NBER WORKING PAPER SERIES

THE NON-COGNITIVE RETURNS TO CLASS SIZE

Thomas Dee
Martin West

Working Paper 13994
http://www.nber.org/papers/w13994

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2008

We thank Jeremy Finn, Rob Hollister, John Tyler and Yona Rubinstein and seminar participants at Harvard, Cornell, Swarthmore, and the 2008 AEFA meetings for helpful discussions and comments. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

The Non-Cognitive Returns to Class Size
Thomas Dee and Martin West
NBER Working Paper No. 13994
May 2008
JEL No. I20,I21,I28

ABSTRACT

Although recent evidence suggests that non-cognitive skills such as engagement matter for academic and economic success, there is little evidence on how key educational inputs affect the development of these skills. We present a re-analysis of follow-up data from the Project STAR class-size experiment and find evidence that early-grade class-size reductions did improve subsequent student initiative. However, these effects did not persist into the 8th grade. Furthermore, the external and, possibly, the internal validity of these inferences is compromised by non-random attrition. We also present a complementary analysis based on nationally representative survey data and a research design that relies on contemporaneous within-student and within-teacher comparisons across two academic subjects. Our results indicate that smaller classes in 8th grade lead to improvements in measures of student engagement with effect sizes ranging from 0.05 to 0.09 and smaller effects persisting two years later. Using the estimated earnings impact of these non-cognitive skills and the direct cost of a class-size reduction, the implied internal rate of return from an 8th-grade class-size reduction is 4.6 percent overall, but 7.9 percent in urban schools.

Thomas Dee
Department of Economics
Swarthmore College
Swarthmore, PA 19081
and NBER
dee@swarthmore.edu

Martin West
Box 1938, 21 Manning Walk
Brown University
Providence, RI 02912
martin_west@brown.edu

1. Introduction

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, particularly for children in early grades. Over the last 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 and Peterson 2007). And Howell et al. (2007) report that 77 percent of American adults would prefer to see new educational dollars spent on reducing class sizes rather than on increasing teacher salaries.

While class-size reduction has strong intuitive appeal among parents and policy makers, its effectiveness (and cost-effectiveness) continues to be debated among researchers. Krueger (2003a, p. 36), for example, asserts 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." Hanushek (2003, p. F92), on the other hand, argues that class-size reductions are an "expensive and generally unproductive policy."

The growing recognition of the importance of non-cognitive skills for later life outcomes may have important implications for this debate. "Non-cognitive skills" is an overarching term for a range of behaviors, habits, and attitudes that are not measured by conventional tests of cognitive ability. Indicators of non-cognitive skills that are highly predictive of long-term educational and labor-market outcomes include such classroom

behaviors as attentiveness, disruptiveness, tardiness, absenteeism, and homework completion. Moreover, unlike IQ, which largely stabilizes while students are still in elementary school, characteristics such as intellectual engagement, motivation, and self-discipline appear to be malleable at later ages. As Carneiro and Heckman (2003) point out, this evidence suggests that evaluations of educational interventions should incorporate analyses of their effects on both cognitive and non-cognitive skills. Yet while 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 paper examines the non-cognitive returns to class size through two complementary studies of its effects on measures of student engagement drawn from teacher and student surveys. First, we provide a reanalysis of follow-up data from Project STAR, a randomized evaluation of class-size reduction in kindergarten through 3rd grade launched in 1985 in Tennessee. Evidence from Project STAR on the cognitive benefits of smaller classes has been extraordinarily influential in building support for class-size reduction in the early grades (Krueger 1999, Krueger and Whitmore 2001, Schanzenbach 2007). We use teacher survey data collected after the conclusion of the experiment, when most STAR students were in 4th and 8th grade, to determine whether taking into account possible long-term non-cognitive benefits provides additional support for class-size reduction policies.

[1] Krueger (2003b, page F58), for example, suggests that existing cost-benefit analyses of class-size reduction probably understate its benefits because it is "likely that school resources influence non-cognitive abilities, which in turn influence earnings." See also Schanzenbach (2007, p. 220), who writes that smaller classes may "improve non-cognitive skills in addition to the cognitive skills measured by standardized test scores."

We also present a complementary analysis based on nationally representative survey data from the National Educational Longitudinal Study of 1988 (NELS:88). To estimate the causal effect of class size on both cognitive and non-cognitive 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 on identical twin pairs (e.g., Ashenfelter and Krueger 1994, Ashenfelter and Rouse 1998, and Rouse 1999). It has also been used with the NELS:88 to isolate the causal effects of teacher traits such as gender (Dee 2005, 2007), race (Dee 2005), and subject-specific qualifications (Dee and Cohodes 2008). Our results based on this strategy indicate that smaller class sizes in 8th grade lead to improvements in several measures of student engagement, with effect sizes ranging from 0.05 to 0.09, and persistent but smaller improvements two years later. We also use data from the 2000 follow-up interview of adult NELS:88 respondents to construct a rough cost-benefit analysis of both general and targeted class-size reductions in the 8th grade in light of its effects on both cognitive and non-cognitive skills

The remainder of the paper 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 intellectual engagement in educational production. The following section presents experimental estimates of the effect of early-grade class-size reductions on the available measures of student engagement in the 4th- and 8th-grade follow-up data from Project STAR. Sections 4 and 5 present our analysis of the effects of 8th grade class size on a range of

engagement measures in the NELS:88 database and compare the costs and benefits of general and targeted class-size reductions in the 8th 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.

2. Literature Review

The Effects of Class Size

The scholarly debate over the effectiveness of class-size reductions has a long history (Glass and Smith 1978). 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 relationship between teacher effectiveness and classroom assignment may also undermine inferences about the effects of smaller classes based on observational data. Largely in an attempt to settle this debate, the state of Tennessee in the 1980s carried out a large experimental evaluation of class-size reduction known as Project STAR (Student/Teacher Achievement Ratio). In the 1985-86 school year, kindergarten students in each of 79 elementary schools were randomly assigned to one of three class types: small, regular, or regular with an full-time teacher's aide. Teachers were also randomly assigned to classrooms within their school. Students who entered a participating school while this cohort of students was in grades 1-3 were added to the experiment and randomly assigned to a classroom. In 4th grade, all students were returned to regularly sized classrooms.

Because of its experimental design, Project STAR arguably provides the best available evidence on the effects of class-size reductions in the early grades. Early

analyses of the experimental data (e.g. Finn et al. 1989, Finn and Achilles 1990) reported positive effects on student achievement of being assigned to a smaller class, but these studies failed to address the potential implications of non-random sample attrition and treatment crossover. In a reanalysis intended to address these lingering concerns, 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. The estimated effect of small-class assignment persisted, but did not increase, as these students remained in smaller classes through the third grade.

Subsequent research (Krueger and Whitmore 2001; see also Finn et al. 2005) indicates that the performance advantage for students assigned to smaller classes decreased after they were returned to regular classes in the fourth grade. However, differences in performance remained evident through 8th 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. For African-American students, the difference in participation rates on college-entrance exams was 8.5 percentage points. Krueger (2003b) compares 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 concludes that the internal rate of return to class-size reductions is roughly 6 percent.

While evidence from Project STAR has been influential in creating enthusiasm for class-size reductions in early grades (Boruch 2002), it has also been criticized on several grounds. Hanushek (2003) argues that the lack of baseline data on the performance of students before entering the experiment makes it impossible to assess

fully the quality of the initial random assignment process and the effects of attrition from the study sample. Hoxby (2000) suggests that the knowledge that they were participating in an experiment may have altered the behavior of participating teachers, making the results an unreliable guide for policy.[2] Finally, the external validity of the experiment may be limited since it was conducted in a single state among schools which volunteered to participate and were large enough to accommodate the research design. Other recent research on American students, using quasi-experimental methods, shows no evidence of class-size effects on student achievement (e.g., Hoxby's 2000 study of Connecticut).[3]

The Effects of Non-cognitive Skills

As this discussion suggests, research on class size—like the great majority of the broader academic literature on skill formation and its consequences—has focused mainly on student performance on tests designed to measure quite specific cognitive skills. Similarly, prominent accountability-based education reforms such as the No Child Left Behind (NCLB) Act and mandatory high-school exit exams also focus largely on test-based measures of cognitive performance. However, common sense suggests that a variety of personality traits and behaviors that are conceptually distinct from cognitive ability, such as motivation, persistence, and engagement, also have an important influence on both educational and labor-market outcomes.

---

[2] Specifically, if teachers suspected that future investments in class-size reductions would be made only if the experiment showed benefits, they may have made special efforts to ensure that students in smaller classes performed well.

[3] There is also a large and rapidly growing literature that uses quasi-experimental methods to estimate the effects of class-size reduction on student achievement internationally (e.g. Case and Deaton 1999, Angrist and Lavy 1999, West and Woessmann 2006, Woessmann and West 2006). For a recent survey of this evidence, see Woessmann 2007).

A number of recent studies have brought new attention to the role of such non-cognitive skills.[4] In particular, Heckman and Rubenstein (2001) note that high school dropouts who successfully complete a General Education Development (GED) test had *lower* wages and schooling levels than other high school dropouts after controlling for measured ability. They conclude 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 both relevant for educational attainment and valued in the labor market. A subsequent study by Heckman et al. (2006) presented evidence that non-cognitive skills had substantive influences on both labor-market outcomes and a variety of risky behaviors. Specifically, this study measured non-cognitive skills with the Rosenberg self-esteem scale and the Rotter "locus-of-control" scale, which measures the extent to which individuals feel that their own actions determine the outcomes they experience (as opposed to chance or their environment).[5]

Several other recent studies have examined the effects of alternative measures of non-cognitive skills on educational and labor market outcomes. For example, Segal (2006a) finds that teacher reports of a student's behavior in 8th grade appear as important for adult earnings as their 8th-grade test scores. Similarly, Deke and Haimson (2006) find that a composite measure of work habits (based on student and teacher reports) has an apparent effect on subsequent educational attainment similar to that of a test-score measure. They also find that the effect of a locus-of-control variable on adult earnings is similar to the effect of a test-score measure. Furthermore, in a study of British data,

[4] However, as these recent studies note, the importance of non-cognitive skills had been recognized in several early studies (e.g., Bowles and Gintis 1976, Edwards 1976, Jencks et al. 1979)

[5] Other recent studies have also concluded that the non-cognitive abilities measured by the Rotter and Rosenberg scales had similar effects on wages and occupational traits (Goldsmith, Veum, and Darity 1997, Waddell 2006).

Blanden et al. (2006) find 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.

Another provocative development in the literature on non-cognitive skills involves new evidence that researchers' distinctions between measures of cognitive and non-cognitive skills may be somewhat illusory. More specifically, two recent studies (Segal 2006b, Borghans et al. 2006) argue that the scores of survey respondents on low-stakes cognitive tests appear to reflect in part non-cognitive skills that are associated with their degree of test effort. In particular, laboratory experiments indicate that subjects whose personalities exhibit high levels of motivation and an internally focused locus of control perform similarly on cognitive tests regardless of the presence of incentives, while other subjects improve their test performance significantly in the presence of stronger incentives.

Student Engagement

The term "non-cognitive skills" clearly encompasses a diverse set of psychological traits. In fact, a principal-components factor analysis discussed in Heckman et al. (2006, footnote 6) confirms that non-cognitive skills reflect multiple latent factors. However, educational psychologists have increasingly highlighted intellectual *engagement* as a "multidimensional construct" with particularly important implications for academic success. In a recent review of this literature, Fredricks et al. (2004) broadly characterize school engagement as an active commitment to education and note that concerns about 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.

Educational researchers have settled on a three-part taxonomy of types of student engagement (Fredricks et al. 2004). “Behavioral engagement” focuses on forms of academic participation such as attendance, not being disruptive, effort, assignment completion, attentiveness in class, and asking questions. “Emotional engagement” consists of student reactions to teachers, peers, and academics in general (e.g., interest, boredom, sadness, and anxiety). Some measures of emotional engagement (e.g., “Is math useful for your future?”) focus on identification with school and are closely related to measures of motivation. The third category, “cognitive engagement”, refers to whether a student has a personal psychological investment in learning. Measures of cognitive engagement are based on student attitudes towards hard work, flexibility in problem solving, and ways of coping with challenges.

The non-cognitive outcome variables used in this study all measure behaviors and attitudes that can be categorized as dimensions of student engagement. There is little evidence on whether smaller classes actually promote the “fusion of behavior, emotion, and cognition” (Fredricks et al 2004, p. 61) that constitutes student engagement. However, smaller classes could conceivably improve all three dimensions of student engagement. For example, smaller classes promote behavioral engagement by allowing teachers to limit disruptive behavior as well as to encourage attentiveness and asking questions. Smaller classes may also help teachers promote emotional engagement in the form of student interest and personal academic identification. Finally, smaller classes

may promote cognitive engagement by allowing teachers to assist students in flexible problem-solving in the face of challenges.

A fundamental, common-sense appeal of the use of student engagement as a non-cognitive skill is that the behaviors and attitudes that make up the construct seem unambiguously instrumental in promoting academic success.[6] In contrast, it should be noted, recent work from the fields of education and psychology suggests a growing consensus against the view that the measures of self-esteem used in several recent economic studies primarily capture 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 et al. (2003, 2004) also argue 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.

3. Evidence from Project STAR

Project STAR was a randomized study of class-size reduction in kindergarten through third grade conducted in 79 Tennessee schools beginning in the 1985-86 academic year (Finn and Achilles 1990, Mosteller 1995). The students and teachers

[6] We also present evidence that our engagement measures have robust positive correlations with multiple measures of academic and labor market success.

assigned to the treatment condition were in classes that had, on average, 16 students while the remaining classrooms averaged 23 students. This influential study remains the only large-scale experiment designed to isolate the causal effect of class size on student outcomes (Schanzenbach 2007). As such, it provides a unique and compelling opportunity to evaluate the effects of class size on non-cognitive student outcomes.

Data on non-cognitive outcomes were collected for a subset of the students who participated in Project STAR in the first and fifth years after the experiment concluded (i.e., when most student participants were in the 4^{th} and 8^{th} grades). Specifically, a team of researchers fielded teacher questionnaires that solicited information on the observable classroom behavior of sampled students (Finn, Fulton, Zaharias and Nye 1989; Finn, Pannozzo and Achilles 2003). In the seminal analysis of the 4^{th}-grade data, Finn, Fulton, Zaharias and Nye (1989) found that the students who were taught in small classes in grades K-3 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, subsequent analyses indicated that these effects did not persist into the 8^{th} grade (Finn, Pannozzo and Achilles 2003, p. 329; Voelkl 1995).

In this section, we present an independent re-analysis of these 4^{th}- and 8^{th}-grade follow-up data. Our analysis addresses three potential shortcomings of the original analysis. First, Finn, Fulton, Zaharias and Nye (1989) classified the treatment status of students by whether they attended small classes *in the third grade*. However, Krueger (1999) reports that about 10 percent of participants were moved from one type of class to another after their initial assignment. This "treatment crossover" constitutes a potentially important source of bias. Our analysis therefore focuses instead on an "intent to treat"

variable (i.e., class-size status upon entry to the experiment) rather than using the potentially compromised measure of class-size status at the *end* of the experiment.

Second, only subsets of Project STAR participants were included in the 4th and 8th-grade follow-up studies. Non-random attrition from these studies could compromise both the external and the internal validity of the resulting inferences. We use auxiliary regressions to assess the pattern of attrition and the balance of observed characteristics across treatment and control states. Third, like other recent Project-STAR analyses (i.e., Krueger 1999, Krueger and Whitmore 2001), we adopt fixed-effects specifications that may be more efficient and that more clearly reflect the nature of the random-assignment process. In particular, our preferred specifications condition on fixed effects unique to each school-by-entry-wave cell because randomization occurred within schools upon entering the study.

It should be noted that, in addition to the teacher-reported follow-up data we study, other non-cognitive traits were also measured while the Project STAR experiment was in progress. More specifically, at the end of each school year during the experiment, the participating students completed a group-administered instrument, the Self-Concept and Motivation Inventory (SCAMIN). The SCAMIN attempts to measure internal personality traits related to self-concept and motivation by asking students to indicate which of a set of pictures would best capture their response to 24 situations (Finn and Achilles 1990). Like other proposed measures of non-cognitive skills that primarily capture self-esteem, it has been criticized for its low correlation with academic

performance and ambiguity over the causal direction of any observed relationships.[7] In addition, Finn and Achilles (1990, p. 562) report that the SCAMIN's "reliability is only moderate in the early grades, with alpha coefficients between .56 and .69 for subscales." Even so, prior analyses of the self-concept and motivation subscales suggest that assignment to a small class had no statistically significant effects on these measures (Finn and Achilles 1990) or effects that were not sustained to grade 3 (Schanzenbach 2007).

The Project STAR and Beyond Database

Our data come from the recently released Project STAR and Beyond Database, which contains information on the full sample of 11,601 students who participated in the experimental phase of the study for at least one year (Finn, Boyd-Zaharias, Fish and Gerber 2007). Project STAR relied on a within-school experimental design: All students entering participating schools in Kindergarten through 3rd grade were randomly assigned to either a small class (intended to be between 13 and 17 students), a regular class (22-25 students), or a regular class with a full-time teacher's aide within their school.[8]

The composition of the sample of schools included in Project STAR reflects both the decisions of schools about whether to participate in the study and eligibility restrictions on school participation. All elementary schools in Tennessee were invited to participate, and 180 schools in 50 districts initially expressed interest. In order to be eligible, however, schools had to have at least 57 Kindergarten students—enough to accommodate two regular classes of 22 students and a small class of 15 students. Among

[7] The correlations of the self-concept and motivation scores with math achievement in the 3rd grade Project STAR sample are -0.004 and -0.015, respectively. The analogous correlations with reading scores are 0.023 and 0.025.

[8] Due to concerns about fairness, students assigned to regular or regular with aide classes in Kindergarten were re-randomized in 1st grade. Students assigned to small classes were unaffected by this modification of the experimental design.

the approximately 100 schools meeting the size restrictions, 79 schools in 42 districts were ultimately selected to participate. Study administrators ensured that inner-city schools (defined as those in metropolitan areas with more than half of their students eligible for a free or reduced price lunch) were overrepresented in the final sample in order to comply with requirements for geographic diversity mandated by the state legislature. In the end, Project STAR schools were larger and enrolled a higher percentage of blacks and poor students than elementary schools statewide.

Our analysis centers on three distinct outcome measures collected in two follow-up studies conducted when most STAR students were in 4th and 8th grades.[9] Specifically, the 4th grade teachers in participating schools rated as many as 19 randomly chosen students who had entered Project STAR schools in Kindergarten or 1st grade on 31 items that comprised the "Student Participation Questionnaire" (SPQ).[10] Each item on the SPQ asked teachers to rate the frequency of occurrence of a particular behavior from "never" (1) to "always" (5). Factor analysis was used to group twenty-five of the SPQ survey items into three subscales measuring student effort (13 items, e.g., "Pays attention in class"), initiative-taking (8 items, e.g., "Does more than just the assigned work"), and non-participatory behavior (4 items, e.g., "Annoys or interferes with peers' work").

The Project STAR and Beyond Database reports simple scale scores created by summing the responses to all of the items within each cluster (after inverting the responses to items asking about negative behaviors). Alpha coefficients for the effort,

---

[9] Each of the three non-cognitive measures based on the SPQ is strongly associated with the current and future academic success as measured by test scores, SAT or ACT participation, performance on the SAT or ACT (conditional on participation), and high school completion.

[10] Although the study protocol called for teachers to complete surveys for only 10 students in order to lessen the burden of participating, 64 of the 262 teachers who participated completed surveys for more than 10 students.

initiative, and non-participatory behavior subscales are .94, .89, and .89, respectively, indicating a high degree of internal consistency (Finn, Folger, and Cox 1991, p. 391). We standardize these scales separately to have a mean of zero and a variance of one.

Four years after the experiment concluded, when most STAR participants were in 8th grade, many of their English and mathematics teachers were both asked to complete a shortened version of the SPQ consisting of 14 items, all but one of which was drawn from the 4th-grade survey.[11] The items were grouped into three additive scales again measuring student effort, initiative, and non-participatory behavior, each of which we again standardize to have a mean of zero and a variance of one.

These measures of effort, initiative, and non-participatory behavior are available for 2,212 students in 75 schools at the 4th-grade level.[12] In 8th grade, they are available for 2,978 students attending 167 schools. While the 4th-grade sample includes only students entering Project STAR schools in kindergarten or 1st grade, the 8th-grade sample includes students from all four entering cohorts.

Table 1 shows how the observed baseline characteristics of these students compare with the full Project STAR sample. In 4th grade, students included in the SPQ are more likely to be White or Asian, more likely to be female, less likely to be born in 1979 (an indicator that they are above-age for their grade), less likely to be eligible for a free lunch, and more likely to have been assigned to a small class. Many of these same differences appear at the 8th grade level, though there no longer seems to be any association between treatment status and inclusion in the SPQ. Auxiliary regressions

[11] The new item, asking how often students were "verbally or physically abusive to the teacher," had been considered less relevant for younger students.

[12] At least one 4th-grade teacher in 71 of the 79 schools originally participating in Project STAR completed the SPQ for at least one student, as did teachers from an additional 4 schools that enrolled Project STAR students but were not in the initial study sample.

based on the specification introduced below provide more direct evidence on the patterns of attrition from the follow-up studies and their implications for the internal and external validity of our results.

Intent-to-Treat Specifications

Given random assignment, the causal effect of being assigned to a small class can be estimated by comparing average outcomes across different class types. All published analyses of the Project STAR have shown that assignment to a class with a teacher aide had no effect on student test scores, a pattern we also see in our analyses of non-cognitive outcomes (results not shown). We therefore ignore this aspect of the experimental design and compare students assigned to small classes to all students in regular classes, regardless of whether an aide was present.

More specifically, we estimate separate models by grade (e.g., 4^{th} and 8^{th}) and by non-cognitive outcome (e.g., each of the three SPQ subscales) that take the following form:

$$(1) \qquad Y_{icsw} = \beta_0 + \beta_1(SMALL_{csw}) + \beta_2\mathbf{X}_{icsw} + \alpha_{sw} + \varepsilon_{icsw}$$

where Y_{icsw} is the dependent variable for student i in classroom c of school s and entry wave w. $SMALL_{csw}$ is a binary indicator for whether a student's *initial* classroom assignment within a school and entry wave was to a small class. The matrix, $\mathbf{X}_{icsw}$, includes controls for student observables (race, gender, free-lunch eligibility, and an indicator for having been born in 1979) and α_{sw} represents a full set of fixed effects unique to each school-by-entry-wave cell. The rationale for conditioning on school fixed effects specific to the entry wave (i.e., the grade of entry into the experiment) is that randomization occurred in this manner (i.e., within school upon entering the experiment).

The error term, ε_{icsw}, is adjusted to reflect heteroscedasticity clustered at the school-by-entry-wave level. In order to provide continuity with the initial analysis of these data, some specifications exclude the school-by-entry-wave fixed effects.

Study attrition

Evidence in Table 1 shows that the observable characteristics of Project STAR students for whom the 4th and 8th-grade SPQ data are available differ notably from the initial Project STAR sample. The propensity for certain types of students to leave the follow-up studies (e.g., male and free-lunch eligible students) clearly reduces the external validity of the results with respect to the full Project STAR sample. Moreover, to the extent that attrition patterns differed between students in the treatment and control groups, it may also undermine the internal validity of our results by creating a situation where the treatment and control groups differ with respect to unobserved traits.

To assess attrition patterns more carefully, we estimated auxiliary regressions based on the full Project STAR sample that have as their dependent variable a dummy variable indicating whether the student was included in the SPQ studies in 4th and 8th grade. The first two columns of Table 2 show that, conditional on school-by-entry-wave fixed effects, females and students who were born in 1979 were significantly more likely to be included in the SPQ samples in both 4th and 8th grade, while poor students were significantly less likely to be included. Because the 4th grade sample was drawn primarily from schools that participated in the experimental phase of Project STAR, the under-

representation of poor students in the 4th grade may reflect higher levels of mobility within this population.[13]

Of potentially greater concern is the fact that for the 4th-grade (but not the 8th-grade) sample, attrition was more likely in the control group.[14] Unless the relative attrition from the control group is random—that is, unrelated to unobserved characteristics that affect student outcomes—our estimates of class-size effects based on equation (1) will be biased. The third and fourth columns of Table 2 provide evidence that the difference in attrition between treatment and control group students was in fact random with respect to observed characteristics. Specifically, the baseline student characteristics that are available to us are uncorrelated with assignment to a small class within the SPQ samples in either grade. However, selection on unobserved characteristics still remains a potential concern. If, for example, the treatment-group students who benefited the most from small classes were more likely to remain in their school, then our estimates will exaggerate any benefits of smaller classes.

Results

Table 3 presents estimates based on equation (1) of the effect of assignment to a small class on each of the three measures of non-cognitive skills drawn from the 4th-grade SPQ. For purposes of comparison with the seminal evaluations of these data, we also present models that do not include school-by-entry-wave fixed effects. Finally, as a robustness check, we also present models that include a fixed effect for each 4th-grade

[13] Additional regressions (not shown) conditional on fixed effects for each 4th-grade school suggest that the selection of students was random with respect to students observed characteristics within schools, as the evaluators intended.

[14] It should be noted that this pattern of differential attrition by treatment status does not appear in the larger group of students who were included in the achievement follow-up studies (e.g., Krueger and Whitmore 2001).

teacher who completed an SPQ. Although teacher quality in the 4th grade is potentially endogenous, we use this specification to ensure that our results are not driven by systematic differences in rating standards across teachers.

Our preferred specification shows that assignment to a small class in grades K-3 was associated with an increase in student initiative as measured in 4th grade of more than 0.1 standard deviations. Although this effect should be interpreted cautiously in light of the concerns about non-random attrition noted above, it is statistically significant and robust to the inclusion of fixed effects for the teacher completing the survey.[15] In contrast, we do not find strong evidence that assignment to a small class led to increases in effort or to decreases in non-participatory behavior. Although both point estimates are in the expected direction, both fall short of conventional levels of statistical significance.

Our results therefore differ from those reported by Finn et al (1989, Table 3), who report statistically significant effect sizes of 0.14, 0.12, and -0.11, respectively, for the effect of smaller classes on initiative, effort, and non-participatory behavior. As discussed above, our analysis differs from theirs in several ways. For example, we control for observed student characteristics and use students' initial classroom assignment rather than their class type in 3rd grade as the treatment variable. However, neither of these changes appears to explain the divergence in our results. Estimates based on the model that makes only these changes are all at least weakly significant and similar in magnitude to those reported in the earlier analysis, as are similar models that include dummy variables measuring the urbanicity (rural, suburban, urban, or inner-city) of each student's entry school.

[15] When 4th grade math and reading test scores are included in the regression, however, the effect of class size on initiative becomes small and statistically insignificant.

However, we find that conditioning on school-by-entry-wave fixed effects leads to substantial reductions in the estimated treatment effects for all three dependent variables. Results based on models that introduce either entry-school or entry-wave fixed effects separately indicate that each of these changes contributes to the reduced estimate across all three outcome variables. This sensitivity suggests that treatment students who appear in the follow-up studies tend to have unobserved school-level and cohort-level traits that predict good non-cognitive outcomes—and that the prior evidence linking small-class assignment to improvements in student effort and behavior should be treated with caution. Nonetheless, it should also be noted that the 95-percent confidence interval for the point estimates that condition on fixed effects includes the point estimate from the model that does not.

As Table 4 indicates, we find no evidence based on our preferred specification that the apparent effects of early class size on student initiative in 4th grade persist to grade 8. In fact, the only (weakly) significant evidence we find of class-size effects on the non-cognitive outcomes we examine runs in a counter-intuitive direction, suggesting that being placed in a small class in grades K-3 *increases* non-participatory behavior as reported by students' 8th-grade math teacher by 0.08 standard deviations. Because we did not observe similar patterns when students were in 4th grade, we consider this result to be a false positive. Taken as a whole, the results in Table 4 strongly suggest that any effects of class size in the early grades on the non-cognitive outcomes measured by the SPQ had fully dissipated by grade 8.

4. Evidence from NELS:88

The use of large, nationally representative survey data to examine the effects of class size on non-cognitive outcomes provides a useful complement to the Project STAR findings for a number of reasons. First, the external validity of results based on a sample of larger public schools from Tennessee for schools in other states is not unambiguously clear. The external (and, possibly, the internal) validity of the longer-term Project STAR results presented above is further compromised by the non-random attrition of study participants who were at risk of low achievement. Second, a complementary data set can 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 only provided evidence 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 (Heckman 2000; Carneiro and Heckman 2003). A third contribution of studying an alternative data set is that longer-term longitudinal data 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.

However, the use of non-experimental data also raises some non-trivial identification challenges. 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 empirical evidence (West and Woessmann 2006), which are confirmed here, indicate that there appears to be

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 a struggling teacher by allowing them to have smaller classes. This study addresses these issues by exploiting the unique features of the student and teacher surveys in a major, nationally representative, longitudinal study, which make it possible to examine class-size effects conditional on both student and teacher fixed effects. The next three sections introduce the relevant data, specifications, and the 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)

NELS:88 is nationally representative, longitudinal survey that began in 1988 with a sample of 24,599 8th-grade students from over 1,000 schools. The two-stage sampling design selected schools first and then approximately 26 students within each participating school (Ingles et al. 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 which 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 with each academic subject.

Our analysis exploits each type of outcome measures. The measures drawn from student surveys were based on three questions 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 didn't 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, NOTLF, and AFASK (Table 5). The teacher perceptions of individual students are based on binary indicators for whether they viewed a particular student as frequently disruptive and 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 5).

In order to examine whether the effects of smaller classes persist, we also utilize 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 over 9,000 base-year students.[16]

All of these non-cognitive outcome measures fit within the broad construct of student engagement as conceptualized by educational psychologists (Fredricks et al. 2004). The teacher-reported measures (i.e, INATT and DISRUPT), which focus on the character of classroom participation, represent forms of "behavioral engagement." In contrast, the student-reported measures (i.e., AFASK, NOTLF, NOTUSE, and TRYH) are more clearly aligned with "emotional engagement" in that they reflect reactions to teachers and peers and identification with school as it relates to motivation and effort.

Table 5 reports the means and standard deviations for these and other variables for which variation might exist after conditioning on student and teacher fixed effects.

[16] The available sample size is smaller largely because only a subset of base-year respondents were 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 8th-grade class size in a subject was unrelated to whether a student took a class in that subject during the follow-up study.

The average class size in this sample is 24.5 with a standard deviation of 5.9.[17] 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, where test score data are available, are more likely to be lower-achieving. Based on the prior class-size literature, which finds 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.

First-difference (FD) 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

[17] It should be noted that we found results similar to those reported here when excluding the small number of class-size outliers (i.e., classes smaller than 8 or larger than 36).

condition on teacher fixed effects that reflect the unobserved teacher quality that may also be correlated with class size.

More specifically, 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., $SIZE_{1it}$):

$$y_{1it} = \boldsymbol{\alpha}\mathbf{X_i} + \beta(SIZE_{1it}) + \boldsymbol{\lambda}\mathbf{Z_{1t}} + \theta_{1t} + \mu_i + \varepsilon_{1it} \quad (2)$$

In equation (2), 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 (2), it would be difficult to estimate β reliably because the error term in equation (2) 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. More specifically, suppose an equation like (2) applies to the student outcomes observed in English or history:

$$y_{2it} = \boldsymbol{\alpha}\mathbf{X_i} + \beta(SIZE_{2it}) + \boldsymbol{\lambda}\mathbf{Z_{2t}} + \theta_{2t} + \mu_i + \varepsilon_{2it}. \quad (3)$$

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

$$(y_{1it} - y_{2it}) = \beta(SIZE_{1it} - SIZE_{2it}) + \boldsymbol{\lambda}(\mathbf{Z_{1t}} - \mathbf{Z_{2t}}) + (\theta_{1t} - \theta_{2t}) + (\varepsilon_{1it} - \varepsilon_{2it}). \quad (4)$$

Estimates based on equation (4) 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 (4) 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 towards *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. While 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 specifically, 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 smaller class in a particular subject if that smaller class also improved student outcomes in another subject. However, we suspect 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 6 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 specifically, the most basic FD specification (i.e., column 2) suggests that smaller classes reduce the extent to which students don't look forward to a subject, don't 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 (though, not disruptive). Smaller classes also appear to increase student test scores. However, the effect size is quite small (i.e., .0022 x 5.87 = 0.013) and statistically insignificant.

The third specification in Table 6 introduces teacher fixed effects (more specifically, teacher fixed effects specific to math-science and English-history subject pairings). Interestingly, the introduction of these controls increases the R^2 in these regressions quite dramatically. More important, the estimated effects of class size on NOTLF, 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 8th 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 0.05 to 0.08. 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 upwards. 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 6 do not overstate the non-cognitive benefits of smaller classes (and may, in fact, understate them). First, the estimated effects of class size on NOTLF, 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 6 actually suggest a pattern of *negative* selection into smaller classes. More specifically, models that include weaker student and teacher controls suggest that smaller class sizes have smaller or even negative benefits (Table 6). 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 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 7 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 percent 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 specifically, we estimated specifications where the non-

cognitive outcome in one academic subject was replaced by the corresponding outcome from a related academic subject.

Table 8 presents the key results from this exercise and focuses on one of the three academic-engagement measures, AFASK.[18] 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 8 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). 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 student engagement with history. Overall, these results imply that the non-cognitive 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 8 suggest that, for all four academic subjects, the

[18] The results of this falsification exercise are similar for NOTUSE and NOTLF. 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 NOTLF 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.

spillover effects to related subjects are relatively small. Like the prior robustness checks, this pattern implies that non-random, within-student selection to smaller classes is not confounding our results.

While 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 8th-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 student 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 student engagement as a form of "skill" development would clearly be strengthened if the subject-specific effects were to persist over time.

Fortunately, 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 9 reports the key results from specifications that estimate the effect of subject-specific class sizes in 8th grade on these longer-term measures. The basic FD specification suggests that a smaller class size in an academic subject during 8th grade implies a statistically significant increase in effort in that subject two years later. The effect size implied by this point estimate (0.032) 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 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, NOTLF, 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 8th graders and the implicit assumption of a common treatment effect for different types of students and schools. However, there are a variety of reasons to suspect that the effects of class size might differ across particular types of students and educational settings. Table 10 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 1 SD decrease 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 x 0.0058), a reduction of roughly 13 percent relative to the gender-specific mean. Similarly, a 1 SD 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 x 0.0109), a reduction of roughly

38 percent 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 0.037 and 0.067, respectively. The estimated effects of class size on the student-engagement measures (i.e., NOTLF, NOTUSE, and AFASK) also differ across the sub-samples. 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.

5. Cost-Benefit Considerations

Our NELS:88 analysis indicates that class-size reductions in the 8th grade lead to statistically significant improvements in several non-cognitive outcomes (i.e., NOTLF, 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.

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 8th grade. The fourth follow-up interview of NELS:88 respondents, which elicited information on both educational attainment and early labor-market experiences, occurred in 2000, when respondents were approximately 26 years old. In order to gauge the possible benefits of 8th grade class-size reductions, we examine the effects of the 8th 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, the comparative results across specifications that introduce additional controls provide evidence on the direction of selection on unobservables.

The fourth follow-up interview included 12,144 respondents. 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 11 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, particularly for college entrance and completion. For example, a 1 SD increase in

NOTLF implies that the probability of completing a bachelor's degree falls by 3.4 percentage points. However, a 1 SD decrease in STEST reduces the probability of completing a bachelor's degree by 16.7 percentage points. Interestingly, one of the measures (AFASK) has a somewhat counterintuitive but weakly significant effect on high school graduation. More specifically, students who are more afraid to ask questions in their academic classes were *more* likely to graduate from high school (though less likely to enter or complete college).

A notable feature of the results in Table 11 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 2006a). The first specification in Table 10 indicates that the estimated effects of the non-cognitive measures are larger in models that exclude the cognitive measure.

In Table 12, we present evidence on how the cognitive and non-cognitive 8th 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 respondent reports that they were engaged in full-time employment in 1999. This variable is defined for the roughly 5,600 respondents who were not students (i.e., those who did not attend a postsecondary institution after January of 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 percent of respondents met this definition of full-time employment. Our second labor-market outcome is the natural log of reported employment earnings for 1999. This measure is defined for the roughly 4,100 respondents 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 (Segal 2006a, Deke and Haimson 2006).

The results in Table 12 indicate that respondents with worse non-cognitive skills in the 8th grade (i.e., higher levels of NOTLF, NOTUSE, and AFASK) are less likely to have been employed full-time in 1999. However, only the effect associated with NOTLF is statistically significant after controlling for the measure of cognitive skills. A 1 SD deviation decrease in NOTLF implies that the probability of full-time employment as a young adult increased by 2.7 percentage points (i.e., roughly 3.2 percent of the mean). Interestingly, this point estimate changes relatively little after conditioning on measures of educational attainment, suggesting these non-cognitive skills have labor-market consequences that are independent of their schooling effects. Lower levels of NOTLF, 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 1 SD decrease in AFASK implies earnings that are approximately 5.4 percent higher. As with the effects of NOTLF on employment, the

estimated effect of AFASK on earnings is similar after controlling for educational attainment.

Comparing Costs and Benefits

The results in Tables 11 and 12 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 8th 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 1 SD 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 1 SD decrease in class sizes would increase the number of classes by 31 percent. Following Krueger (2003b), 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 2002-03 expenditures per pupil in public schools and

adjusting for inflation. Our estimate of the direct per-pupil cost of a 1 SD class-size reduction is then simply 31 percent of this estimate.

To construct a comparable estimate of the monetized benefits of an 8th-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 measure, which appears to have had the clearest impact on earnings. More specifically, using the point estimate from model (3) in Table B, a 1 SD decrease in class size would reduce AFASK by 0.089 (i.e., 0.014 x -5.8675). Using the estimate from column (3) in Table 11, this decrease in AFASK implies that earnings would grow by 0.48 percent (i.e., -.0541 x -0.089). As in Krueger (2003b), we assumed that this earnings impact 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). This age-earnings profile is represented in Figure 1. We then calculated the present discounted values of a 0.48 percent 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 13 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 percent discount rate and 1 percent 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.[19] The internal rates of return for this class-size investment range from 3.6 to 5.6 percent, depending on the assumed productivity growth (i.e., 0 to 2 percent).

The results in Table 10 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 1 SD 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 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).[20] Using the results from Tables 10 and 12, we estimated that a 1 SD class-size reduction in 8th grade would increase earnings by 0.97 percent. That estimated increase reflects the effects of the class-size reduction on both AFASK and STEST. Table 14 presents the present discounted value of this earnings increase under different assumptions about the discount rate and productivity growth. Not surprisingly, the urban-specific results suggest that a class-size investment appears cost-effective under a broader range of assumptions. For example, assuming a 5 percent discount rate and 1 percent 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 percent under the assumption of 1 percent productivity growth.

---

[19] 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.

[20] More specifically, we adjusted costs upward by 3.7 percent, a correction based on data from Table 86 of the 2006 Digest of Education Statistics.

Overall, these results suggest that the apparent cost-effectiveness of an 8th 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 appropriately (e.g., the relevant discount rate). For example, Krueger (2003c) and Summers (2003) discuss 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 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).[21] Finally, an additional uncertainty is that our estimates of the effect of a class-size reduction (e.g., Tables 6 and 10) 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.

[21] Interestingly, the fourth follow-up NELS:88 survey included questions about volunteering and voting. Increases in the 8th grade non-cognitive measures are associated with statistically significant increases in these forms of civic participation.

6. Conclusions

The prevalence of class-size reduction policies in public schools is a powerful testament to their public 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 re-analysis of follow-up data from the Project STAR class-size experiment suggests that assignment to a small class led to improvements in teacher-reported measures of student initiative. However, these estimated effects did not persist to later grades and may be compromised by non-random attrition from the follow-up studies. Our quasi-experimental analysis of nationally representative data on 8th graders provided more definitive evidence that reductions in class size improve some non-cognitive skills related to student engagement. Furthermore, we find qualified evidence that 8th-grade class-size reductions may be cost-effective, in light of the apparent long-term labor-market benefits of these non-cognitive skills. While our cost-benefit comparisons are sensitive to their underlying assumptions, it is notable that 8th-grade class-size reductions appear to be particularly cost-effective when targeted in urban schools. A final and substantive caveat worth underscoring is 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.

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 policy-makers 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 promising 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 indeed, the most attractive way in which 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. Further research may uncover policies and practices that are both effective and, quite possibly, more cost-effective in this regard.

REFERENCES

Angrist, Joshua D. and Victor Lavy. 1999. Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement *Quarterly Journal of Economics* 114: 533-575.

Ashenfelter, Orley and Alan Krueger. 1994. Estimating the returns to schooling using a new sample of twins. *American Economic Review* 84: 1157-1173.

Ashenfelter, Orley and Cecilia Rouse. 1998. Income, schooling and ability: Evidence from a new sample of identifical twins. *Quarterly Journal of Economics* 113: 253-284.

Baumeister, R. F., Campbell, J. D., Krueger, J. I. and 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. and Vohs, K. D. 2004. Exploding the Self-Esteem Myth. *Scientific American*, 20 December.

Boruch, Robert. 2002. The Virtues of Randomness. *Education Next* 2(3): 37-41.

Bowles, Samuel and Herbert Gintis. 1976. *Schooling in Capitalist America: Educational reform and the contradictions of economic life*. New York: Basic Books.

Carneiro, Pedro and James J. Heckman. 2003. Human Capital Policy. In James J. Heckman and Alan B. Krueger. *Inequality in America: What Role for Human Capital Policies?* Cambridge, Mass: MIT Press.

Case, Ann and Angus Deaton. 1999. School Inputs and Educational Outcomes in South Africa. *Quarterly Journal of Economics* 114:1047-1084.

Dee, Thomas S. 2005. A Teacher Like Me: Does Race, Ethnicity or Gender Matter? *American Economic Review* 95(2): 158-165.

Dee, Thomas S. 2007. Teachers and the Gender Gaps in Student Achievement. *Journal of Human Resources* 42(3): 528-554.

Dee, Thomas S. and Sarah Cohodes. 2008. Out-of-Field Teaching and Student Achievement: Evidence from 'Matched-Pairs' Comparisons. *Public Finance Review* 36(1): 7-32.

Deke, John and Joshua Haimson. 2006. Valuing Student Competencies: Which Ones Predict Postsecondary Educational Attainment and Earnings, and for Whom? Princeton, NJ: Mathematica Policy Research.

Duncan, Greg J. and Rachel Dunifon. 1998. 'Soft-skills' and long-run labor market success. *Research in Labor Economics* 17: 123-149.

Dunifon, Rachel, Greg J. Duncan and Jeanne Brooks-Gunn. 2001. As Ye Sweep, So Shall Ye Reap. *American Economic Review* 91(2): 150-154.

Education Commission of the States. 2005. State Class-Size Reduction Measures, Denver, Colorado: Education Commission of the States.

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

Finn, Jeremy D. and Charles M. Achilles. 1990. Answers and Questions about Class Size: A Statewide Experiment. *American Educational Research Journal* 27(3): 557-577.

Finn, Jeremy D., John Folger, and Deborah Cox. 1991. Measuring Participation Among Elementary School Students. *Educational and Psychological Measurement* 51(393-402).

Finn, Jeremy D., DeWayne Fulton, Jayne Zaharias, and Barbara A. Nye. 1989. *Peabody Journal of Education*. 67(1): 75-84.

Finn, Jeremy D., Susan B. Gerber, and Jayne Boyd-Zaharias. 2005. Small Classes in the Early Grades, Academic Achievement, and Graduating from High School. *Journal of Educational Psychology* 97(2): 214-223.

Finn, Jeremy D., Gina M. Pannozzo, and Charles M. Achilles. 2003. The 'Why's' of Class Size: Student Behavior in Small Classes. *Review of Educational Research* 73(3): 321-368.

Finn, Jeremy D., Jayne Boyd-Zaharias, Reva M. Fish, and Susan B. Gerber. 2007. Project STAR and Beyond: Database User's Guide. Lebanon, Tennessee: Hero's, Incorporated.

Fredricks, Jennifer A., Phyllis C. Blumenfeld, and Alison H. Paris. 2004. School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research* 74(1): 59-109.

Glass, Gene and Mary L. Smith. 1978. Meta-Analysis of the Relationship of Class Size and Student Achievement. *Educational Evaluation and Policy Analysis* 1(1): 2-16.

Hanushek, Eric A. 1999. Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis*, 21(2): 143-163.

Hanushek, Eric A. 2003. The Failure of Input-Based Schooling Policies. *Economic Journal* 113(1): F64-F98.

Heckman, James J. 2000. Policies to Foster Human Capital. *Research in Economics* 54: 3-56.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. The Effects of Cognitive and Non-cognitive Abilities on Labor Market Outcomes and Social Behaviors. *Journal of Labor Economics* 24(3): 411-482.

Heckman, James J and Yona Rubenstein. 2001. The Importance of Non-cognitive Skills: Lessons from GED Testing Program, *American Economic Review* 91(2): 145-149.

Howell, William G., Martin R. West, and Paul E. Peterson. 2007. What Americans Think About Their Schools. *Education Next*, 7(4): 12-26.

Hoxby, Caroline M. 2000. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics* 115: 1239-1285.

Ingels, Steven J., Sameer Y. Abraham, Rosemary Karr, Bruce D. Spencer and Martin R. Frankel. 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, 1990.

Jacob, Brian 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.

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

Jepsen, Christopher and Steven Rivkin. 2002. What is the Tradeoff between Smaller Classes and Teacher Quality. NBER Working Paper 9205. Cambridge, Mass.: National Bureau of Economic Research.

Kim, James S., The Relative Influence of Research on Class-Size Policy. In Tom Loveless and Frederick M. Hess, eds., *Brookings Papers on Education Policy, 2006/2007*. Washington, DC: Brookings Institution Press.

Krueger, Alan B. 1999. Experimental Estimates of Education Production Functions. Quarterly Journal of Economics 114: 497-532.

Krueger, Alan B. 2003a. Inequality, too much of a good thing. In James J. Heckman and Alan B. Krueger. *Inequality in America: What Role for Human Capital Policies?* Cambridge, Mass: MIT Press.

Krueger, Alan B. 2003b. Economic Considerations and Class Size. *Economic Journal* 113: F34-F63.

Krueger, Alan B. 2003c. Responses. In James J. Heckman and Alan B. Krueger. *Inequality in America: What Role for Human Capital Policies?* Cambridge, Mass: MIT Press.

Krueger, Alan B., and Whitmore, Diane 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.

Lazear, Edward P. 2001. Educational Production. *Quarterly Journal of Economics* 116(3): 777-803.

Mosteller, F. 1995. The Tennessee study of class size in the early school grades. *The Future of Children* 5:113-127.

Rouse, Cecilia 1999. Further estimates of the economic return to schooling from a new sample of twins. *Economics of Education Review* 18: 149-157.

Segal, Lydia. 2006a. Motivation, Test Scores, and Economic Success. Harvard Business School. Mimeo.

Segal, Lydia. 2006b. Classroom Behavior. Harvard Business School. Mimeo.

Schanzenbach, Diane Whitmore. 2007. What Has Research Learned from Project STAR? In Tom Loveless and Frederick M. Hess, eds., *Brookings Papers on Education Policy, 2006/2007*. Washington, DC: Brookings Institution Press.

Summers, Lawrence H. 2003. Comments. In James J. Heckman and Alan B. Krueger. *Inequality in America: What Role for Human Capital Policies?* Cambridge, Mass: MIT Press.

Voelkl, Kristen E. 1995. Class Size and Classroom Participation. Paper presented at the National Symposium on Class Size Research, London, England.

West, Martin R. and Paul E. Peterson. 2007. *School Money Trials: The Legal Pursuit of Educational Adequacy*. Washington, DC: Brookings Institution Press.

West, Martin R. and Ludger Woessmann. 2006. Which School Systems Sort Weaker Students into Smaller Classes? International Evidence. *European Journal of Political Economy* 22:944-968.

Woessmann, Ludger. 2007. International Evidence on Expenditure and Class Size: A Review. In Tom Loveless and Frederick M. Hess, eds., *Brookings Papers on Education Policy, 2006/2007*. Washington, DC: Brookings Institution Press.

Woessmann, Ludger and Martin R. West. 2006. Class-size Effects in School Systems Around the World *European Economic Review* 50(3): 695-736.

Table 1- Project STAR: Descriptive Statistics for the Grade 4 and 8 Participation Studies

	Grade 4		Grade 8		Project STAR	
Variable	Obs	Mean (SD)	Obs	Mean (SD)	Obs	Mean (SD)
Initiative	2,212	0.000 (1.000)	2,978	0.000 (1.000)	-	-
Effort	2,212	0.000 (1.000)	2,978	0.000 (1.000)	-	-
Non-participatory behavior	2,212	0.000 (1.000)	2,978	0.000 (1.000)	-	-
White/Asian (1= Yes)	2,212	0.762	2,978	0.733	11,467	0.623
Female (1=Yes)	2,212	0.499	2,978	0.529	11,581	0.471
Birth year 1979 (1=Yes)	2,212	0.309	2,977	0.339	11,533	0.340
Free lunch (1=Yes)	2,193	0.410	2,925	0.404	11,334	0.550
Small class (1=Assigned to small class)	2,212	0.333	2,978	0.267	11,601	0.261

Notes: Standard deviations for continuous variables are reported in parentheses. The free lunch variable measures whether the student was eligible for a free or reduced price lunch during his or her entry year.

Table 2 – Project STAR: Assessing Attrition and Balance

Independent variable	Dependent variable: Study Participation		Dependent variable: Small-class Assignment	
	Grade 4	Grade 8	Grade 4	Grade 8
Small Class	0.065††	0.001	-	-
	(0.013)	(0.010)		
White/Asian	-0.000	-0.007	0.003	0.000
	(0.013)	(0.015)	(0.038)	(0.035)
Female	0.019††	0.064††	0.019	0.001
	(0.006)	(0.008)	(0.018)	(0.018)
Birth year 1979	0.016*	0.020†	0.017	0.006
	(0.008)	(0.010)	(0.021)	(0.020)
Free lunch	-0.069††	-0.110††	-0.003	-0.039
	(0.011)	(0.013)	(0.027)	(0.025)
School-by-entry-wave fixed effects?	Yes	Yes	Yes	Yes
P-value for joint significance test	0.000	0.000	0.742	0.540
Sample Size	11,183	11,183	2,193	2,924

Notes: Standard errors, adjusted for school-by-entry-wave clustering, are reported in parentheses. The sample for the models presented in the first two columns includes all Project STAR participants with valid data for each observed student characteristic. The final row reports the p-value for a joint test of the significance of the independent variables listed.

* Statistically significant at the 10-percent level

† Statistically significant at 5-percent level

†† Statistically significant at 1-percent level

Table 3 – Project STAR: Effects of Small Class Assignment on Grade 4 Engagement

Independent Variable	Dependent Variable: Initiative			Dependent Variable: Effort			Dependent Variable: Non-part. Behavior		
Small Class	0.179††	0.113†	0.129††	0.136††	0.072	0.075	-0.102*	-0.043	-0.049
	(0.047)	(0.045)	(0.050)	(0.046)	(0.043)	(0.049)	(0.052)	(0.055)	(0.058)
White/Asian	-0.120	-0.033	-0.038	0.032	-0.032	0.017	-0.130†	-0.193*	-0.288††
	(0.080)	(0.111)	(0.095)	(0.070)	(0.094)	(0.089)	(0.061)	(0.106)	(0.095)
Female	0.289††	0.265††	0.265††	0.432††	0.412††	0.398††	-0.522††	-0.514††	-0.496††
	(0.042)	(0.043)	(0.044)	(0.038)	(0.040)	(0.042)	(0.045)	(0.049)	(0.049)
Birth year 1979	-0.247††	-0.149††	-0.148††	-0.214††	-0.100†	-0.084	0.085*	0.007	-0.005
	(0.056)	(0.052)	(0.054)	(0.056)	(0.050)	(0.053)	(0.049)	(0.050)	(0.051)
Free lunch	-0.392††	-0.380††	-0.338††	-0.353††	-0.343††	-0.314††	0.122††	0.163††	0.119†
	(0.057)	(0.056)	(0.062)	(0.049)	(0.051)	(0.056)	(0.041)	(0.043)	(0.049)
School-by-entry-wave fixed effects?	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Teacher fixed effects?	No	No	Yes	No	No	Yes	No	No	Yes
R^2	0.076	0.204	0.372	0.095	0.219	0.371	0.083	0.180	0.344

Notes: Standard errors, adjusted for school-by-entry-wave clustering, are reported in parentheses.
* Statistically significant at the 10-percent level
† Statistically significant at 5-percent level
†† Statistically significant at 1-percent level

Table 4 – Project STAR: Effects of Small Class Assignment on Grade 8 Engagement

Independent Variable	Dependent Variable (as reported by Math teacher):			Dependent Variable (as reported by English teacher):		
	Initiative	Effort	Non-part. Behavior	Initiative	Effort	Non-part. Behavior
Small Class	-0.038	-0.053	0.080*	-0.026	-0.037	-0.033
	(0.045)	(0.043)	(0.045)	(0.053)	(0.045)	(0.044)
White/Asian	0.123	0.201†	-0.395††	-0.126	0.160†	-0.411††
	(0.078)	(0.084)	(0.100)	(0.085)	(0.069)	(0.092)
Female	0.177††	0.353††	-0.463††	0.301††	0.516††	-0.500††
	(0.039)	(0.037)	(0.041)	(0.042)	(0.039)	(0.038)
Birth year	-0.101†	-0.096†	-0.021	-0.104†	-0.060	0.068
1979	(0.044)	(0.045)	(0.044)	(0.041)	(0.045)	(0.044)
Free lunch	-0.132††	-0.147††	0.148††	-0.202††	-0.266††	0.124†
	(0.041)	(0.070)	(0.049)	(0.050)	(0.046)	(0.049)
R^2	0.227	0.231	0.209	0.273	0.262	0.239

Notes: Standard errors, adjusted for school-by-entry-wave clustering, are reported in parentheses. All models condition on school-by-entry-wave fixed effects.
* Statistically significant at the 10-percent level
† Statistically significant at 5-percent level
†† Statistically significant at 1-percent level

Table 5 – Samples Means, 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
NOTLF	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

Table 6 – NELS:88: Estimated effects of class size on noncognitive and cognitive student outcomes

	First-difference (FD) 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
NOTLF	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	
Control variables										
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	

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.

* Statistically significant at the 10-percent level

† Statistically significant at the 5-percent level

‡ Statistically significant at the 1-percent level

Table 7 – NELS:88: Selection on classroom and teacher observables

		First-difference (FD) estimates			
Dependent variable	(1)	(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
Control variables					
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

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

Table 8 – 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	Change in dependent variable: Science AFASK replaced by math	Change in dependent variable: Reading AFASK replaced by history	Change in dependent variable: History AFASK replaced by reading
Class size in math	0.0153‡	0.0024	0.0145‡	0.0145‡	0.0109†
	(0.0051)	(0.0044)	(0.0051)	(0.0051)	(0.0051)
Class size in science	0.0083*	0.0082*	0.0064	0.0045	0.0104*
	(0.0048)	(0.0048)	(0.0053)	(0.0051)	(0.0053)
Class size in English	0.0201‡	0.0171‡	0.0170‡	0.0104*	0.0103‡
	(0.0053)	(0.0050)	(0.0054)	(0.0055)	(0.0053)
Class size in history	0.0115†	0.0124†	0.0069	0.0111†	0.0040
	(0.0052)	(0.0051)	(0.0060)	(0.0052)	(0.0051)
p-value (H_0: $\beta_M = \beta_S = \beta_E = \beta_H$)	0.3669	0.0920	0.3968	0.5666	0.1998

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

* Statistically significant at the 10-percent level

† Statistically significant at the 5-percent level

‡ Statistically significant at the 1-percent level

Table 9 – Estimated effects of grade-8 class size on subsequent effort, 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

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-percent level
† Statistically significant at the 5-percent level
‡ Statistically significant at the 1-percent level

Table 10 – NELS:88: Class Size Effects by Student and School Traits

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

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.

* Statistically significant at the 10-percent level

† Statistically significant at the 5-percent level

‡ Statistically significant at the 1-percent level

TABLE 11 – Estimated effect of noncognitive and cognitive measures on educational attainment, NELS:88 Fourth Follow-up

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

Standard errors, adjusted for school-level clustering, are reported in parentheses.
* Statistically significant at the 10-percent level
† Statistically significant at the 5-percent level
‡ Statistically significant at the 1-percent level

TABLE 12 – Estimated effects of noncognitive and cognitive measures on labor-market outcomes, NELS:88 Fourth Follow-up

	Full-time Employment (1999)				ln(1999 Earnings)			
Independent variables	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
NOTLF	-0.0217‡	-0.0179†	-0.0269‡	-0.0223†	-0.0273†	-0.0210	-0.0216	-0.0144
	(0.0081)	(0.0081)	(0.0094)	(0.0095)	(0.0133)	(0.0134)	(0.0152)	(0.0151)
STEST	-	0.0457‡	0.0368‡	0.0156*	-	0.0905‡	0.0859‡	0.0414‡
		(0.0068)	(0.0079)	(0.0087)		(0.0099)	(0.0124)	(0.0139)
NOTUSE	-0.0160†	-0.0103	-0.0130	-0.0077	-0.0086	0.0022	0.0051	0.0155
	(0.0077)	(0.0078)	(0.0090)	(0.0091)	(0.0108)	(0.0109)	(0.0135)	(0.0134)
STEST	-	0.0456‡	0.0370‡	0.0158*	-	0.0917‡	0.0876‡	0.0429‡
		(0.0069)	(0.0079)	(0.0088)		(0.0099)	(0.0124)	(0.0139)
AFASK	-0.0168†	-0.0093	-0.0077	-0.0075	-0.0631‡	-0.0495‡	-0.0541‡	-0.0512‡
	(0.0073)	(0.0073)	(0.0086)	(0.0088)	(0.0121)	(0.0122)	(0.0158)	(0.0158)
STEST	-	0.0453‡	0.0374‡	0.0153*	-	0.0844‡	0.0791‡	0.0345†
		(0.0069)	(0.0079)	(0.0087)		(0.0099)	(0.0128)	(0.0145)
Dependent mean	0.84				10.19			
Control variables								
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

Standard errors, adjusted for school-level clustering, are reported in parentheses.
* Statistically significant at the 10-percent level
† Statistically significant at the 5-percent level
‡ Statistically significant at the 1-percent level

Figure 1 - Age-Earnings Profile - 2007 March CPS

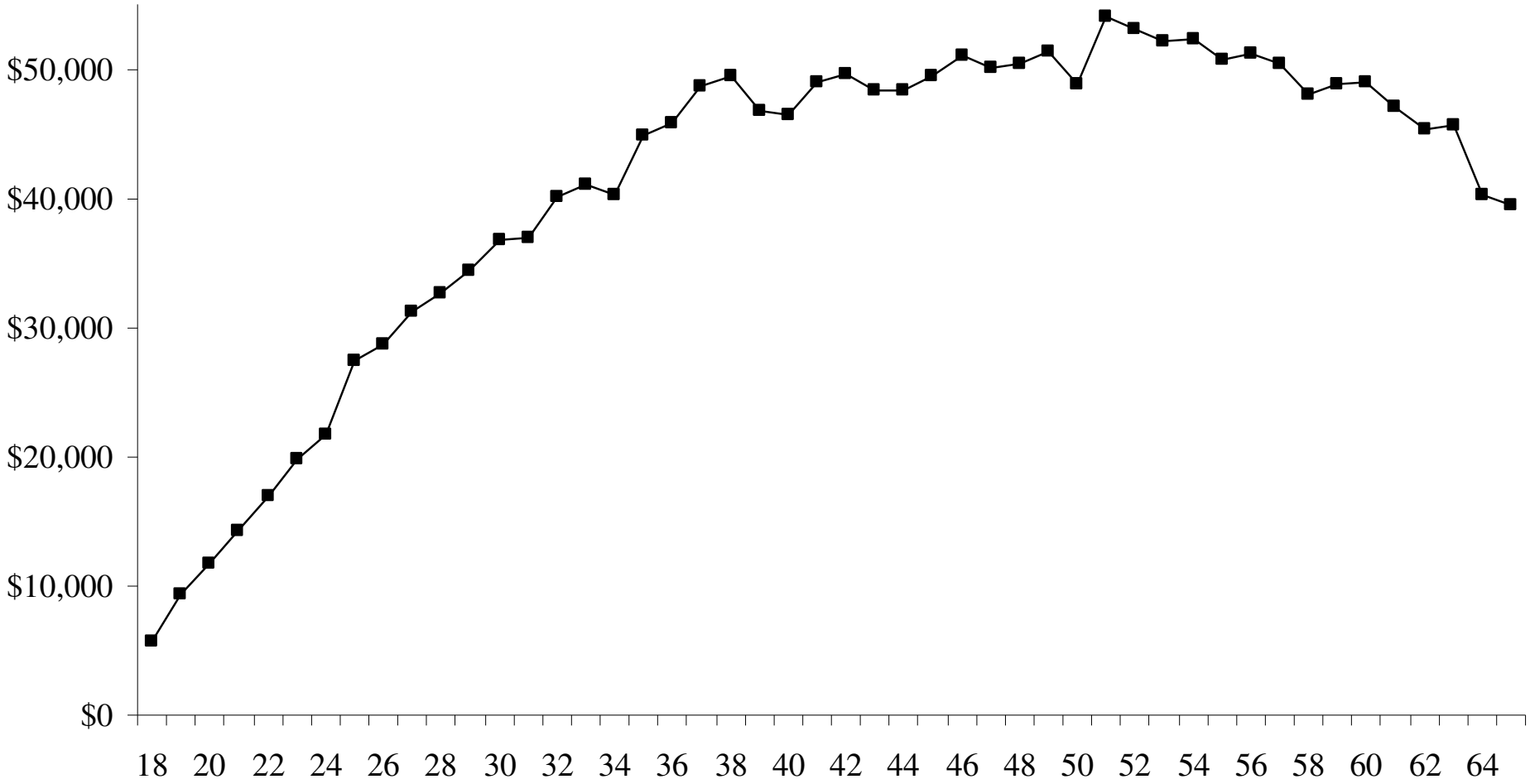

Table 13 - Present Discounted Value of Increased Earnings from Reducing 8th-Grade Class Size by 1 Standard Deviation

	Assumed Productivity Growth Rate		
Discount 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

Notes: 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 1 SD 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 1 SD class-size reduction is estimated as $3,392 in 2006 dollars.

Table 14 - Present Discounted Value of Increased Earnings from Reducing 8th-Grade Class Size by 1 Standard Deviation in Urban Schools

	Assumed Productivity Growth Rate		
Discount 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

Notes: 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 1 SD 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 1 SD class-size reduction is estimated as $3,157 in 2006 dollars.