Method

Participants: To catalogue change in test scores from students from 4th to 8th grade, scores were needed from both time-points. We were unable to obtain 4th-grade test scores for 485 students in our sample. Some cognitive data were missing from an additional 195 students due to time constraints during data collection.

Processing Speed: The number of items completed correctly in 2 minutes was the dependent measure for each task. These measures were merged to yield a standardized processing speed score in accordance with WISC-IV scoring procedures. All data were scored within the 13 years 0 months and 13 years 3 months age range in order to standardize scoring.

Working Memory: Each display included 3–9 targets (blue circles), 1–5 circular distractors (red), and 1–9 color distractors (blue triangles), which varied independently. Students were told not to use their fingers or to write anything down before the "?" prompt. Span was determined by scoring each response (correct numbers in the correct order) and defined as the largest load in which at least 2 out of 3 problems were correctly answered. For example, a student would be given a span of 3 if they got 2 or 3 correct at loads of 1, 2 and 3, but only 1 or 0 correct at a load of 4.

Fluid Reasoning: Scoring was in accordance with TONI scoring guidelines. Ceiling was reached when 3 of 5 consecutive items were incorrect; basal was the highest item after 5 consecutive correct responses; the number correct was counted within the range of the basal and ceiling scores to yield a raw score, which was converted to a standard score based on the average age of the children in the sample, 13.

Validation: A subset of students were tested individually on version B of the TONI (n = 17) and a computer-based, self-paced count span WM task (n = 15). Performance for these students was highly correlated between group testing in schools and individualized testing in the laboratory testing (WM: r=.56, p=.029; FR: r=.55, p=.023; Figure S2).

Behavioral data collection: Prior to each measure, students completed practice problems with the proctor. Only when students completed the practice problems and had a chance to ask questions did the proctor move on. Because timing was important, students were told not to pick up their pencils or turn pages until instructed to do so by the proctor. When the time limit was up the proctor would say "pencils down, hands up."

Data Analysis

Relations of schools to MCAS scores and cognitive measures: We conducted an analysis of variance to determine the share of the overall variation in student achievement and cognitive skills explained by the school attended in 8th grade. For a given score, we fit the following multilevel model:

$$\text{(I)} \quad Y_{is} = \alpha A_{i,t-4} + \pi X_i + (\upsilon_s + \varepsilon_{is}) \,,$$

where Y_{is} represents a given test score or measure of cognitive skill for student *i* in school *s*. We included lagged 4th -grade scores in math and ELA, $A_{i,t-4}$. We also included a vector of student demographic characteristics, X_i, consisting of controls for gender, race, age, free and reduced-priced lunch status, limited English proficiency, and special education status. This multilevel model allowed us to estimate unique variances associated with our two-level error structure where individual students are nested within

schools. Estimates of the variance of υ_s over the total error variance of the model provided the share of the overall variation in each outcome that was at the school level.

As depicted in Figures 2b-d, we tested the robustness of these results in three ways. First, we refit model (I) and included controls for 8thgrade MCAS scores in math and ELA for models with cognitive outcomes, and our composite cognitive measure as a control for models with test score outcomes. Second, we replicated the analyses just described but excluded 4thgrade test scores. Finally, we refit model (I) but excluded students attending the two exam schools.

Average improvement in test scores by school: To show which schools in our sample were associated with the largest average gains on standardized tests, we plotted the average student growth percentile (SGP) separately for each school and indicated whether it was an over-subscribed charter, undersubscribed charter, exam, or traditional school. Student growth percentiles are a metric of growth in academic achievement calculated by identifying all students in the state whose previous scores are similar and comparing these students' scores on the next grade-level test. The growth of each student is measured relative to that of other students with similar test-score histories and expressed as a percentile. We took the average SGP for all of the students in the school for grades 6-8 in 2011, the year students in our sample were studied, as a estimate of overall school relation to SGP.

Figures:

Figure S1: Lottery Applicant Sample

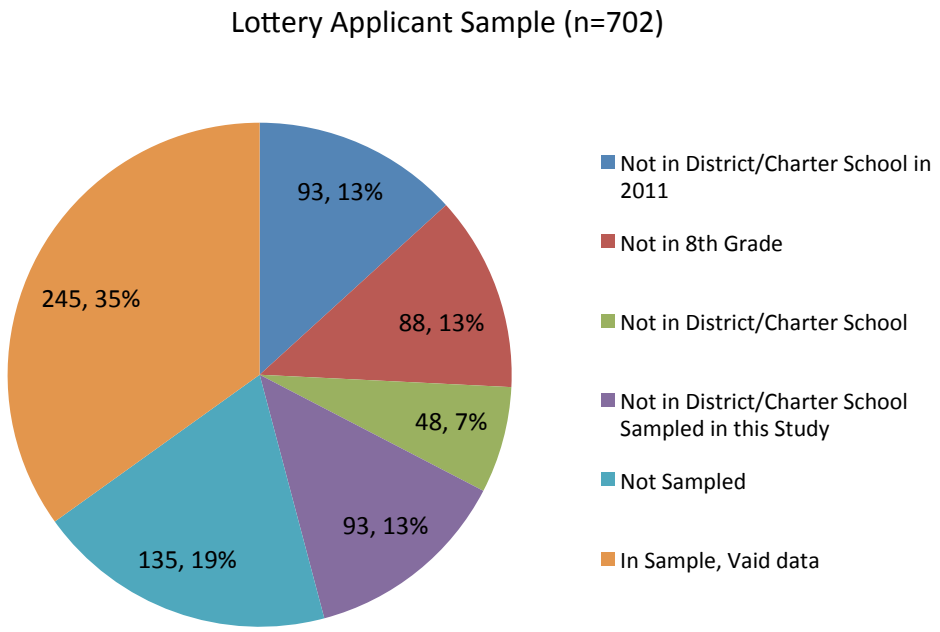

Figure S2: Classroom and in-lab measures of working memory and fluid reasoning

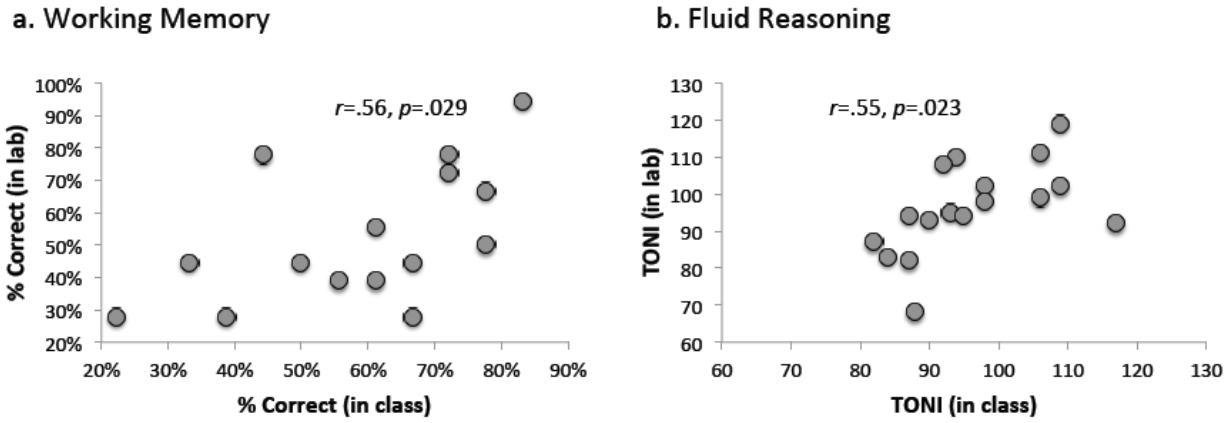

Figure S3: Path Analysis 4th and 8th grade MCAS

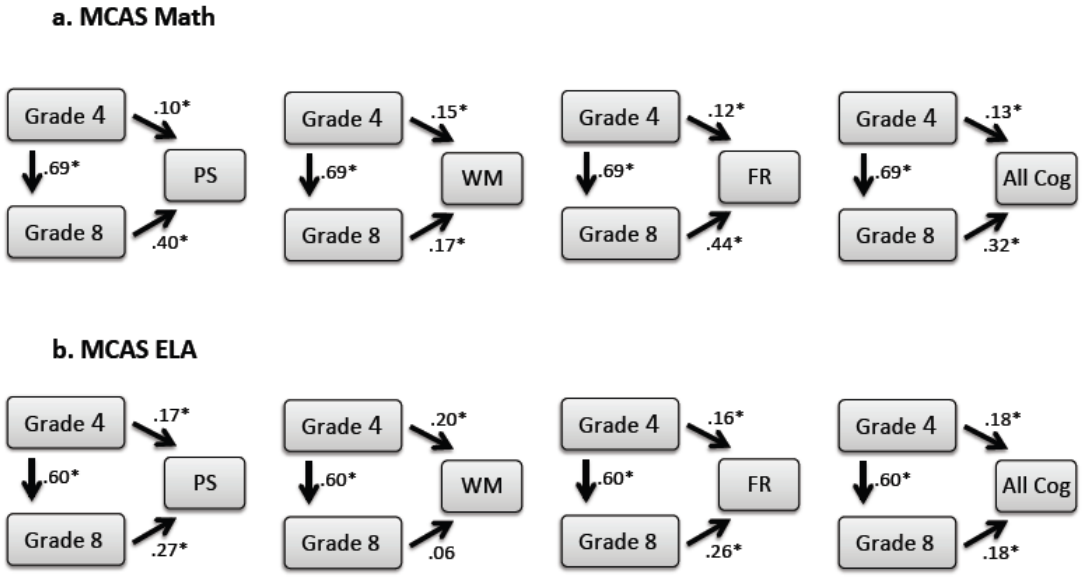

Figure S4: Average improvement on test scores by school.