teachers were surveyed (i.e., a math or science teacher and an English or history teacher). The teachers were selected by randomly assigning each school to one of the four possible subject pairings of math and science with English and history. Teachers provided information on themselves (e.g., certification, education, and experience) and the size of their sampled classes.

In combination, the teacher and student surveys in NELS:88 provide three types of student-outcome measures that are specific to each of the two academic subjects taught by sampled teachers. First, NELS:88 collected direct cognitive assessments based on subject tests completed by students. Second, both of the surveyed teachers provided their subjective assessment of the performance and behavior of each sampled student. For example, the teachers answered questions about whether a sampled student was frequently inattentive or disruptive in class. And, third, the student survey in NELS:88 solicited information on each student's intellectual engagement and effort with respect to each academic subject.

Our analysis exploits each type of outcome measure. The indicators drawn from student surveys are based largely on three questions that students were asked about their engagement in each of four academic subjects (i.e., math, science, English, and history). Specifically, students were asked to indicate their level of agreement with statements about whether the subject was not useful for their future, whether they did not look forward to the subject, and whether they were afraid to ask questions in their class on that subject. There were four possible categorical responses to these questions (i.e., *strongly agree*, *agree*, *disagree*, and *strongly disagree*). These responses were assigned values of 1 to 4 so that higher values implied lower levels of engagement, and they were standardized within subjects to create the variables NOTUSE, NOLKFD, and AFASK (see Table 1). The teacher perceptions of

TABLE 1

Sample Means, National Education Longitudinal Study of 1988 (NELS:88) Base-Year Sample

Variable	Description	Mean	Standard deviation	Sample size
NOTUSE	Subject not useful for my future	0.0018	0.9953	32,152
NOLKFD	Do not look forward to subject	0.0008	0.9945	32,246
AFASK	Afraid to ask questions in subject class	−0.0061	0.9936	32,197
DISRUPT	Student is frequently disruptive	0.1368	0.0019	33,018
INATT	Student is consistently inattentive	0.2255	0.0023	32,962
TRYH	Frequency of trying hard in subject (1st follow-up)	−0.0071	1.0012	18,612
STEST	Test score in subject	0.0219	0.9976	32,646
CLSSIZE	Class size	24.5067	5.8675	33,162
OTHRACE	Teacher of opposite race/ethnicity	0.3172	0.0025	33,802
OTHSEX	Teacher of opposite gender	0.5028	0.0025	33,802
SCERTIFD	Teacher certified by state in subject	0.8838	0.0017	33,802
PCTLEP	% classmates with limited English proficiency	0.0141	0.0718	31,362
SUBJECT1	English	0.2576	0.0024	33,802
SUBJECT2	History/social studies class	0.2424	0.0023	33,802
SUBJECT3	Mathematics class	0.2568	0.0024	33,802
SUBJECT4	Science class	0.2432	0.0023	33,802

individual students are based on binary indicators for whether they viewed a particular student as frequently disruptive or consistently inattentive (DISRUPT and INATT). The test score measure (STEST) is the cognitive assessment based on the subject for which a teacher was sampled and is standardized by subject (Table 1).

To examine the persistence of any effects of smaller classes, we also use a subject-specific, non-cognitive outcome reported by the subset of students participating in the 1990 follow-up survey (when most were in 10th grade). The student survey administered in 1990 did not include the same battery of questions as the base-year survey. However, it did include a closely related measure of student effort. Specifically, with respect to each of four academic subjects, participants in the first follow-up survey were asked, "How often do you try as hard as you can?" We numbered the five possible responses (which ranged from *never* to *almost every day*) 1 to 5 and standardized them separately within each subject to create the variable TRYH. These paired-subject data are available for more than 9,000 base-year students.[10]

These diverse NELS variables reflecting effort, motivation, self-confidence, and self-control clearly correspond to the broad array of work habits and behavioral traits commonly identified as non-cognitive skills (e.g., ter Weel, 2008). In fact, some of the same data and variables have been used in recent non-cognitive studies (e.g., Segal, 2008a). Recent studies in the parallel literature on school engagement have also used some of the same NELS variables and data (e.g., Glanville & Wildhagen, 2007). It is not surprising that we find that the variables we use tend to cluster into the behavioral engagement (i.e., DISRUPT and INATT) and the psychological engagement (i.e., NOTLF, NOTUSE, AFASK) constructs identified in this literature.[11]

Table 1 reports the means and standard deviations for these and other variables for which variation might exist after conditioning on student and teacher fixed effects. Other variables identify whether the student and teacher share the same race or gender, whether the teacher has state certification in the given subject, and the share of a student's classmates who have limited English proficiency. The base-year sample from NELS:88 includes 19,396 students from public schools. However, this sample is limited to 33,802 student-by-subject observations because two teacher questionnaires are available for only 16,901 of these students. More than half of the students for whom two teacher questionnaires are unavailable are also missing data on test scores. Students missing two teacher surveys are also more likely to be minorities and,
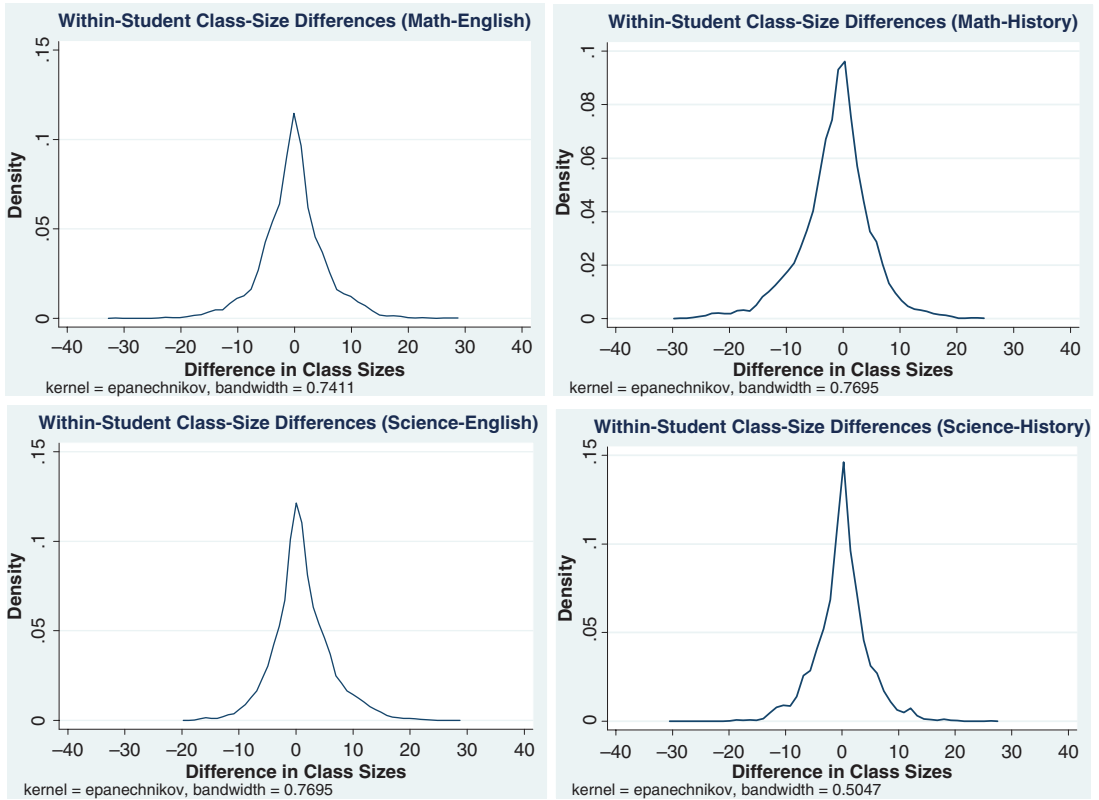
28

FIGURE 1. *Kernel density estimates, within-student class-size differences by academic subjects, National Education Longitudinal Study of 1988 (NELS:88).*

where test score data are available, are more likely to be lower achieving. Based on the prior class-size literature, which found that class-size reductions are more effective among disadvantaged students, we would expect this sample reduction to bias our results against finding larger class-size effects.

The average class size in this sample is 24.5 with a standard deviation of 5.9. However, our research design relies on the within-student variation in class sizes across contemporaneous academic subjects. Figure 1 presents kernel density estimates that illustrate the distribution of the within-student class-size differences for each of the four possible academic-subject pairings. In all four cases, these distributions are symmetrically distributed over zero and exhibit a plausible amount of variance. Specifically, fewer than 2% of the observations have within-student class-size differences larger than 15 in absolute value. And the results reported here are similar to those

based on models that exclude these outliers. The potential internal-validity concerns stemming from our reliance on this within-student variation are examined below.

First-Difference Specifications

The design of the NELS:88 surveys implies that each student-outcome measure is contemporaneously observed twice (i.e., once in each of two academic subjects) along with the corresponding class size. The matched-pairs nature of these data makes it possible to construct within-student comparisons that purge the influence of student-specific unobservables that are invariant across subjects (e.g., unobserved student traits that may influence class-size assignments). Furthermore, because teachers sometimes taught multiple classes that were part of the NELS:88 sample, it is also possible to condition on teacher fixed effects that reflect the unobserved

29

teacher quality that may also be correlated with class size.

In particular, assume that the math or science outcome observed for student i who is with teacher t (i.e., y_{1it}) is a function of observed student traits, $\mathbf{X}_i$, and the size of the student's class with teacher t (i.e., $SIZE_{1it}$):

$$y_{1it} = \alpha\mathbf{X}_i + \beta(SIZE_{1it}) + \lambda\mathbf{Z}_{1t} + \theta_{1t} + \mu_i + \varepsilon_{1it}. \tag{1}$$

In equation (1), the terms μ_i, θ_{1t}, and ε_{1it} are, respectively, a student fixed effect, a teacher fixed effect, and a mean-zero error term adjusted to accommodate school-level clustering. And the term $\mathbf{Z}_{1t}$ consists of the other observed determinants of y_{1it}, which vary at the level of the classroom and teacher. These variables include fixed effects for the subject of the class and other observed attributes of the teacher and the classroom. In a conventional cross-sectional study based on equation (1), it would be difficult to estimate β reliably because the error term in equation (1) would include confounding teacher and student effects (i.e., θ_{1t} and μ_i). However, the availability of a second contemporaneous observation makes it possible to estimate β conditional on student and teacher fixed effects. Suppose an equation like (1) applies to the student outcomes observed in English or history:

$$y_{2it} = \alpha\mathbf{X}_i + \beta(SIZE_{2it}) + \lambda\mathbf{Z}_{2t} + \theta_{2t} + \mu_i + \varepsilon_{2it}. \tag{2}$$

First differencing equations (1) and (2) yields the following:

$$(y_{1it} - y_{2it}) = \beta(SIZE_{1it} - SIZE_{2it}) + \lambda(\mathbf{Z}_{1t} - \mathbf{Z}_{2t}) + (\theta_{1t} - \theta_{2t}) + (\varepsilon_{1it} - \varepsilon_{2it}). \tag{3}$$

Estimates based on equation (3) identify the effects of class size conditional on all the subject-invariant determinants unique to individual students and teachers. However, these inferences could still be biased by subject-specific student traits as well as by unobserved classroom traits associated with class size. For example, our results would overstate the beneficial effects of smaller classes on the intellectual engagement of students if students with a tendency to like a particular subject were more likely to be assigned

to a smaller class in that subject. Similarly, if smaller class sizes were associated with important classroom traits (e.g., a lower share of peers with limited English proficiency), estimates based on equation (3) would overstate the benefits of smaller classes.

We address these concerns partly by examining the robustness of our results to conditioning on various observables (e.g., characteristics of classroom peers) in addition to student and teacher fixed effects. The pattern suggested by this evidence is generally one of negative selection into smaller classes. In particular, the pattern of selection on observables suggests that students with a propensity toward lower intellectual engagement with a particular academic subject are actually more likely to be assigned to a smaller class in that subject. These results imply that the inferences based on equation (3) would, if anything, imply a lower bound on the true non-cognitive benefits of class size reductions. We also examine the internal validity of estimates based on equation (3) in several other ways. For example, some of our specifications control for the possible influence of subject-specific propensities for good non-cognitive outcomes by conditioning on the student's test score in that subject. Although test scores are potentially endogenous with respect to class size, this specification provides a useful robustness check for our main results.

We also present evidence on whether small classes in one subject create empirically meaningful spillover benefits in closely related subjects. For example, we examine whether a lower class size in math appears to influence non-cognitive outcomes in science. This evidence is of interest mainly because it provides information on the nature of the educational production function. However, it also provides an indirect robustness check of our main results. More specific, some spillover effects of smaller classes might be expected. However, if the "other-subject" effects of smaller classes were large relative to the own-subject effect, it would suggest that students with a propensity to do well in related subjects (e.g., math and science) were simply more likely to be assigned to such classes. Alternatively, the existence of even modest spillover effects could imply that our research design understates the true benefits of smaller classes.

That is, our within-student comparisons would understate the effects of a smaller class in a particular subject if that smaller class also improved student outcomes in another subject. However, we suspect that this is not an important concern both because of the other-subject results we report and because the sampling design for the teacher surveys in NELS:88 always paired disparate academic subjects (i.e., math and science were paired with either English or history).

Baseline Results

Table 2 presents the estimated effects of class size on the non-cognitive and cognitive student outcomes and across different specifications. The results in column (1) are based on a specification that includes several student, teacher, and classroom observables (e.g., race, gender, socioeconomic status, teacher experience, etc.) as well as school fixed effects. The results of this within-school specification suggest that smaller class sizes actually reduce student test scores, increase the perceived disruptiveness and inattentiveness of students, and lower their levels of academic engagement. However, the subsequent first-difference (FD) estimates indicate that these counterintuitive results reflect the non-random sorting of students (and, to a lesser extent, teachers) to classrooms of different sizes.

More specific, the most basic FD specification (i.e., column 2) suggests that smaller classes reduce the extent to which students do not look forward to a subject, do not see it as useful for their future, and are afraid to ask questions. Similarly, smaller classes reduce the chance that a given student is inattentive (although not disruptive). Smaller classes also appear to increase student test scores, although the effect size is quite small (i.e., $.0022 \times 5.87 = .013$) and statistically insignificant.

The third specification in Table 2 introduces teacher fixed effects (more specific, teacher fixed effects specific to math–science and English–history subject pairings). It is interesting that the introduction of these controls increases the R^2 in these regressions quite dramatically. More important, the estimated effects of class size on NOLKFD, NOTUSE, and AFASK increase substantially after introducing teacher fixed effects. The apparent bias relative to the prior specification

is consistent with students who have poor academic engagement with a subject being more likely to be assigned to a relatively small class and a teacher who is particularly effective at promoting engagement in that subject. However, the estimated effect of class size on INATT falls somewhat in this specification and becomes statistically insignificant (*p* value = .107). This pattern of results is similar in specifications that introduce controls for teacher and classroom observables (i.e., PCTLEP, OTHRACE, OTHSEX, and SCERTIFD).

Robustness Checks

Overall, these results indicate that the benefits of smaller classes for eighth graders are concentrated in their effects on the three student-reported measures of emotional engagement. The effect sizes implied by these point estimates range from roughly .05 to .08. Yet, there are several reasons that these modest effects could actually overstate the benefits of smaller classes. For example, all of our first-difference models condition on student fixed effects that are, by assumption, invariant across subjects. In a situation where students who are likely to have high degrees of engagement in a particular subject are more likely to be assigned to smaller classes in that subject, the estimated benefits of smaller classes would be biased upward. Similarly, the apparent benefits of smaller classes would be misleading if smaller classes were associated with classroom traits such as higher quality peers.

However, several types of evidence suggest that the estimates reported in Table 2 do not overstate the non-cognitive benefits of smaller classes (and may, in fact, understate them). First, the estimated effects of class size on NOLKFD, NOTUSE, and AFASK are robust in a specification that introduces STEST, a subject-specific (and endogenous) variable as a control. Second, the comparative results across the specifications in Table 2 actually suggest a pattern of negative selection into smaller classes. More specific, models that include weaker student and teacher controls suggest that smaller class sizes have smaller or even negative benefits. This pattern of selection on observables implies that students with a propensity for worse non-cognitive outcomes are more likely to be assigned to smaller

31

TABLE 2

National Education Longitudinal Study of 1988 (NELS:88): Estimated Effects of Class Size on Non-Cognitive and Cognitive Student Outcomes

| | | | First-difference estimates | | | | | | | |
| | (1) | | (2) | | (3) | | (4) | | (5) | |
Dependent variable	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2
NOLKFD	0.0023 (0.0016)	0.0888	0.0056‡ (0.0020)	0.0166	0.0117‡ (0.0032)	0.3366	0.0126‡ (0.0034)	0.3470	0.0112‡ (0.0034)	0.3626
NOTUSE	−0.0053‡ (0.0016)	0.0649	0.0039† (0.0018)	0.0142	0.0085‡ (0.0031)	0.2664	0.0095‡ (0.0033)	0.2738	0.0086† (0.0034)	0.2788
AFASK	−0.0031* (0.0017)	0.0612	0.0081‡ (0.0017)	0.0024	0.0140‡ (0.0030)	0.2723	0.0152‡ (0.0032)	0.2834	0.0142‡ (0.0033)	0.2906
DISRUPT	−0.0040‡ (0.0006)	0.1123	−0.0010 (0.0007)	0.0005	−0.0005 (0.0011)	0.3212	−0.0003 (0.0012)	0.3305	−0.0005 (0.0012)	0.3342
INATT	−0.0034‡ (0.0007)	0.1129	0.0034‡ (0.0009)	0.0029	0.0021 (0.0013)	0.3481	0.0017 (0.0014)	0.3513	0.0016 (0.0014)	0.3562
STEST	0.0246‡ (0.0019)	0.2965	−0.0022 (0.0014)	0.0237	−0.0029 (0.0020)	0.2928	−0.0018 (0.0021)	0.2996	n/a	
Sample sizes (range)	29,724–31,140		15,478–15,911		15,478–15,911		13,865–14,586		13,394–14,035	
Controls										
Student observables	x									
School fixed effects	x									
Student fixed effects			x		x		x		x	
Teacher fixed effects					x		x		x	
Teacher/classroom observables	x						x		x	
Subject test score									x	

Note. Each reported coefficient is from a separate regression. Standard errors, adjusted for school-level clustering, are reported in parentheses. All models include gender-specific subject fixed effects.

*$p < 0.1$; †$p < 0.05$; ‡$p < 0.01$.

TABLE 3

National Education Longitudinal Study of 1988 (NELS:88): Selection on Classroom and Teacher Observables

Dependent variable	(1)	First-difference estimates			
		(2)	(3)	(4)	(5)
PCTLEP	−0.0010‡ (0.0002)	−0.0010‡ (0.0003)	−0.0002 (0.0002)	−0.0002 (0.0002)	−0.0002 (0.0002)
SCERTIFD	0.0070‡ (0.0009)	0.0037‡ (0.0014)	n/a	n/a	n/a
Novice teacher (1–3 years experience)	−0.0001 (0.0007)	−0.0002 (0.0012)	n/a	n/a	n/a
Controls					
School fixed effects	x				
Student fixed effects		x	x	x	x
Teacher fixed effects			x	x	x
Teacher/classroom observables	x			x	x
Subject test score					x

Note. Standard errors, adjusted for school-level clustering, are reported in parentheses. All models include gender-specific subject fixed effects.
*Statistically significant at the 10% level. †Statistically significant at the 5% level. ‡Statistically significant at the 1% level.

classes. The existence of negative selection into smaller classes implies that, to the extent that these inferences are biased, they understate the true non-cognitive benefit of smaller classes.

Table 3 presents further evidence on this point by reporting the estimated effects of class size in auxiliary regressions where PCTLEP, SCERTIFD, and a binary measure for novice teachers (i.e., 1 to 3 years of experience) are the dependent variables. The results from models that condition on school or student fixed effects indicate that smaller class sizes imply a statistically significant increase in the percentage of classroom peers who have limited English proficiency and a statistically significant decrease in the likelihood of having a teacher who is state-certified in the given subject. In models that condition on teacher fixed effects, this pattern of selection on observables becomes smaller and statistically insignificant with respect to PCTLEP and is not defined with respect to the teacher traits.

Spillover Effects and Persistence

An implied assumption of our FD research design is that the benefits of a small class in one subject do not have empirically meaningful implications for outcomes in other subjects. As noted above, this assumption may be a reasonable one

because the sampling design for the teacher survey implies that observations specific to math and science classes are always paired with observations of either English or history classes. However, whether class-size reductions in one academic subject create benefits in more closely related academic subjects is an interesting and policy-relevant question. We examined this issue directly by estimating the effect of class size in a particular subject (i.e., class size interacted with subject-specific fixed effects) on the outcomes in a related but different subject. More specific, we estimated specifications where the non-cognitive outcome in one academic subject was replaced by the corresponding outcome from a related academic subject.

Table 4 presents the key results from this exercise and focuses on one of the three academic-engagement indicators, AFASK.[12] The baseline model reports the results of a specification where the dependent variables were unaltered (i.e., own-subject effects). Those results indicate that the estimated effect of class size on AFASK is largest in math and English. However, the hypothesis that these four coefficients are equal cannot be rejected. The remaining results in Table 4 suggest that small classes in one academic subject led to benefits in closely related academic subjects (i.e., the four key estimates are all positive).

33

TABLE 4

National Education Longitudinal Study of 1988 (NELS:88): Estimated Effects of Class Size on AFASK by Academic Subject

Independent variable	Baseline measures	Change in dependent variable			
		Math AFASK replaced by science	Science AFASK replaced by math	Reading AFASK replaced by history	History AFASK replaced by reading
Class size in math	$0.0153^{\ddagger}$ (0.0051)	0.0024 (0.0044)	$0.0145^{\ddagger}$ (0.0051)	$0.0145^{\ddagger}$ (0.0051)	$0.0109^{\dagger}$ (0.0051)
Class size in science	0.0083^{*} (0.0048)	0.0082^{*} (0.0048)	0.0064 (0.0053)	0.0045 (0.0051)	0.0104^{*} (0.0053)
Class size in English	$0.0201^{\ddagger}$ (0.0053)	$0.0171^{\ddagger}$ (0.0050)	$0.0170^{\ddagger}$ (0.0054)	0.0104^{*} (0.0055)	$0.0103^{\ddagger}$ (0.0053)
Class size in history	$0.0115^{\dagger}$ (0.0052)	$0.0124^{\dagger}$ (0.0051)	0.0069 (0.0060)	$0.0111^{\dagger}$ (0.0052)	0.0040 (0.0051)
p value (H_0: $\beta_M = \beta_S = \beta_E = \beta_H$)	0.3669	0.0920	0.3968	0.5666	0.1998

Note. Standard errors, adjusted for school-level clustering, are reported in parentheses. All models include gender-specific subject fixed effects and student and teacher fixed effects.
*Statistically significant at the 10% level. †Statistically significant at the 5% level. ‡Statistically significant at the 1% level.

However, these estimated effects are all relatively small and, in 3 of the 4 cases, statistically insignificant. The only exception is that a smaller English class implies a relatively small but weakly significant increase in school engagement with history. Overall, these results imply that the noncognitive benefits of smaller classes are largely concentrated in the particular subject taught with a smaller class size.

In addition to providing evidence on the nature of the educational production function, these results also provide a useful ad hoc falsification exercise for the basic FD identification strategy employed in this study. In particular, if the other-subject effects of class size had been comparatively large, it would have underscored concerns about whether students with a propensity for good non-cognitive outcomes in a broad subject area (e.g., math and science) are more likely to be assigned to smaller classes in those subjects. Instead, the results in Table 4 suggest that, for all four academic subjects, the spillover effects to related subjects are relatively small. Like the prior robustness checks, this pattern implies that non-random, within-student selection into smaller classes is not confounding our results.

Although reassuring with respect to the internal validity of our main results, the subject-specific nature of the class-size effects also raises the question of whether the apparent effects of

eighth-grade class size on engagement persist over time or whether they simply reflect the interaction between classroom environments and fixed student traits. It is worth noting that even transient effects on school engagement could have policy relevance, to the extent that our measures are, in fact, instrumental to subsequent academic success. However, the interpretation of these effects on school engagement as a form of "skill" development would clearly be strengthened if the subject-specific effects were to persist over time.

It is fortunate that the subject-specific questions about the frequency of trying hard (TRYH), which were asked of students participating in the first follow-up study, allow us to address this question. Table 5 reports the key results from specifications that estimate the effect of subject-specific class sizes in eighth grade on these longer term measures. The basic FD specification suggests that a smaller class size in an academic subject during eighth grade implies a statistically significant increase in effort in that subject 2 years later. The effect size implied by this point estimate (.032) is smaller than the effect size for contemporaneous grade 8 measures. In models that introduce teacher fixed effects as well as other controls (e.g., OTHRACE, OTHSEX, SCERTIFD, and PCTLEP), this effect is somewhat larger but becomes weakly significant because

34

TABLE 5

Estimated Effects of Grade 8 Class Size on Subsequent Effort, National Education Longitudinal Study of 1988 (NELS:88) First Follow-Up Survey

Specification	$\hat{\beta}$	R^2	Sample size
Student fixed effects, gender-specific, subject fixed effects	−0.0054‡ (0.0019)	0.0032	9,046
Previous model and teacher fixed effects	−0.0061* (0.0035)	0.3181	9,046
Previous model and teacher & classroom observables	−0.0067* (0.0037)	0.3218	8,174
Previous model and subject test scores	−0.0065* (0.0039)	0.3275	7,898

Note. The dependent variable is TRYH, a standardized measure for the frequency of student-reported effort in an academic subject during the first follow-up interview. Standard errors, adjusted for school-level clustering, are reported in parentheses. *Statistically significant at the 10% level. †Statistically significant at the 5% level. ‡Statistically significant at the 1% level.

of a large increase in the standard error. However, this weakly significant result is also robust to conditioning on subject-specific test scores from the base year.

Treatment Heterogeneity

Overall, the results based on the NELS:88 data suggest that assignment to a smaller class improves several of the non-cognitive measures (i.e., NOTUSE, NOLKFD, and AFASK) and that these results cannot be explained by the presence of confounding student or classroom unobservables. In fact, the pattern of selection is such that these results may actually understate the true non-cognitive benefits of smaller classes. The results with respect to teacher observations (i.e., DISRUPT and INATT) and cognitive scores (i.e., STEST), on the other hand, were less dispositive.

All of these results were based on the full analytical sample of NELS:88 eighth graders and the implicit assumption of a common treatment effect for different types of students and schools. However, there is a variety of reasons to suspect that the effects of class size might differ across particular types of students and educational settings. Table 6 presents evidence on this issue by presenting the estimated effects of class size on each of the non-cognitive and cognitive measures for samples defined by various student and school traits.

Several aspects of these results are worth underscoring. For example, these results imply that a decrease of one standard deviation in boys' class sizes would reduce the probability

that a boy is viewed as frequently inattentive by 3.5 percentage points (i.e., $6.0 \times .0058$), a reduction of roughly 13% relative to the gender-specific mean. Similarly, a one standard deviation decrease in the class sizes of Hispanic students would reduce the probability that a Hispanic student is seen as disruptive by 6.5 percentage points (i.e., $6.0 \times .0109$), a reduction of roughly 38% relative to the Hispanic-specific mean. The estimated effect of class size on subject-specific test scores is statistically significant among girls and in urban schools with effect sizes of .037 and .067, respectively. The estimated effects of class size on the student-engagement measures (i.e., NOLKFD, NOTUSE, and AFASK) also differ across the subsamples. For example, the class-size effects on these outcomes are particularly large in urban schools. However, it should also be noted that these distinctions are in most cases small relative to the sampling variation.

Cost-Benefit Considerations

Our NELS:88 analysis indicates that class-size reductions in the eighth grade lead to statistically significant improvements in several non-cognitive outcomes (i.e., NOLKFD, NOTUSE, and AFASK). Furthermore, the educational gains from class-size investments appear to be larger and more extensive in certain targeted settings (e.g., urban schools). However, class-size reductions also involve costly, upfront expenditures. Whether these benefits justify further expenditures is, in large part, an empirical question. In this section, we present some qualified evidence on the relevant cost-benefit comparisons.

35

TABLE 6
National Education Longitudinal Study of 1988 (NELS:88): Class Size Effects by Student and School Traits

Dependent variable	Boys	Girls	Black	Hispanic	Low SES	High SES	Urban	Suburban	Rural
NOLKFD	0.0122† (0.0059)	0.0110† (0.0054)	0.0124 (0.0145)	0.0161 (0.0121)	0.0139‡ (0.0053)	0.0125† (0.0054)	0.0195‡ (0.0069)	0.0085* (0.0052)	0.0110† (0.0048)
NOTUSE	0.0099* (0.0058)	0.0098* (0.0054)	0.0224* (0.0134)	0.0067 (0.0117)	0.0054 (0.0051)	0.0123† (0.0054)	0.0138† (0.0070)	0.0089* (0.0051)	0.0053 (0.0047)
AFASK	0.0102† (0.0051)	0.0134† (0.0053)	0.0270† (0.0158)	0.0267† (0.0135)	0.0169‡ (0.0051)	0.0151‡ (0.0048)	0.0145† (0.0068)	0.0116† (0.0045)	0.0159‡ (0.0051)
DISRUPT	0.0013 (0.0020)	−0.0011 (0.0015)	−0.0038 (0.0053)	0.0109† (0.0054)	−0.0014 (0.0019)	−0.0008 (0.0018)	−0.0014 (0.0028)	0.0008 (0.0016)	−0.0014 (0.0017)
INATT	0.0047† (0.0023)	−0.0007 (0.0019)	0.0071 (0.0066)	0.0094 (0.0059)	0.0021 (0.0024)	0.0027 (0.0020)	0.0008 (0.0031)	0.0013 (0.0018)	0.0038* (0.0023)
STEST	−0.0003 (0.0036)	−0.0065† (0.0033)	−0.0114 (0.0077)	−0.0085 (0.0080)	−0.0028 (0.0029)	−0.0051 (0.0037)	−0.0117† (0.0045)	−0.0002 (0.0033)	−0.0009 (0.0029)

Note. Standard errors, adjusted for school-level clustering, are reported in parentheses. All models include gender-specific subject fixed effects, student fixed effects, and teacher fixed effects. SES = socioeconomic status.

*Statistically significant at the 10% level. †Statistically significant at the 5% level. ‡Statistically significant at the 1% level.

Non-Cognitive Skills and Long-Term Outcomes

The longitudinal nature of NELS:88 makes it possible to examine the long-term consequences of improvements in cognitive and non-cognitive skills as measured in the eighth grade. The fourth follow-up interview of NELS:88 participants, which elicited information on both educational attainment and early labor-market experiences, occurred in 2000, when participants were approximately 26 years old. To gauge the possible benefits of eighth-grade class-size reductions, we examine the effects of the eighth-grade non-cognitive and cognitive skill measures (standardized and averaged across all four subjects) on these outcomes. This type of correlational evidence raises important identification problems, which, as in similar studies, are not addressed here. However, our analysis does improve upon much of the prior evidence by conditioning on school fixed effects. Furthermore, comparing the results across specifications that introduce additional controls can provide evidence on the direction of selection on unobservables.

The fourth follow-up interview included 12,144 participants. However, the exclusion of those who were not base-year participants from public schools and those for whom base-year cognitive and non-cognitive data are unavailable reduces the sample size to approximately 8,300. Our results condition both on measures of student observables (i.e., dummy variables for gender, race, ethnicity, and age) and on dummy variables that identify a variety of family traits. The family measures consist of unrestricted dummy variables for multiple categories of family composition (7 categories), family size (10), parental education (8), family income (16), and language-minority status (2). Our measures of educational attainment consist of dummy variables for high-school completers (excluding GED completers), matriculants at 4-year colleges, and those who have completed bachelor's degrees.

Table 7 presents the estimated effect of each non-cognitive measure on educational attainment in specifications that also condition on STEST. Overall, these results suggest that both cognitive and non-cognitive skills have statistically significant effects on educational attainment. However, the effect sizes associated with the non-cognitive measures are smaller than those associated with the cognitive measure, in particular for college entrance and completion. For example, a one standard deviation increase in NOLKFD implies that the probability of completing a bachelor's degree falls by 3.4 percentage points. However, a one standard deviation decrease in STEST reduces the probability of completing a bachelor's degree by 16.7 percentage points. It is interesting that one of the measures (AFASK) has a somewhat counterintuitive but weakly significant effect on high school graduation. More specific, students who are more afraid to ask questions in their academic classes were actually more likely to graduate from high school (although less likely to enter or complete college).

A notable feature of the results in Table 7 is that the estimated effects of STEST tend to decrease after conditioning on family observables and school fixed effects. However, the estimated effects of the non-cognitive measures tend to grow in absolute value after conditioning on these controls. This pattern of selection on observables suggests that these results, if anything, understate the effects of the non-cognitive measures on educational attainment. These results may also understate the effects of non-cognitive skills to the extent that the low-stakes test measure included in these regressions also reflects non-cognitive skills (e.g., Segal, 2008b). The first specification in Table 6 indicates that the estimated effects of the non-cognitive measures are larger in models that exclude the cognitive measure.

In Table 8, we present evidence on how the cognitive and non-cognitive eighth-grade measures are related to labor market outcomes as reported in the fourth follow-up survey. Our first labor-market outcome is a binary indicator for whether the participant reports that he or she was engaged in full-time employment in 1999. This variable is defined for the roughly 5,600 participants who were not students (i.e., those who did not attend a postsecondary institution after January 1999) and who reported data on hours and weeks worked. We define full-time employment as having worked 40 or more weeks and 35 or more hours in a typical week. Roughly 80% of participants met this definition of full-time employment. Our second labor-market outcome is the natural log of reported

37

TABLE 7

Estimated Effect of Non-Cognitive and Cognitive Measures on Educational Attainment, National Education Longitudinal Study of 1988 (NELS:88) Fourth Follow-Up

Independent variable	High school graduate			College entrant			Bachelor's degree		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
NOLKFD	−0.0209‡ (0.0057)	−0.0153‡ (0.0057)	−0.0190‡ (0.0062)	−0.0478‡ (0.0077)	−0.0304‡ (0.0071)	−0.0341‡ (0.0077)	−0.0393‡ (0.0071)	−0.0247‡ (0.0065)	−0.0337‡ (0.0072)
STEST	—	0.0659‡ (0.0043)	0.0618‡ (0.0044)	—	0.2078‡ (0.0061)	0.2050‡ (0.0069)	—	0.1707‡ (0.0060)	0.1669‡ (0.0069)
NOTUSE	−0.0258† (0.0054)	−0.0165‡ (0.0053)	−0.0175‡ (0.0058)	−0.0740‡ (0.0070)	−0.0448‡ (0.0065)	−0.0447‡ (0.0071)	−0.0552‡ (0.0060)	−0.0311‡ (0.0057)	−0.0349‡ (0.0063)
STEST	—	0.0648‡ (0.0043)	0.0607‡ (0.0048)	—	0.2042‡ (0.0061)	0.2012‡ (0.0069)	—	0.1685‡ (0.0060)	0.1645‡ (0.0069)
AFASK	−0.0038 (0.0046)	0.0088* (0.0046)	0.0089* (0.0051)	−0.0581‡ (0.0066)	−0.0200‡ (0.0064)	−0.0229‡ (0.0071)	−0.0457‡ (0.0058)	−0.0140‡ (0.0056)	−0.0165‡ (0.0064)
STEST	—	0.0684‡ (0.0043)	0.0647‡ (0.0049)	—	0.2062‡ (0.0063)	0.2033‡ (0.0071)	—	0.1698‡ (0.0062)	0.1663‡ (0.0071)
Dependent mean		0.87			0.51			0.30	
Controls									
Student observables	x	x	x	x	x	x	x	x	x
Family observables	x	x	x	x	x	x	x	x	x
School fixed effects			x			x			x

Note. Standard errors, adjusted for school-level clustering, are reported in parentheses.

*Statistically significant at the 10% level. †Statistically significant at the 5% level. ‡Statistically significant at the 1% level.

TABLE 8

Estimated Effects of Non-Cognitive and Cognitive Measures on Labor-Market Outcomes, National Education Longitudinal Study of 1988 (NELS:88) Fourth Follow-Up

Independent variable	Full–time employment (1999)				ln (1999 earnings)			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
NOLKFD	−0.0217‡ (0.0081)	−0.0179† (0.0081)	−0.0269‡ (0.0094)	−0.0223† (0.0095)	−0.0273† (0.0133)	−0.0210 (0.0134)	−0.0216 (0.0152)	−0.0144 (0.0151)
STEST	—	0.0457‡ (0.0068)	0.0368‡ (0.0079)	0.0156* (0.0087)	—	0.0905‡ (0.0099)	0.0859‡ (0.0124)	0.0414‡ (0.0139)
NOTUSE	−0.0160† (0.0077)	−0.0103 (0.0078)	−0.0130 (0.0090)	−0.0077 (0.0091)	−0.0086 (0.0108)	0.0022 (0.0109)	0.0051 (0.0135)	0.0155 (0.0134)
STEST	—	0.0456‡ (0.0069)	0.0370‡ (0.0079)	0.0158* (0.0088)	—	0.0917‡ (0.0099)	0.0876‡ (0.0124)	0.0429‡ (0.0139)
AFASK	−0.0168† (0.0073)	−0.0093 (0.0073)	−0.0077 (0.0086)	−0.0075 (0.0088)	−0.0631‡ (0.0121)	−0.0495‡ (0.0122)	−0.0541‡ (0.0158)	−0.0512† (0.0158)
STEST	—	0.0453‡ (0.0069)	0.0374‡ (0.0079)	0.0153* (0.0087)	—	0.0844‡ (0.0099)	0.0791‡ (0.0128)	0.0345† (0.0145)
Dependent mean		0.84				10.19		
Controls								
Student observables	x	x	x	x	x	x	x	x
Family observables	x	x	x	x	x	x	x	x
School fixed effects			x	x			x	x
Educational attainment				x				x

Note. Standard errors, adjusted for school-level clustering, are reported in parentheses.

*Statistically significant at the 10% level. †Statistically significant at the 5% level. ‡Statistically significant at the 1% level.

39

employment earnings for 1999. This measure is defined for the roughly 4,100 participants who had full-time employment in 1999 and who responded to the earnings question. An average hourly wage can be imputed using the earnings data and the data on hours and weeks worked, and results based on this measure are similar to those reported here. However, we report the results based on the annual earnings measure because it has less measurement error (Deke & Haimson, 2006; Segal, 2008b).

The results in Table 8 indicate that participants with worse non-cognitive skills in the eighth grade (i.e., higher levels of NOLKFD, NOTUSE, and AFASK) are less likely to have been employed full-time in 1999. However, only the effect associated with NOLKFD is statistically significant after controlling for the measure of cognitive skills. A one standard deviation decrease in NOLKFD implies that the probability of full-time employment as a young adult increased by 2.7 percentage points (i.e., roughly 3.2% of the mean). It is interesting that this point estimate changes relatively little after conditioning on measures of educational attainment, suggesting that these non-cognitive skills have labor-market consequences that are independent of their schooling effects. Lower levels of NOLKFD, NOTUSE, and AFASK are also associated with higher earnings. However, only the effect associated with AFASK is statistically significant after conditioning on the cognitive test-score measure. A one standard deviation decrease in AFASK implies earnings that are approximately 5.4% higher. As with the effects of NOLKFD on employment, the estimated effect of AFASK on earnings is similar after controlling for educational attainment.

Comparing Costs and Benefits

The results in Tables 7 and 8 indicate that the non-cognitive skills most clearly shaped by exposure to smaller classes are highly predictive of subsequent educational attainment and may also generate some targeted labor-market benefits among young adults. However, whether these benefits justify class-size reductions is not clear. Investments in smaller classes involve costly upfront expenditures but generate benefits that are realized only over the subsequent years. We

provide some evidence on this issue by using our NELS:88 results to compare the costs and benefits of reducing class sizes in the eighth grade. These comparisons necessarily involve a number of important assumptions and caveats, which we discuss after presenting our basic results. The normative interpretation of our cost-benefit comparisons appears sensitive to reasonable differences in the relevant parameters and assumptions. Nonetheless, we view this qualified evidence as policy relevant because it suggests whether class-size reductions appear remotely cost-effective and underscores some of the key issues relevant to understanding this issue more clearly.

First, we estimated the per-pupil cost of a one standard deviation decrease in class size as $3,392 in 2006 dollars. To construct this estimate, we first noted, using the NELS:88 data in Table 5, that a one standard deviation decrease in class sizes would increase the number of classes by 31%. Following Krueger (2003a), we assumed that the cost of a class-size reduction would be proportional to expenditures per pupil. We estimated expenditures per pupil in 2006 dollars ($10,774) by taking the 2002–2003 expenditures per pupil in public schools and adjusting for inflation. Our estimate of the direct per-pupil cost of a one standard deviation class-size reduction is then simply 31% of this estimate.

To construct a comparable estimate of the monetized benefits of an eighth-grade class-size reduction, we calculated the present discounted value of the increased earnings implied by this investment. In particular, we focused on the AFASK indicator, which appears to have had the clearest effect on earnings. More specific, using the point estimate from model (3) in Table 2, a one standard deviation decrease in class size would reduce AFASK by .089 (i.e., .014 × −5.8675). Using the estimate from column (3) of the earnings results panel in Table 8, this decrease in AFASK implies that earnings would grow by 0.48% (i.e., −.0541 × −.089). As in Krueger (2003a), we assumed that this earnings effect would exist from age 18 to 65. To calculate the present discounted value of this earnings increase, we identified employment earnings by year of age for members of the civilian labor force, aged 18 to 65, who responded to the March 2007 Current Population Survey (CPS).
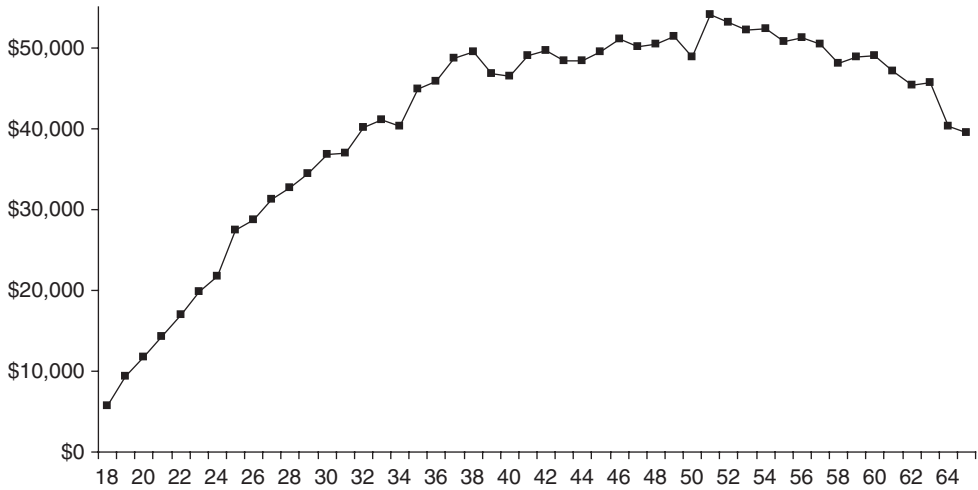
FIGURE 2. *Age-earnings profile—March 2007 Current Population Survey.*

TABLE 9

Present Discounted Value of Increased Earnings From Reducing Eighth-Grade Class Size by One Standard Deviation

Discount rate	Assumed productivity growth rate		
	0%	1%	2%
0.02	$5,167	$6,941	$9,433
0.05	$2,376	$3,060	$3,986
0.08	$1,247	$1,548	$1,548
0.11	$727	$876	$1,065
Internal rate of return	0.036	0.046	0.056

Note. The estimated increase in earnings is based on the age-earnings profile of labor-force participants from the 2007 March CPS, the estimated effect of a one standard deviation class-size decrease on AFASK (Table 6, column 4) and the estimated effect of AFASK on earnings (Table 11, column 3). The direct cost of a one standard deviation class-size reduction is estimated as $3,392 in 2006 dollars.

This age-earnings profile is represented in Figure 2. We then calculated the present discounted values of a 0.48% increase in earnings under different assumptions about the discount rate and the productivity-related growth in earnings. These increased earnings are assumed to begin 5 years after the class-size investment (i.e., at age 18).

Table 9 presents the results. The increased earnings implied by the class-size reduction exceed the cost of this reduction only for lower values of the discount rate or more generous assumptions about productivity growth. For example, assuming a 5% discount rate and 1% productivity growth, the present discounted value of the increased earnings is $3,060, roughly $300 more than the cost. The internal rate of return

(i.e., the discount rate that would equate the present discounted value of costs and benefits) provides a useful way to summarize the results.[13] The internal rates of return for this class-size investment range from 3.6% to 5.6%, depending on the assumed productivity growth (i.e., 0% to 2%).

The results in Table 6 suggest that targeted investments in class-size reductions may be more unambiguously cost-effective. In particular, this could be so for urban schools where class-size reductions appear to improve both cognitive and non-cognitive skill measures. We estimated the cost of a one standard deviation reduction in urban class sizes at $3,157 in 2006 dollars. This estimate reflects an upward adjustment in costs to reflect the higher costs per pupil in urban schools

41

TABLE 10

Present Discounted Value of Increased Earnings From Reducing Eighth-Grade Class Size by One Standard Deviation in Urban Schools

Discount rate	Assumed productivity growth rate		
	0%	1%	2%
0.02	$10,423	$14,003	$19,031
0.05	$4,793	$6,173	$8,042
0.08	$2,515	$3,123	$3,916
0.11	$1,467	$1,768	$2,149
Internal rate of return	0.069	0.079	0.090

Note. The estimated increase in earnings is based on the age-earnings profile of labor-force participants from the March 2007 Current Population Survey, the estimated effect of a one standard deviation class-size decrease on AFASK and STEST (Table 9) and the estimated effect of AFASK and STEST on earnings (Table 11, column 3). The direct cost of a one standard deviation class-size reduction is estimated as $3,157 in 2006 dollars.

as well as the fact that the standard deviation for class size is smaller among the urban schools in NELS:88 (i.e., 5.69).[14] Using the results from Tables 6 and 8, we estimated that a one standard deviation class-size reduction in eighth grade would increase earnings by 0.97%. That estimated increase reflects the effects of the class-size reduction on both AFASK and STEST. Table 10 presents the present discounted value of this earnings increase under different assumptions about the discount rate and productivity growth. It is not surprising that the urban-specific results suggest that a class-size investment appears cost-effective under a broader range of assumptions. For example, assuming a 5% discount rate and 1% productivity growth, the benefit from the class-size investment (i.e., $6,173) is nearly twice its estimated cost. Stated differently, the internal rate of return for a class-size investment is 7.9% under the assumption of 1% productivity growth.

Overall, these results suggest that the apparent cost-effectiveness of an eighth-grade class-size reduction is sensitive to whether the investment is targeted where it would appear to be most effective (e.g., urban schools) and to reasonable disagreements about how to compare costs and benefits (e.g., the relevant discount rate). For example, Krueger (2003c) and Summers (2003) discussed whether the appropriate benchmark for an investment of this sort should be the long-term real interest rate on government bonds, the average real return on the stock market, or the pre-tax profit rate. Other substantive issues complicate a comparison of costs and benefits

even further. For example, the estimated direct cost of a class-size reduction would understate the true cost of this investment to the extent that the tax mechanisms used to raise this revenue generate deadweight loss (Summers, 2003). Furthermore, these cost-benefit comparisons also ignored the possible general-equilibrium consequences of a broad investment in smaller classes. In particular, a pervasive effort to reduce class sizes might be compromised, at least in the short term, by rising salaries, lower quality teachers, and inadequate facilities. However, it should also be noted that the benefit calculations may understate the true benefits of class-size investments because they ignored any positive externalities (e.g., through improved civic engagement and reductions in criminal behavior).[15] Finally, an additional uncertainty is that our estimates of the effect of a class-size reduction (e.g., Tables 2 and 6) turn on an identification strategy that compares a student contemporaneously across two academic subjects with different class sizes. However, this source of variation could conceivably overstate or understate the true effects of a class-size reduction across multiple academic subjects.

Conclusions

The prevalence of class-size reduction policies in public education is a powerful testament to their popular appeal. However, the research base has provided more limited and sometimes conflicting evidence on the likely cost-effectiveness of broad class-size reductions. This study addressed one of

the most important gaps in this literature by examining the effects of class size on non-cognitive student outcomes that appear to have important educational and labor-market implications.

Our quasi-experimental analysis of nationally representative data on eighth graders indicates that reductions in class size are associated with improvements in the available measures of non-cognitive skills related to psychological, but not behavioral, engagement with school. Evidence of negative selection into smaller classes based on observed student characteristics suggests that, if anything, our estimates understate the true non-cognitive benefit of smaller classes. Furthermore, we find qualified evidence that eighth-grade class-size reductions may be cost-effective, in light of the apparent long-term labor-market benefits of these non-cognitive skills. Although our cost-benefit comparisons are sensitive to their underlying assumptions, it is notable that eighth-grade class-size reductions appear to be particularly cost-effective when targeted in urban schools.

The relative cost-effectiveness of targeted class-size reductions is worth underscoring in light of the concern that broad initiatives to reduce class sizes may be implemented in haphazard ways and have implications for teacher quality that are not captured by these results. For example, a 1996 California policy that provided financial incentives for school districts statewide to reduce class size resulted in a dramatic increase in the share of novice and unqualified teachers and widespread facilities shortages (Schrag, 2007). Jepsen and Rivkin (2002) provided evidence that class-size reduction improved third-grade test scores but that these benefits were offset in high-minority schools by a decline in teacher quality. Targeting class-size policies to schools most likely to benefit would reduce the possibility of negative unintended consequences.

Our analysis also adds to the growing literature indicating that non-cognitive skills matter for subsequent academic and labor-market success. Taken as a whole, this body of evidence strongly suggests that policymakers and researchers should consider ways to encourage schools to promote these skills. The results we have presented here imply that targeted class-size reductions are one viable policy lever. In contrast, accountability-style policies that reward or sanction schools explicitly based on the types of teacher- and student-reported measures of non-cognitive skills that we have examined here would, in all likelihood, perform poorly because they would be easy to game.

However, this does not mean that class-size reductions are the only way, or even the most attractive way, to promote such skills. The non-cognitive effects of other reform-oriented policies, from test-based accountability to school choice programs to efforts to improve teacher quality, are not well understood. Indeed, among the interventions included in the U.S. Department of Education's What Works Clearinghouse (WWC), only one topic area—character education—had outcome measures that approximate non-cognitive skills (i.e., the "knowledge, behavior, and attitudes" outcome category), and only one study reviewed in this category met WWC standards (U.S. Department of Education, 2006). The intervention evaluated in that study is a classroom curriculum designed to promote social skills and to reduce social norms related to violence and drugs. An evaluation of the curriculum in six high schools found that it promoted self-efficacy and emotional competency, and its costs are modest relative to those of class-size reductions. Further research may uncover additional policies and practices that are both effective and, quite possibly, more cost-effective than class-size reduction in this regard.

Notes

1. Schanzenbach (2007), for example, wrote that smaller classes may "improve non-cognitive skills in addition to the cognitive skills measured by standardized test scores" (p. 220).

2. Rockoff (2009) reviewed a large number of class-size experiments in the early 20th century.

3. Follow-up studies (Finn, Gerber, & Boyd-Zaharias, 2005; Krueger & Whitmore, 2001) indicated that the performance advantage of students assigned to smaller classes decreased after they were returned to regular classes in the fourth grade. However, differences in performance remained evident through eighth grade, and students who had been assigned to smaller classes in kindergarten were 3.7 percentage points more likely to take college-entrance exams in high school.

4. There is also a rapidly growing literature that uses quasi-experimental methods to estimate the

effects of class-size reduction on student achievement internationally (e.g., Angrist & Lavy, 1999; Case & Deaton, 1999; Woessmann & West, 2006). For a recent survey of this evidence, see Woessmann (2007).

5. Our study examines the reduced-form effects of smaller classes on non-cognitive skills and does not attempt to distinguish among the corresponding structural mechanisms.

6. The seminal analyses of these follow-up data focused on the effects of attending a small class rather than the effect of the assignment to a small class (i.e., the "intent to treat") and did not use a regression specification that paralleled the random assignment process, which was done within school/entry-wave cells. Dee and West (2008, Table 2) also found that the internal validity of the non-cognitive results from Project STAR may have been compromised because students who had been assigned to the treatment condition were significantly more likely to appear in the fourth-grade study sample, as were females and students ineligible for free lunches.

7. However, as these recent studies noted, the importance of non-cognitive skills had been recognized in several early studies (e.g., Bowles & Gintis, 1976; Edwards, 1976; Goldsmith, Veum, & Darity, 1997; Jencks et al., 1979).

8. The "locus of control" measure is not conducive to our within-student identification strategy. It should also be noted that recent studies raise doubts about whether self-esteem constitutes an important non-cognitive skill. The alleged difficulty with conventional measures of self-esteem is that they encompass traits such as narcissism and defensiveness that may be detrimental for long-term success. Baumeister, Campbell, Krueger, and Vohs (2003, 2004) also argued that the direction of causality between self-esteem and various outcomes has not been established and that interventions designed to promote self-esteem have generally been ineffective (or even counterproductive). Furthermore, laboratory experiments suggest that increases in self-esteem do not generally improve task performance (Baumeister et al., 2003). Nonetheless, it remains possible that efforts to boost self-esteem could be effective when they reinforce meaningful achievements instead of being pursued indiscriminately.

9. Fredricks et al. (2004) also identified a third construct, cognitive engagement, which refers to whether a student has a personal psychological investment in learning. Measures of cognitive engagement are based on student attitudes toward hard work, flexibility in problem solving, and ways of coping with challenges.

10. The available sample size is smaller largely because only a subset of base-year participants was included in the follow-up survey. However, this variable is also undefined for students who reported not taking a course in the given academic subject. We found that eighth-grade class size in a subject was unrelated to whether a student took a class in that subject during the follow-up study.

11. Specifically, DISRUPT and INATT correlate at .47 (behavioral engagement) and NOTUSE and NOTLF correlate at .38 (psychological engagement). The variable AFASK is not as strongly correlated with NOTUSE and NOTLF (.08 for NOTUSE, and .14 for NOTLF), but we consider it an indicator of psychological engagement because it measures an educationally relevant affective response to the classroom environment. It is not surprising that aggregating these variables into two construct measures yields class-size effects similar to the variable-specific results reported here (i.e., statistically insignificant effects on behavioral engagement but statistically significant effects on psychological engagement).

12. The results of this falsification exercise are similar for NOTUSE and NOLKFD. However, AFASK appears to provide a more powerful test because the effects of class size are more even across subjects. In particular, the effects of math class sizes on NOTUSE and NOLKFD are relatively small. However, with regard to all three non-cognitive measures, the hypothesis that the effects of class size are the same across subjects cannot be rejected.

13. However, the standard caveats about internal rates of return should be noted. For example, it can be misleading when judging the net benefits of projects of different scales. The internal rate of return can also take on multiple values. However, the latter concern is unlikely in this situation, which involves one upfront cost and a stream of benefits.

14. More specific, we adjusted costs upward by 3.7%, a correction based on data from Table 86 of the 2006 Digest of Education Statistics.

15. It is interesting that the fourth follow-up NELS:88 survey included questions about volunteering and voting. Increases in the eighth-grade non-cognitive measures are associated with statistically significant increases in these forms of civic participation.

References

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on student achievement. *Quarterly Journal of Economics*, *114*, 533–575.

Ashenfelter, O., & Krueger, A. (1994). Estimating the returns to schooling using a new sample of twins. *American Economic Review*, *84*, 1157–1173.

Ashenfelter, O., & Rouse, C. (1998). Income, schooling and ability: Evidence from a new sample of

identical twins. *Quarterly Journal of Economics*, *113*, 253–284.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest*, *4*(1), 1–44.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2004, December 20). Exploding the self-esteem myth. *Scientific American*. Retrieved from http://www.scientificamerican.com/article.cfm?id=exploding-the-self-esteem

Blanden, J., Gregg, P., & Macmillan, L. (2007). Accounting for intergenerational income persistence: Noncognitive skills, ability, and education. *Economic Journal*, *117*(519), C43–C60, 03.

Borghans, L., ter Weel, B., & Weinberg, B. A. (2008). Interpersonal styles and labor market outcomes. *Journal of Human Resources*, *43*(4), 815–818.

Bowles, S., & Gintis, H. (1976). *Schooling in capitalist America: Educational reform and the contradictions of economic life*. New York: Basic Books.

Case, A., & Deaton, A. (1999). School inputs and educational outcomes in South Africa. *Quarterly Journal of Economics*, *114*, 1047–1084.

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*, *42*(3), 528–554.

Dee, T. S., & West, M. R. (2008). *The non-cognitive returns to class size* (NBER Working Paper 13994). Cambridge, MA: National Bureau of Economic Research.

Deke, J., & Haimson, J. (2006). *Valuing student competencies: Which ones predict postsecondary educational attainment and earnings, and for whom?* Princeton, NJ: Mathematica Policy Research.

Education Commission of the States. (2005). *State class-size reduction measures*. Denver, CO: Author.

Edwards, R. C. (1976). Individual traits and organizational incentives: What makes a good worker? *Journal of Human Resources*, *11*(1), 51–68.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, *27*(3), 557–577.

Finn, J. D., Fulton, D., Zaharias, J., & Nye, B. A. (1989). Carry-over effects of small classes. *Peabody Journal of Education*, *67*(1), 75–84.

Finn, J. D., Gerber, S. B., & Boyd-Zaharias, J. (2005). Small classes in the early grades, academic achievement, and graduating from high school. *Journal of Educational Psychology*, *97*(2), 214–223.

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

Fortin, N. M. (2008). The gender wage gap among young adults in the United States: The importance of money versus people. *Journal of Human Resources*, *43*(4), 884–918.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109.

Glanville, J. L., & Wildhagen, T. (2007). The measurement of school engagement. *Educational and Psychological Measurement*, *67*(6), 1019–1041.

Glass, G., & Smith, M. L. (1978). Meta-analysis of the relationship of class size and student achievement. *Educational Evaluation and Policy Analysis*, *1*(1), 2–16.

Goldsmith, A. H., Veum, J. R., & Darity Jr., W. D. (1997). The impact of psychological and human capital on wages. *Economic Inquiry, 35*(4), 815–829.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *Economic Journal*, *113*(1), F64–F98.

Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics*, *54*, 3–56.

Heckman, J. J., & Rubenstein, Y. (2001). The importance of non-cognitive skills: Lessons from GED testing program. *American Economic Review*, *91*(2), 145–149.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and non-cognitive abilities on labor market outcomes and social behaviors. *Journal of Labor Economics*, *24*(3), 411–482.

Howell, W. G., West, M. R., & Peterson, P. E. (2007). What Americans think about their schools. *Education Next*, *7*(4), 12–26.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics*, *115*, 1239–1285.

Ingels, S. J., Abraham, S. Y., Karr, R., Spencer, B. D., & Frankel, M. R. (1990). *National Education Longitudinal Study of 1988 Base Year: Student component data file user's manual*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Jencks, C., et al. (1979). *Who gets ahead? The determinants of economic success in America*. New York: Basic Books.

Jepsen, C., & Rivkin, S. (2002). *What is the tradeoff between smaller classes and teacher quality?* (NBER Working Paper 9205). Cambridge, MA: National Bureau of Economic Research.

Kim, J. S. (2007). The relative influence of research on class-size policy. In T. Loveless & F. M. Hess

(Eds.), *Brookings papers on education policy, 2006/2007* (pp. 273–295). Washington, DC: Brookings Institution Press.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, *114*, 497–532.

Krueger, A. B. (2003a). Economic considerations and class size. *Economic Journal*, *113*, F34–F63.

Krueger, A. B. (2003b). Inequality, too much of a good thing. In J. J. Heckman, A. B. Krueger, & B. M. Friedman (Eds.), *Inequality in America: What role for human capital policies?* (pp. 1–75). Cambridge, MA: MIT Press.

Krueger, A. B. (2003c). Responses. In J. J. Heckman, A. B. Krueger, & B. M. Friedman (Eds.), *Inequality in America: What role for human capital policies?* (pp. 293–312). Cambridge, MA: MIT Press.

Krueger, A. B., & Whitmore, D. M. (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.

Kuhn, P., & Weinberger, C. (2005). Leadership skills and wages. *Journal of Labor Economics*, *23*(3), 395–434.

Lazear, E. P. (2001). Educational production. *Quarterly Journal of Economics*, *116*(3), 777–803.

Ouazad, A. (2008, September 12). *Assessed by a teacher like me: Race, gender, and subjective evaluations* (INSEAD Working Paper No. 2008/57/EPS). Available from http://ssrn.com/abstract=1267109

Rockoff, J. (2009). Field experiments in class size from the earlier twentieth century. *Journal of Economic Perspectives*, *23*(4), 211–230.

Rouse, C. (1999). Further estimates of the economic return to schooling from a new sample of twins. *Economics of Education Review*, *18*, 149–157.

Schanzenbach, D. W. (2007). What has research learned from Project STAR? In T. Loveless & F. M. Hess (Eds.), *Brookings papers on education policy, 2006/2007* (pp. 205–228). Washington, DC: Brookings Institution Press.

Schrag, P. (2007). Policy from the hip: Class-size reduction in California. In T. Loveless & F. M. Hess (Eds.), *Brookings papers on education policy, 2006/2007* (pp. 229–243). Washington, DC: Brookings Institution Press.

Segal, C. (2008a). Classroom behavior. *Journal of Human Resources*, *43*(4), 783–814.

Segal, C. (2008b). *Motivation, test scores, and economic success* (UPF Working Paper 1124). Barcelona: Universitat Pompeu Fabra, Department of Economics and Business.

Summers, L. H. (2003). Comments. In J. J. Heckman, A. B. Krueger, & B. M. Friedman (Eds.), *Inequality in America: What role for human capital policies?* (pp. 285–291). Cambridge, MA: MIT Press.

ter Weel, B. (2008). The noncognitive determinants of labor market and behavioral outcomes: Introduction to the symposium. *Journal of Human Resources*, *43*(4), 729–737.

Urzua, S. (2008). Racial labor market gaps: The role of abilities and schooling choices. *Journal of Human Resources*, *43*(4), 919–971.

U.S. Department of Education. (2006). *What Works Clearinghouse intervention report: Too good for drugs and violence*. Washington, DC: U.S. Department of Education, Institute for Education Sciences.

Voelkl, K. E. (1995). *Class size and classroom participation*. Paper presented at the National Symposium on Class Size Research, London, England.

Waddell, G. R. (2006). Labor-market consequences of poor attitude and low self-esteem in youth. *Economic Inquiry, 44*(1), 69–97.

West, M. R., & Peterson, P. E. (2007). *School money trials: The legal pursuit of educational adequacy*. Washington, DC: Brookings Institution Press.

West, M. R., & Woessmann, L. (2006). Which school systems sort weaker students into smaller classes? International evidence. *European Journal of Political Economy*, *22*, 944–968.

Woessmann, L. (2007). International evidence on expenditure and class size: A review. In T. Loveless & F. M. Hess (Eds.), *Brookings papers on education policy, 2006/2007* (pp. 245–272). Washington, DC: Brookings Institution Press.

Woessmann, L., & West, M. R. (2006). Class-size effects in school systems around the world. *European Economic Review*, *50*(3), 695–736.

Authors

THOMAS S. DEE is Professor of Public Policy and Economics and Research Professor of Education at the University of Virginia. His research focuses on the analysis of designed and natural experiments to inform issues in education policy.

MARTIN R. WEST is Assistant Professor of Education at the Harvard Graduate School of Education and the Deputy Director of the Program on Education Policy and Governance at Harvard's Kennedy School. His research analyzes the politics of education in the United States and the impact of education reforms on student outcomes.