Student Abilities During the Expansion of US Education: Online Appendix

Lutz Hendricks* Todd Schoellman†

January, 2014

Abstract

Details are provided for the construction of the moments from the Census and the NLSY. Details are also provided for the time series data on test scores of college-bound and non-college-bound high school seniors. These details include references for the underlying studies containing the raw data; metadata on the studies; a description of the process of transforming the results of the studies into a single time series; and robustness results.

*University of North Carolina, Chapel Hill. Address: UNC at Chapel Hill, Department of Economics, 6C Gardner Hall, CB 3305, Chapel Hill, NC 27599-3305. Phone: (919) 966-5328. E-mail: lutz@lhendricks.org.

†Arizona State University. Contact: Department of Economics, W.P. Carey School of Business, 501 E. Orange Street, Tempe, AZ 85287. Phone: (480) 965-7995. E-mail: todd.schoellman@gmail.com.

A Construction of Moments from the Census

Samples. The census is taken every ten years, which provides data from many birth cohorts spanning multiple censuses. The analysis is confined to the cohorts that are exactly 40 years of age in the 1950–2000 censuses, or the cohorts born every ten years from 1910–1960. Focusing on one age eliminates any problems associated with comparability of educational data and wages at different ages. The data used are the public-use micro data files available from Ruggles et al. (2009). These files are a 1% sample from 1950–1970 and 5% samples thereafter. In 1950, only sample line individuals report wages and hours worked. This reduces the effective sample size to only one quarter of the 1960 sample. Table 1 shows descriptive statistics for each census year.

Educational attainment. Educational attainment is derived from two variables. The variable used in the 1990 and 2000 censuses is EDUCREC (detailed), which records the degrees obtained by respondents. High school graduates includes those with GEDs, while some college includes those with two-year degrees. For earlier censuses the only available variable is HIGRADE (again detailed), which records the number of years of education a person has obtained. Those with fewer than 12 years of schooling are coded as high school dropouts; those with exactly 12 years as high school graduates; those with 13–15 years as having some college; and those with at least 16 years of schooling, college graduates.

Cohorts that respond to both the HIGRADE and EDUCREC questions (in two different censuses) typically have different measured attainment for the two questions, even if they are old enough that significant changes in actual school attainment are unlikely. Goldin and Katz (2008) use Current Population Survey data to produce a more detailed concordance between EDUCREC and HIGRADE questions that they use between 1980 and 1990. The raw responses are used in this paper for two reasons. First, the concordance is likely to vary by year as the structure of education changes, especially within each of the four discrete categories. Second, the focus here is on the large-scale movements, such as the near-universality of high school graduation and the increase in college attendance. Since most of those identified as dropouts in the 1910 cohort report less than 11 years of schooling, it is quite likely that they did not achieve a high school degree, let alone start college. By contrast the 1960 cohort answered directly about degree completion. It is thus likely that the major trends are real and are not the artifact of changing data collection.

Wages. Hourly wages are calculated as the ratio of wage and salary income (INCWAGE) to annual hours worked. Annual work hours are the product of weeks per year times hours

per week. For consistency, intervalled weeks and hours are used for all years. Usual hours worked per week are used where available. Wages are computed only for persons who report working "for wages" (CLASSWKR) and who work between 520 and 5110 hours per year. All dollar figures are converted into year 2000 prices using the Bureau of Labor Statistics' consumer price index (CPI) for all wage earners (all items, U.S. city average).

B Construction of Moments from the NLSY79

Sample. The sample includes white males. Individuals whose school completion status is unclear are dropped. Likewise, individuals who completed schooling past the age of 34 or who did not participate in the ASVAB aptitude tests are dropped. Observations are weighted.

Schooling. The sample is divided into the four school categories according to the highest degree completed. Those who attended two-year colleges only are classified as having "some college". The last year in school is defined as the start of the first three-year spell without school enrollment.

Wages. Hourly wages are calculated as the ratio of labor income to annual hours worked. Labor income includes wages, salaries, bonuses, and two-thirds of business income. Wage observations prior to the last year of school enrollment, with hours outside the range $[520, 5110]$, or with wage levels outside the range of $[0.02, 100]$ times the median wage are dropped. Wages are deflated by the CPI.

Log-wages are regressed on experience and region of residence separately for each year and school group to remove variation due to demographic characteristics not captured by the model. Log-wages are also regressed on years of schooling within school group to remove within-school-group variation in wages (such as the difference between high school dropouts who completed tenth versus eleventh grade). The resulting adjusted (residual) wages are used throughout.

In order to be consistent with the Census data, all results are reported for wages as of age 40. Some people are not interviewed at age 40. Interpolation between age 39 and age 41 wages is used in such cases.

AFQT. The measure of standardized test score in the NLSY79 is the 1980 Armed Forces Qualification Test (AFQT) percentile rank (variable R1682). The AFQT aggregates various

ASVAB aptitude test scores into a scalar measure. The tests cover numerical operations, word knowledge, paragraph comprehension, and arithmetic reasoning (see NLS User Services (1992) for details). AFQT scores are regressed on the age at which the test was administered and the residual is used to remove the effects of age at testing.

C Test Scores and College Attendance Data

C.1 Construction of the Basic Data and Trends

A central claim of the paper is that there is a growing gap in test scores between students who continue to college and those who finish with high school graduation. This claim rests primarily on data drawn from numerous studies performed by psychological and educational researchers between World War I and 1965. These figures are augmented with calculations made using the nationally representative studies that became available after 1960. This appendix documents the details of the pre-1965 studies and how they are harmonized into the data points used in the paper.

The studies used vary somewhat in size, scope, and methodology. However, it is useful to describe a typical study, its results, and how its results are used. Deviations from the norm are discussed below. Generally, a researcher interested in conducting a study secured the aid and cooperation of the educational authorities of a state. Most researchers were affiliated with the education department or equivalent at a major state university, which facilitated such cooperation. Through his or her connections, the researcher would arrange for many or all of the students of a state to take an exam in a fairly short period of time. In most cases, this would be in the senior year of high school, although a couple of studies used tests given in the junior year and one went as early as eighth grade. These tests were collected and scored.

The second phase of the study involved collecting data on further schooling. Here there is some methodological divergence. Some studies – typically those done in the spring of the senior year of high school – simply asked students about their plans for college. Others arranged to collect schooling data by following up at a later date. In practice later typically meant one year later, although some authors waited several years, and the modern longitudinal data allow researchers to follow up into the children's thirties or forties. For studies that followed up, some tracked students directly, but this was the most challenging method. More common and more feasible alternatives involved asking students' parents or the school administration at their former high school what had become of them.

These two phases gave researchers the raw data on schooling and test scores that provide the motivating fact for this paper. The raw data do not seem to exist for any paper published before 1960. Instead, what remains are published cross-tabulations of the two variables. This includes ranges of test scores (or percentiles of test scores) and, for each, the number or fraction of students who stop their education with high school graduation versus continue to college. The results are derived from a collection of tabulations or calculations along these lines from a number of journal articles, books, dissertations, and unpublished technical reports.

These tabulations or calculations are standardized to a common metric. The benchmark metric follows Taubman and Wales. Ranges of test scores are converted to ranges of percentiles (if they are not reported that way). The average percentile rank of the two education groups is computed by assigning to each range the midpoint and then taking the weighted average across test score groups. A simplified example may clarify the procedure. Suppose that there are only two categories, above median and below median, and that 60% of those above median went to college but only 40% below median did so. In this case, the average percentile rank for college students would be $0.6 \times 75 + 0.4 \times 25 = 55$th percentile. Following a similar calculation, the average percentile rank for high school graduates would be 45th.

This simple example hides a few complications that arise in standardizing the studies. Two are particularly common. First, some studies that ask students about their future plans allow the students to report that they do not know or are unsure. Such students are excluded entirely from calculations. Evidence provided below shows that this decision is not critical for the results. Second, there is time-varying heterogeneity in the set of post-high school educational opportunities that needs to be handled consistently. The approach used here is to include in the college group students who attend college, university, normal schools, or technical schools. Normal schools are what are now called teachers' colleges. Technical schools historically were schools that focused on scientific or industrial careers, and includes for example MIT. Any short (less than one year) or vocational school, including the old business schools (which taught typing and so forth), are excluded from the group of college.

Table 2 gives the resulting average percentile rank for each group for all of the studies. Figure 1a plots the results against birth cohort; Figure 1b plots the gap $\bar{A}_c - \bar{A}_{nc}$ against birth cohort. A quadratic trend is fit for each figure. The quadratic trend for the difference suggests that the two groups were separated by roughly ten percentage points at the turn of the century, but that the gap grew to twenty-five to thirty percentage points for cohorts

born after World War II.

Figure 1: Changes in Test Scores Over the Twentieth Century

(a) Level

(b) Difference

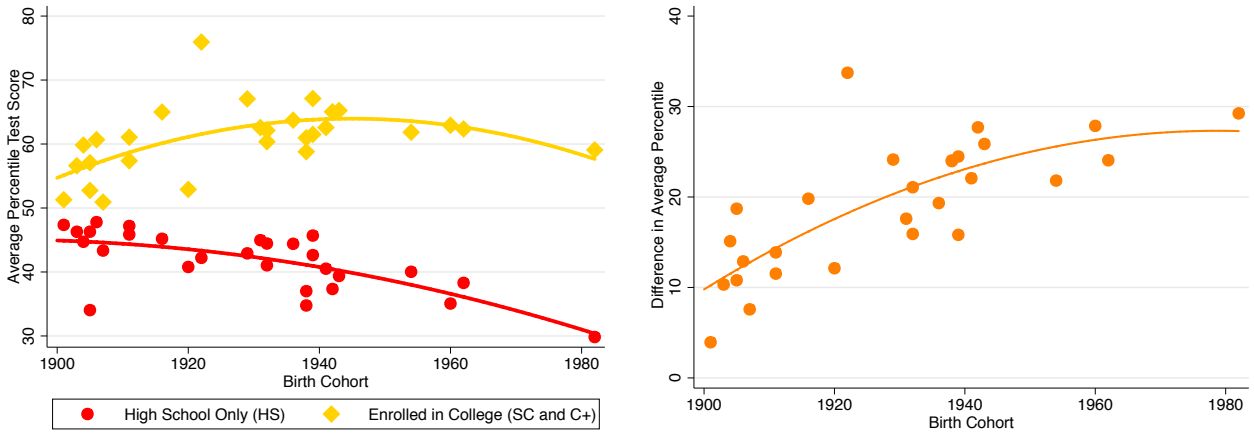


Table 2 lists the references for the data and the data points that result from the calculations described above. It also lists some of the most important metadata on the studies, which is useful for understanding how studies deviate from the typical study described above and for thinking about the comparability of the results from different studies. The Location column gives the geographic scope of the study. Most covered an entire state, but two covered cities or regions of a state. For the later period it is sufficient to focus on representative samples with national scope. The Breadth column explains the sampling framework. In some cases the researchers cooperated with the state's educational authorities and procured statewide results covering practically all schools and students; such cases are denoted as statewide. In other cases schools were sampled or all schools were invited to participate but a non-trivial fraction did not. Such cases are denoted as samples; the sample is called large if it contains more than roughly half of the state's students at the time. For most cases the researchers carefully documented that the schools were diversified geographically and that they represented the school size distribution fairly well. Finally, a couple of the early studies were conducted using schools that had in some way expressed an interest in the study and hence were convenient to the researcher. Such cases are denoted as selected because there is no notion of representativeness of the schools.

The studies used a number of different testing instruments, which are recorded in the column Test. Although the early studies appear to use a wide variety of instruments, in truth most were using IQ tests strongly influenced by the Army Alpha test or the Stanford-Binet

test that were then fashionable. Because of this, they all measure very similar concepts and have very similar results. Franzen (1922) documents that administering the Mentimeter, Terman, Otis, and Haggerty exams to the same student yields highly correlated results; the cross-exam correlations ranged from 0.7–0.91. These results are in the same range as the cross-exam correlations for modern tests as documented in Herrnstein and Murray (1994) and discussed in the main paper. Later in the period, the ACE became the dominant exam for educational testing. It was used for six different studies in the sample.

One potential concern with the time series comparisons is that the quality of the different tests used over time may have varied systematically. For example, the rising test score gap could be explained by a constant ability gap but an increase in test quality, as judged by the validity or signal-to-noise ratio. The available evidence suggests that this is not the case. Garrett (1949) and Fishman and Pasanella (1960) are extensive reviews of the empirical literature on the predictive validity of test scores. The most commonly used metric is the correlation between test scores in high school and subsequent college grades.[1] Both studies report similar correlations that differ little between tests labelled as achievement or ability tests. Garrett (1949) cites a median correlation of 0.49 for achievement tests and 0.47 for ability tests; Fishman and Pasanella (1960) produces almost identical estimates. Referring to the specific tests in Table 2, the Otis and the Terman (two of the earliest tests) have average correlations of 0.48 and 0.46; the ACE (the most commonly used test) is reported as 0.49. In some cases the earliest studies produced independent estimates of the validity of their test; for example, Book (1922) reports that his test has a correlation with high school grades of 0.49 (for a single high school, reported originally in Rice (1920)), while Colvin and MacPhail (1924) reports a correlation of 0.43 with high school grades.

To put these figures into perspective it is useful to compare them to similar results from more modern exams. Perhaps the most widely used predictive instrument today is the SAT. That test has reported correlations with college grades of 0.35 to 0.45 (Morgan, 1989; Kobrin et al., 2008), in the same range as the older exams. Another useful comparison is the correlation between AFQT and high school grades, which Borghans et al. (2011) report at 0.54, just slightly higher than typical figures from previous cohorts. This comparison is useful because AFQT score is the test score used for the NLSY cohorts. These metrics suggest that it is useful to compare results from the different testing programs because there is little variation over time in how well the exams predict pertinent outcomes.

The column Cohort gives the birth cohort that the data represent. In most cases

[1]Specifically, most studies use first semester or freshman year grades. The typical procedure is to compute the correlation college-by-college, and then to take the average across colleges.

this is computed as year of graduation – 18. The column Type captures the variation in how school data were collected. Researchers used one of two main designs. In the first, they asked students while in high school about their intentions to attend college; this is labeled the prospective method. In the second, they follow up with students or other knowledgeable parties about the students' actual school-going behavior; this is labeled the follow-up method. Many studies followed up during the next academic year; these are labeled as 1-year follow-ups. Any that followed up over a longer term (from two years later through adulthood) are called several-year follow-ups. Several of the studies mixed or re-examined the importance of using these different approaches, including primarily Odell (1927) and Barker (1937). Barker (1937) provides the clearest classification of the data. He shows that, in terms of mean test score, the groups can be ranked in order as:

1. Those who attend college immediately after high school.
2. Those who attend college after a delay.
3. Those who indicate a plan to attend college but do not follow through.
4. Those who never plan to or attend college.

The result is that studies that follow up after several years are needed to portray test score gaps precisely. One-year follow-up studies tend to overstate the gap because those who attend in the first year are the highest scoring, while prospective studies tend to understate it because they include some students who do not score as well and will ultimately not attend college. The data in Odell (1927) is useful for quantifying this point, since he provides cross-tabulations of test scores and college-going, measured both using plans and through a one-year follow-up. The former indicates a test score gap of 6.5 percentage points while the latter indicates a gap of 12.8 percentage points; the preceding discussion suggests that a true estimate that would be yielded by a multi-year follow-up would lie somewhere in the middle. An important take-away is that the plausible range is small relative to the total trend change in the gap observed during this time period.

The baseline metric is the average percentile score, computed using the reported cross-tabulations between test score and schooling. These cross-tabulations are discretized versions of the true joint distribution, and hence the accuracy of the calculation is affected by the level of available detail. The number of test score bins is recorded in the column No. Bins. For example, a study that grouped test scores into deciles would be coded as having ten bins. For the calculations from the NLSY79 and NLSY97 the underlying raw data were used, and hence the number of bins is meaningless.

This completes the description of the studies and their metadata. The next section describes exploration of the robustness of the findings along a number of dimensions.

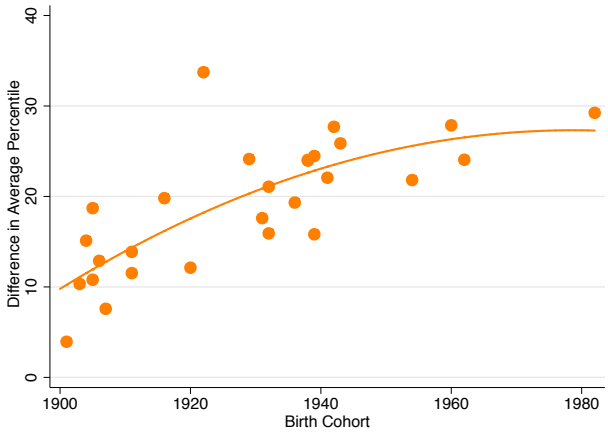
C.2 Robustness

The baseline metric of the ability gap is the difference in the average test score percentile of those who go to college and those who do not. To compute this metric it is necessary to use the discretized test score distributions as explained in the last section. This section begins with evidence that the main conclusion is robust to this step.

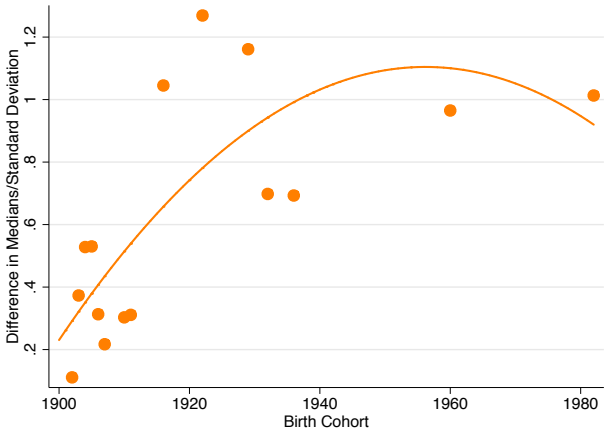
The findings are robust to considering an alternative metric labeled the standardized difference (reported as S.D. in Table 2). The standardized difference is the gap in median or mean test scores between college-bound and non-college-bound high school seniors divided by the standard deviation in test scores among the two groups pooled.[2] An additional benefit of focusing on this metric is that it allows for the use of information from several studies that do not report cross-tabulations of test scores, including especially the seminal study of Learned and Wood (1938). Figure 2 compares the results from the two metrics. Each shows a pronounced rise in the test score difference between the two groups, although the two metrics disagree somewhat on the timing of when the test score gap widened.

Figure 2: Robustness: Alternative Metrics

(a) Difference in Mean Percentile



(b) Normalized Difference in Median Score

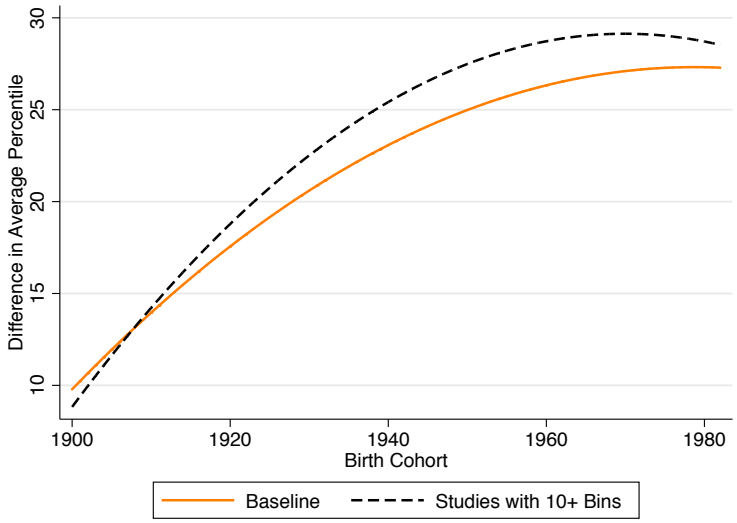


The findings are also robust to using the baseline metric but handling samples with a

[2] A few cases provide the interquartile range and not the standard deviation. Since test scores are closely approximated by a normal distribution, it is possible to use the stable relationship that the standard deviation is 0.7413 of the interquartile range to approximate the standard deviation closely.

small number of bins in various ways. The baseline metric can only generate test score gaps between college-bound and non-college-bound high school graduates using between-bin differences because it assigns the two groups the same test score conditional on being in the same test score bin. Although this is unlikely to be a significant problem when the original study reported twenty bins, it has the potential to affect calculations based on studies with three or four bins. Two alternative series are provided. The first uses only studies with ten or more bins. The second uses a linear probability model to estimate the probability of attending college as a function of test score, where the number of data points in the regression is equal to the number of bins in the original study. A linear functional form is parsimonious and fits the results of most studies well. The average percentile test score for college-goers and non-college-goers can then be estimated using the parameter estimates from the linear probability model. This approach is useful because the linear function accounts for within-bin test score gaps by assuming that they follow the same underlying process as the between-bin test score gaps observed in the data. This seems the natural assumption since at some level the decision by researchers on how to group the data is arbitrary. A quadratic trend is estimated for each of these alternative ways of producing the baseline metric.

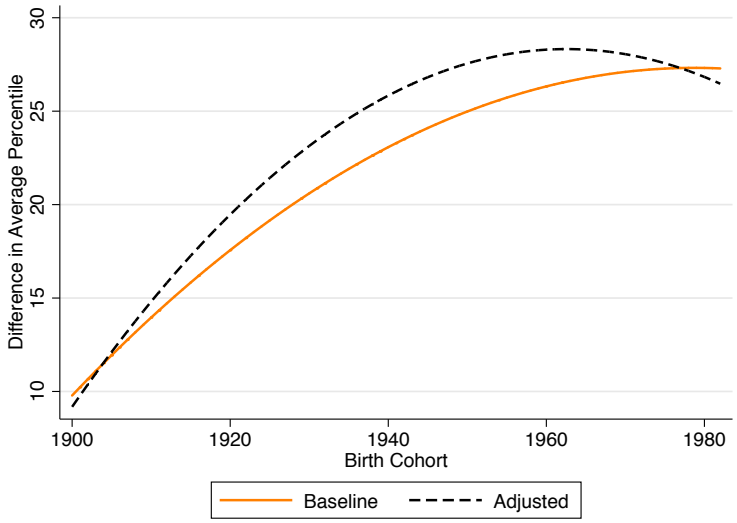
Figure 3: Baseline vs. Studies with 10+ Bins



The results of these two robustness checks are given in figures 3 and 4. The main message from the two figures is quite similar. Each shows that the rise in test scores is robust to different ways of trying to discard or adjust for the small number of bins in some

of the studies. Indeed, both suggest that performing such an adjustment leads to a larger measured rise in the test score gaps between the two groups. The main result is not sensitive to changes along this margin.

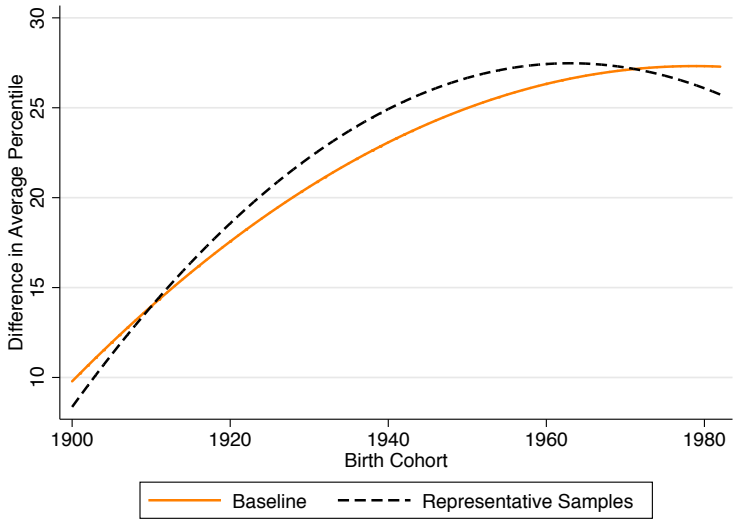
Figure 4: Baseline vs. Linear-Adjusted Test Score Gaps



The next check is to show that the results are robust to sampling procedures. Although many early studies made careful efforts to sample a diverse set of schools in terms of geography, metropolitan status, and so on, some did not. Further, it is impossible to verify that the sampling was ideal since the raw data no longer exist. However, sampling procedures are not likely to be an issue for two sets of studies. First, some studies sampled essentially the entire state, minus students who were absent on the given day. Second, some of the modern studies (such as the HS&B or the NLSY) have carefully designed, nationally representative samples. Figure 5 compares the quadratic trend implied by these samples to that implied by using all the samples. They are quite similar, suggesting that the results are not driven by the sampling framework.

The remaining robustness checks are intended to show that the basic conclusion survives focusing on subsets of studies that are more comparable along certain dimensions. First, the results apply in the time series within a state, for states that were tested multiple times. Likewise, the results apply in the time series within a particular test, for tests that were administered multiple times. These results are given in Figure 6. Figure 6a shows the result only for states that were tested multiple times at least ten years apart, which includes Wisconsin, Iowa, and Minnesota, as well as national studies. The states are labeled to make

Figure 5: Statewide or Nationally Representative Samples

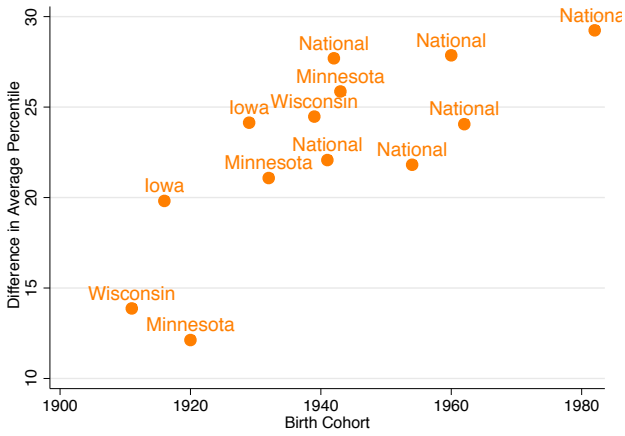


the comparison clearer. This figure shows that large test score gaps arose even within given states. Figure 6b shows the results only for tests that were used multiple times at least ten years apart. Again, tests are labeled to make the comparison easier. This figure shows that large test score gaps arose even within given tests.

Finally, the previous section documented that different studies used different methods to collect schooling data for students and that the collection method used affects the measured test score gap. The final robustness check shows that the results apply within a school data collection methodology. To demonstrate this, a linear trend is fitted separately for each of the three methodologies: prospective, 1-year follow up, and several year follow-up. A linear trend is used because there are only eight studies with two of these designs, which makes estimating a quadratic unreasonable. Figure 7 shows that all the subsets of studies agree on a robust upward trend in test score gaps. The estimated coefficient is statistically significant for all of the study designs except for 1-year follow-ups; that trend is also significant after excluding the clear outlier of Livesay (1942)'s Hawaii study. The results are robust to the school data collection methodology.

Figure 6: Robustness: Subsamples of the Studies

(a) States with Repeat Observations



(b) Tests with Repeat Observations

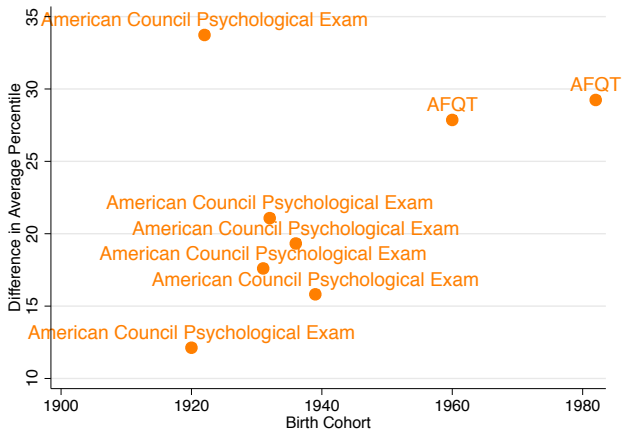


Figure 7: Robustness: Different Methods of Collecting School Data

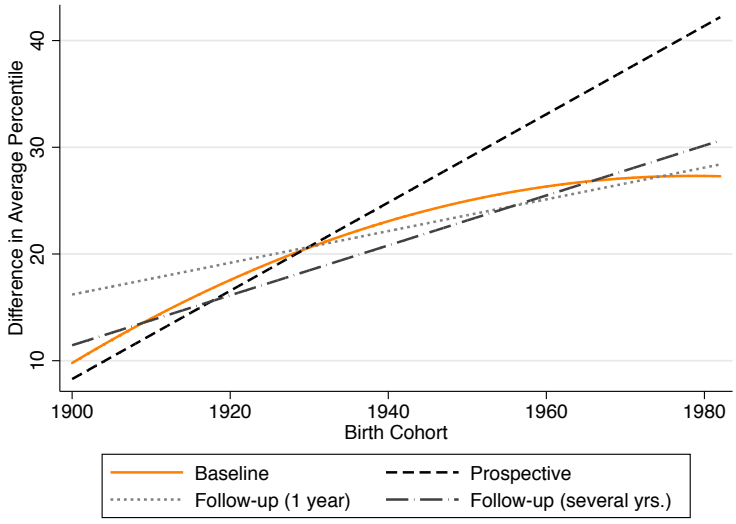


Table 1: Summary statistics: Census data

	Census Year					
	1950	1960	1970	1980	1990	2000
Number of Observations	17,503	74,744	72,873	379,087	539,145	562,262
Fraction <HS	59%	45%	35%	20%	10%	10%
Fraction HS	23%	32%	35%	39%	25%	31%
Fraction SC	9%	11%	12%	17%	32%	30%
Fraction C+	9%	12%	18%	25%	33%	29%
$w_{<HS}$	9.6	13.0	15.3	12.4	12.7	12.0
w_{HS}	11.4	15.7	18.3	16.4	16.7	15.2
w_{SC}	13.4	18.0	21.1	18.6	19.3	18.1
w_{C+}	16.5	22.7	28.5	25.1	26.5	25.4
College wage premium	0.37	0.37	0.44	0.42	0.47	0.52

Table 2: Test Score-Education Data Sources and Details

Source	Location	Breadth	Test	Cohort	Type	No. Bins	$\bar{A}_c$	$\bar{A}_{nc}$	S.D.
Book (1922)	Indiana	Statewide	Indiana University Intelligence Scale	1901	Prospective	10	51.3	47.4	
Batson (1922)	South Dakota	Selected	South Dakota Group Intelligence Test	1902	Prospective				0.111
OBrien (1928)	Kansas	Large Sample	Terman Group Test of Mental Ability	1903	Follow-up (several yrs.)	17	56.6	46.3	0.373
OBrien (1928)	Kansas	Large Sample	Terman Group Test of Mental Ability	1904	Follow-up (several yrs.)	17	59.8	44.7	0.528
Mann (1924)	North Carolina	Selected	Mentimeter	1905	Prospective	20	52.7	34.0	
Colvin and MacPhail (1924)	Massachusetts	Large Sample	Brown University Psychological Exam	1905	Prospective	3	57.1	46.3	0.530
Odell (1927)	Illinois	Large Sample	Otis Test of Mental Ability	1906	Follow-up (1 year)	15	60.7	47.8	0.313
Ames (1926)	Montana	Large Sample	Otis Test of Mental Ability	1907	Prospective	13	50.9	43.3	0.217
Learned and Wood (1938)	Pennsylvania	Large Sample	Otis Test of Mental Ability	1910	Follow-up (several yrs.)				0.303
Benson (1942)	Minneapolis	Large Sample	Haggerty Intelligence Examination	1911	Follow-up (several yrs.)	15	57.4	45.9	0.311
Henmon and Holt (1931)	Wisconsin	Statewide	Ohio State University Psychological Test	1911	Follow-up (1 year)	20	61.1	47.2	
Barker (1937)	Iowa	Large Sample	Iowa Every-Pupil Test	1916	Follow-up (several yrs.)	8	65.0	45.2	1.045
Wolfle and Smith (1956)[a]	Minnesota	Large Sample	American Council Psychological Exam	1920	Follow-up (several yrs.)	10	52.9	40.8	
Livesay (1942)[b]	Hawaii	Statewide	American Council Psychological Exam	1922	Follow-up (1 year)	20	75.9	42.2	1.269
Phearman (1948)	Iowa	Large Sample	Iowa Tests of Educational Development	1929	Follow-up (1 year)	11	67.0	42.9	1.161
Morehead (1950)	Arkansas	Large Sample	American Council Psychological Exam	1931	Follow-up (1 year)	4	62.6	45.0	
Berdie (1954)	Minnesota	Statewide	American Council Psychological Exam	1932	Prospective	20	62.1	41.0	0.698
White (1952)	Northeast Ohio	Sample	Not Specified	1932	Follow-up (1 year)	3	60.4	44.4	
Jones (1956)	Arkansas	Statewide	American Council Psychological Exam	1936	Follow-up (1 year)	19	63.7	44.4	0.693
Cowen (1957)	New York City	Sample	New York State Scholastic Ability Test	1938	Prospective	6	58.8	34.8	
Cowen (1957)	New York (upstate)	Large Sample	New York State Scholastic Ability Test	1938	Prospective	6	61.0	37.0	
Little (1958)	Wisconsin	Statewide	Henmon-Nelson Test of Mental Ability	1939	Follow-up (1 year)	10	67.1	42.6	
Stroup and Andrew (1959)	Arkansas	Large Sample	American Council Psychological Exam	1939	Follow-up (1 year)	3	61.5	45.7	
Nam and Cowhig (1962)	National	Sample	Various[c]	1941	Follow-up (1 year)	4	62.6	40.5	
Flanagan et al. (1964)	National	Sample	Composite Aptitude Test, Project Talent	1942	Follow-up (1 year)	15	65.0	37.3	
Berdie and Hood (1963)	Minnesota	Statewide	Minnesota Scholastic Aptitude Test	1943	Follow-up (1 year)	10	65.2	39.4	
Fetters et al. 1977	National	Sample	Composite Score, NLS72	1954	Follow-up (1 year)	3	61.8	40.0	
Authors' Calculation	National	Sample	AFQT, NLSY79	1960	Follow-up (several yrs.)		62.9	35.1	
Gardner (1987)	National	Sample	Cognitive Test, HS&B	1962	Follow-up (1 year)	4	62.3	38.3	
Donovan and Herrington (2013)	National	Sample	AFQT, NLSY97	1982	Follow-up (several yrs.)		59.1	29.8	

[a] The estimates of $\bar{A}_c$ and $\bar{A}_{nc}$ are from Taubman and Wales (1972), who received detailed information from the study's authors through personal correspondence; this information is not available. However, the details of when and how the study were conducted are available through the listed reference.

[b] Data are from this source; some of the details of how and when the study were conducted are drawn from other papers by the author using the same data (Livesay, 1941a,b). Livesay (1932) and Livesay (1936) provide background that helps explain why this point is such an outlier. The University of Hawaii used test scores to screen applicants from 1922 onward. By the time of the reported cohort, high grades and a letter of recommendation from the school principal were required for admission for low-scoring students. Given Hawaii's isolation from the continental U.S. and the fact that the University of Hawaii was at the time the only university in the state, this policy likely explains the large test score gap.

[c] The authors collected both the score and the test on which it was recorded, and used equivalence tables to map the different testing programs together.

References

Ames, W.R., 1926. Intelligence of High School Seniors in Montana. Ph.D. thesis. University of Wisconsin–Madison.

Barker, R.W., 1937. The Educational and Vocational Careers of High School Graduates Immediately Following Graduation in Relation to Their Scholastic Abilities. Master's thesis. State University of Iowa.

Batson, W.H., 1922. The south dakota group intelligence test for high school. School and Society 15, 311–315.

Benson, V.E., 1942. The intelligence and later scholastic success of sixth-grade pupils. School and Society 55, 163–167.

Berdie, R.F., 1954. After High School – What? University Of Minnesota Press, Minneapolis, Minnesota.

Berdie, R.F., Hood, A.B., 1963. Trends in Post-High School Plans Over and 11-Year Period. Student Counseling Bureau, University of Minnesota, Minneapolis, Minnesota.

Book, W.F., 1922. The Intelligence of High School Seniors as Revealed By a Statewide Mental Survey of Indiana High Schools. Macmillan, New York.

Borghans, L., Golsteyn, B.H., Heckman, J., Humphries, J.E., 2011. Identification problems in personality psychology. Personality and Individual Differences 51, 315–320.

Colvin, S.S., MacPhail, A.H., 1924. Intelligence of Seniors in the High Schools of Massachusetts. 9, Government Printing Office, Washington, DC.

Cowen, P.A., 1957. Needs and Facilities in Higher Education in New York State. The University of the State of New York and the State Education Department, Albany, NY.

Donovan, K., Herrington, C., 2013. Factors affecting college attainment and student ability in the u.s. since 1900. Mimeo, Notre Dame.

Fetters, W.B., Dunteman, G.H., Peng, S.S., 1977. Fulfillment of Short-Term Educational Plans and Continuance in Education. National Longitudinal Study of High School Seniors. Department of Health, Education, and Welfare, Education Division, National Center for Education Statistics, Washington, D.C.

Fishman, J.A., Pasanella, A.K., 1960. College admission-selection studies. Review of Educational Research 30, 298–310.

Flanagan, J.C., Davis, F.B., Dailey, J.T., Shaycoft, M.F., Orr, D.B., Goldberg, I., Neyman Jr, C.A., 1964. The American High-School Student. Project Talent Office, University of Pittsburgh.

Franzen, R., 1922. Attempts at test validation. The Journal of Educational Research 6, 145–158.

Gardner, J.A., 1987. Transition from high school to postsecondary education: Analytical studies. Center for Education Statistics, Office of Educational Research and Improvement, US Department of Education, Washington, D.C.

Garrett, H.F., 1949. A review and interpretation of investigations of factors relatd to scholastic success in colleges of arts and sciences and teachers colleges. The Journal of Experimental Education 18, 91–138.

Goldin, C., Katz, L.F., 2008. The Race Between Education and Technology. The Belknap Press of Harvard University Press.

Henmon, V.A.C., Holt, F.O., 1931. A Report on the Administration of Scholastic Aptitude Tests to 34,000 High School Seniors in Wisconsin in 1929 and 1930: Prepared for the Committee on Cooperation, Wisconsin Secondary Schools and Colleges. 1786, Bureau of Guidance and Records of the University of Wisconsin, Madison, Wisconsin.

Herrnstein, R.J., Murray, C., 1994. The Bell Curve: Intelligence and Class Structure in American Life. Free Press.

Jones, T.M., 1956. Comparisons of Test Scores of High School Graduates of 1954 who Go to College with Those who Do Not Go: And a Study of Certain Factors Associated with Going to College. Ph.D. thesis. University of Arkansas.

Kobrin, J.L., Patterson, B.F., Barbuti, S.M., Mattern, K.D., Shaw, E.J., 2008. Validity of the SAT for predicting first-year college grade point average. Technical Report. College Board Research Report.

Learned, W.S., Wood, B.D., 1938. The Student and His Knowledge. The Carnegie Foundation for the Advancement of Teaching, New York.

Little, J.K., 1958. A Statewide Inquiry Into Decisions of Youth About Education Beyond High School. Technical Report. School of Education, University of Wisconsin.

Livesay, T.M., 1932. A Study of Public Education in Hawaii. University of Hawaii.

Livesay, T.M., 1936. Racial comparisons in performance on the american council pscyhological examination. Journal of Educational Psychology 27, 631–634.

Livesay, T.M., 1941a. Intelligence of high-school seniors in hawaii. Journal of Educational Psychology 32, 377.

Livesay, T.M., 1941b. Test intelligence and future vocation of high school seniors in hawaii. Journal of Applied Psychology 25, 679.

Livesay, T.M., 1942. Test intelligence and college expectation of high school seniors in hawaii. The Journal of Educational Research 35, 334–337.

Mann, G., 1924. Selective influence of desire to attend college. The High School Journal 7, 8–9.

Morehead, C.G., 1950. What's happening to our high-school seniors? Journal of Arkansas Education 23, 12–27.

Morgan, R., 1989. Analysis of the Predictive Validity of the SAT and High School Grades from 1976 to 1985. Technical Report. College Board Report.

Nam, C.B., Cowhig, J.D., 1962. Factors Related to College Attendance of Farm and Nonfarm High School Graduates, 1960. Technical Report. Series Census-ERS P-27 No. 32.

NLS User Services, 1992. Nls79 profiles of american youth. addendum to attachment 106. Http://www.nlsinfo.org/ordering/display_db.php3.

OBrien, F.P., 1928. Mental ability with reference to selection and retention of college students. The Journal of Educational Research 18, 136–143.

Odell, C.W., 1927. Are College Students a Select Group? University of Illinois, Urbana, Illinois.

Phearman, L.T., 1948. Comparisons of High School Graduates Who Go to College with Those Who Do Not Go to College. Ph.D. thesis. State University of Iowa.

Rice, E.A., 1920. A Study of the Correlation Between Scholastic Success and Scores Made on Intelligence Tests. Master's thesis. Indiana University.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C.A., Goeken, R., Hall, P.K., King, M., Ronnander, C., 2009. Integrated public use microdata series: Version 4.0 [machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor]. Available online at `http://usa.ipums.org/usa/`.

Stroup, F., Andrew, D.C., 1959. Barriers to College Attendance. Technical Report. Office of Education, U.S. Department of Health, Education, and Welfare. Magnolia, Arkansas.

Taubman, P., Wales, T., 1972. Mental Ability and Higher Educational Attainment in the Twentieth Century. National Bureau of Economic Research.

White, R.C., 1952. These Will Go to College. Press of Western Reserve University.

Wolfle, D., Smith, J.G., 1956. The occupational value of education for superior high-school graduates. The Journal of Higher Education 27, 201–232.