ACHIEVEMENT GAP ESTIMATES AND DEVIATIONS FROM CARDINAL COMPARABILITY

ERIC R. NIELSEN

THE FEDERAL RESERVE BOARD

Abstract. This paper assesses the sensitivity of standard empirical methods for measuring group differences in achievement to violations in the cardinal comparability of achievement test scores. The paper defines a distance measure over possible weighting functions (scalings) of test scores. It then constructs worst-case bounds for the bias in the estimated achievement gap (or achievement gap change) that could result from using the observed rather than the true test scale, given that the true and observed scales are no more than a certain distance from each other. The paper next estimates these worst-case weighting functions for black/white and high-/low-income achievement gaps and gap changes using several commonly employed surveys. The results of this empirical exercise suggest that cross-sectional achievement gap estimates tend to be quite robust to scale misspecification. In contrast, achievement gap change estimates seem to be quite sensitive to the choice of test scale. The paper next extends the bounding methodology to study bias in regression coefficients when the left-hand side variable is incorrectly scaled. The same survey data suggest that regression coefficients relating income to achievement in the cross-section are quite robust to scale-misspecification, while first differences in regression coefficients appear to be much more fragile. Standard empirical methods do not robustly identify the sign of the trend in achievement inequality between students from different racial groups and income classes. JEL Codes: C18, I24, I26

1. Introduction

Researchers frequently use test-score data to assess group differences in achievement. The vast majority of such investigations assume that some known normalization renders test scores cardinally comparable in the sense that a given score change has the same meaning throughout the range of possible scores. Furthermore, such investigations typically assume that a given test score has the same meaning across different surveys, student ages, or time periods.[1] Neither of these comparability assumptions are well motivated by either economic or psychometric theory. If either fails, standard estimates of achievement gaps and achievement gap changes

Date: October, 2015.

Seminar participants at the Copenhagen Education Network, the DC Education Working Group, and the Federal Reserve contributed valuable feedback and suggestions for this project. Rick Ogden provided excellent research assistance. The views and opinions expressed in this paper are solely those of the author and do not reflect those of the Board of Governors or the Federal Reserve System. Contact: Division of Research and Statistics, Board of Governors of the Federal Reserve System, Mail Stop 97, 20th and C Street NW, Washington, D.C. 20551. eric.r.nielsen@frb.gov. (202) 872-7591.

[1] Consider SAT scores. If SAT scores are comparable over time, a student who earned a 600 on the math section in 1980 should have the same achievement as a student who earned a 600 in 2010. If the SAT has a cardinal (interval) scale, then a student who improves her math score from 400 to 500 has improved by the same amount as a student whose score increased from 600 to 700.

("gaps/changes") may be severely biased. Such estimates are no longer even guaranteed to correctly identify the sign of the achievement gap/change.

In a parallel working paper I show how to make achievement comparisons using only the ordinal content of test scores. That paper shows that the cardinal/ordinal distinction has real importance; standard cardinal methods suggest that the achievement gap between youth from high- and low-income households widened or changed ambiguously in recent decades, whereas ordinal methods indicate just the opposite.[2]

The two necessary conditions for ordinal statistics to unambiguously identify achievement differences are quite demanding. First, it must be possible to place test scores on a common scale so that a given score corresponds to the same underlying level of achievement regardless of the year, cohort, or age group from which the score was drawn.[3] Second, various first-order stochastic dominance conditions must hold between the relevant test-score distributions.[4] These conditions will not be met for many economically interesting achievement comparisons.

The stringency of the necessary conditions for valid ordinal inference means that many interesting achievement comparisons are inherently scale dependent. In these situations we really cannot determine with certainty the sign of an achievement gap/change without leaning more or less heavily on some particular cardinalization of achievement. Since test scores are unlikely to be valid cardinal measures, should researchers simply plead ignorance when ordinal estimates are inconclusive?

There are good reasons to resist such radical agnosticism. Test scales may not be perfectly cardinal, yet they may still carry useful cardinal information. For example, suppose we are comparing three students with SAT scores of 1000, 1500, and 1510. It seems plausible that the student with a 1500 is closer to the 1510 student than she is to the 1000 student, even if the true differences are not exactly proportional to 10 and 500. Eschewing cardinality completely may throw away a lot of useful information, unnecessarily decreasing one's power

[2] In particular, ordinal analysis of data on student achievement in 1980 1997 strongly suggests a decrease in the income-achievement gap, while cardinal methods applied to the same data suggest a flat or increasing gap. Data comparing cohorts from 1990 and 2002 yield an ambiguous ordinal gap change and an ambiguous or increasing cardinal gap change.

[3] Many standardized tests are renormed every year, violating the common-scale assumption. I abstract from this problem in the theory sections of this paper. In my empirical work I take great care to use scores that allow one to rank students from different surveys against each other consistently.

[4] In particular, the "high" group score distribution must first-order dominate the "low" group score distribution within a given year/cohort for the sign of the cross-sectional achievement gap to be unambiguous. For an achievement gap change to be unambiguous, the high group in the earlier period must first-order dominate the high group in the later period, and the low group in the later period must first-order dominate the low group in the earlier period.

to detect achievement differences. Intuitively, if a known test scale is "almost" cardinal, cardinal statistical tests may correctly identify the sign of an achievement gap/change and have greater power than ordinal tests. In contrast, if the test scale used is actually very far from the true scale, then cardinal methods may misidentify achievement gaps/changes and ordinal methods should be used instead.

In order to operationalize this intuitive tradeoff, I define in this paper a distance measure that quantifies how far apart are two candidate test scales. Next, I suppose that nothing is known about the true test scale other than that it lies within a fixed distance of the observed scale. I then search for the unobserved true scale satisfying the hypothesized distance restriction that maximizes the difference between the observed and true achievement gap/change. By studying the worst-case bias as a function of the hypothesized distance between the true and observed scales, I can assess the sensitivity of standard methods to scale misspecification.

I derive closed-form expressions for the test scales that maximize and minimize the true gap/change relative to the observed gap/change. The worst-case weighting functions are all piecewise-linear, with flat regions (where changes in observed test scores are uninformative) and cardinal regions (where changes in observed test scores map linearly to changes in true achievement). Furthermore, the weighting functions often feature discontinuous jumps where a small change in the observed test score corresponds to a large change in true achievement.

I estimate the worst-case weighting functions and resulting biases for black/white and high/low-income achievement gaps/changes in the National Longitudinal Surveys of Youth (NLSY) 1979 and 1997 and the National Education Longitudinal Surveys (NELS/ELS) 1990 and 2002. The cross-sectional achievement gap estimates are quite robust in these data. It is often not possible to find a rescaling that flips the sign of a given gap estimate, no matter the distance restriction. In other cases, the minimum distance needed for the observed scale to misidentify the sign of the true gap is very large. For instance, to flip the sign of the black/white reading achievement gap in the NLSY97, the weights placed on test scores by the true and observed scales must differ by at least 2 standard-deviation units somewhere on the range of the observed scores. In contrast, gap-change estimates are typically much more sensitive to scale deviations. The sign of every gap-change I analyze can be flipped given a sufficiently large distance restriction. Furthermore, the minimum distances required to affect a sign flip are often quite small. For example, if the true and observed scale are allowed to differ by only 0.15 standard deviations somewhere on their (normalized) support, the sign of the income-achievement gap change for reading may be misidentified in the NELS/ELS data.

Test scores are also often used as outcome variables in regression models, either to estimate school/teacher value-added effects or to assess how strongly achievement is related to some socioeconomic or demographic variable of interest. Regression-based methods also assume that test scores are cardinal measures of achievement and therefore may produce biased estimates if the scale of achievement has been incorrectly specified. The bounding methodology developed for mean differences can be modified easily to study bias in ordinary least squares (OLS) or instrumental variables (IV) regression coefficients when the left-hand side variable is incorrectly scaled. Empirically, I estimate the robustness of OLS regression coefficients relating household income to student achievement in the NLSY data. I find that the regression coefficients describing the cross-sectional relationships between income and achievement are uniformly very robust to scale misspecification. It is never possible to flip the sign of these coefficients by rescaling the test scores. However, estimates of the change in the association between income and achievement from the NLSY79 to the NLSY97 are not robust. It is always possible to reverse the estimated trend in this association, and sometimes only minor changes to the observed test scale are sufficient to affect such a reversal.

Although the main applications studied in this paper are quite specific, the techniques introduced here can be easily adapted to study robustness in a number of other empirical applications. The methodology can be applied to any situation in which either mean differences or regressions are used on a variable that does not have a clearly-defined, cardinally-interpretable scale. Other potential applications include measuring group differences in self-reported happiness, group differences in non-cognitive skills, and group differences in poverty rates assessed using deprivation indicators. In order to keep the length and scope of this paper manageable, such empirical extensions are left for future work.

My empirical results cast serious doubt on research that uses cardinal methods to measure time trends in achievement inequality. Some of the most well-studied achievement trends estimated with very widely used data are not robust to minor rescalings of test scores. Since there are not good reasons to prefer the observed test scale to other, similar scales, such estimates are not credible. Researchers assessing changes in achievement inequality over time should be much more circumspect in their deployment of standard cardinal methods and should use ordinal methods where possible.

This paper is not entirely negative, as I develop tools that allow researchers to assess whether a particular cardinal estimate is sensitive to the choice of scale. These tools are straightforward to apply and do not require more data than would be used in standard empirical calculations. If standard methods turn out to be robust in a particular setting, then there is no need to proceed to less familiar and less powerful ordinal approaches.

Ultimately, whether a test scale should be used cardinally depends on the judgment of the researcher. A given test scale may be cardinal for some applications and not cardinal for others. For instance, a test score which gives the percentage of an alphabet that a student knows is cardinal by definition if the outcome of interest is the percentage of the alphabet known. If the outcome of interest, however, is adult earnings capacity, literacy, or virtually any long-run outcome, then such a test scale is likely not cardinal even if the scores strongly predict outcomes. The bounding methods developed in this paper allow one to effectively parametrize one's uncertainty about the cardinality of a test scale in a given application and assess the robustness of standard approaches as a function of that uncertainty.

The rest of the paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 lays out the notation, defines the necessary mathematical objects, and justifies the normalizations and simplifications employed. Section 4 derives the worst-case weighting functions for a general class of achievement gap/change estimates and shows how to extend these results to linear OLS and IV regression. Section 5 assesses the sensitivity of a number of achievement gap/change estimates to cardinal deviations using the NLSY and NELS/ELS data. Section 6 investigates the sensitivity of regression coefficients to scale misspecification in the NLSY data. Section 7 discusses estimation error, inference, and measurement error. Section 8 concludes. Appendices A through E contain figures, estimates, proofs, and additional discussion.

2. Literature Review

The economics literature using cardinal methods to assess group differences in achievement is vast. Fryer and Levitt[9, 10], Clotfelter, Ladd, and Vigdor[6], Duncan and Magnuson[8], Hanushek and Rivkin[11], and Neal[18], among many others, use cardinal methods to assess changes in black/white achievement inequality in the United States.[5] Reardon[23] employs cardinal methods to argue that the gap in achievement between high- and low-income youth has widened tremendously over the past several decades. Research assessing school and teacher

[5] Neal[18] does recognize, however, that "[a]chievement has not natural units," and so he also analyzes the percentile rankings of black versus white test takers.

performance through value-added models (VAMs) and papers estimating the productivity of various inputs such as class size and teacher quality on student achievement also typically assume that test scores are cardinal measures.[6]

This paper is not the first to argue that normalized test scores are not cardinal measures of achievement. In psychometrics, Stevens[27] and Lord[17] argue that most psychometric test scores are inherently ordinal. In economics, Cunha and Heckman[7], along with many others, argue for "anchoring" test scores on interpretable life outcomes to avoid using test scores cardinally. Lang[16], Bond and Lang[4], Cascio and Staiger[5], Reardon[22], and Nielsen[20] all discuss the sensitivity of standard achievement gap/change estimates to order-preserving transformations of test scores. The analysis in Bond and Lang[4] is particularly relevant to this paper. These authors search over a fairly general class of order-preserving transformations of test scores in order to find rescalings that maximize and minimize the apparent change in black/white achievement inequality through the first several years of school. In spirit, my paper is also quite similar to a working paper (currently not posted) from Schroeder and Yitzhaki[24] that investigates whether sign reversals are possible in regressions in which the outcome variable is self-reported life satisfaction. In addition to considering a different empirical application, their paper differs from mine in that it focuses primarily on regressions and does not attempt to construct bounds on how badly misspecified a scale must be in order to generate a sign reversal.

Economists and policymakers are usually not really interested in the test scores themselves, but rather are interested in the (social) value of the achievement represented by the test scores. This formulation yields an isomorphism between measuring achievement gaps and using social welfare functions to rank income distributions. In this context, Atkinson[3] shows that first-order stochastic dominance (FOSD) is both necessary and sufficient for all increasing social welfare functions to agree on the ranking of two distributions, while all concave functions will rank identically under second-order dominance. Aaberge, Havnes, and Mogstad[1] extend the analysis to consider ranking distributions under dominance of any order. Applying these results to test scores requires imposing conditions on the social welfare function that are less plausible for test scores than for income. For example, the relationship between test scores and life outcomes may be quite convex, so that social welfare may not be convex in test scores

[6] For example, Krueger[15] and Hoxby[13] both use test scores cardinally to estimate the effect of class size on student achievement. Value-added methodologies such as those expounded in Raudenbush[21] and elsewhere likewise suppose that (normalized) test scores are cardinally comparable.

even if it is concave in life outcomes.[7] Furthermore, if academic achievement is itself the object of interest, there is no particular reason to impose that the true scale is a concave function of the observed scale.

3. Notation, Assumptions, Definitions

This section introduces the notation and assumptions I maintain throughout the paper. I introduce assumptions and concepts specific to a particular application as needed.

Consider a population of students with real-valued test scores s distributed according to cumulative density function (cdf) F. Let $W_0(s)$ be the true value of the underlying achievement corresponding to test score s. My preferred framing conceives of W_0 as the composition of several conceptually distinct maps: the map from test scores to true achievement, the map from true achievement to life outcomes, and the map from life outcomes to welfare. Even assuming that the choice of the welfare function is uncontroversial, the first of these maps is not knowable and the second is very difficult to estimate even with very rich data.[8] An alternative framing simply considers W_0 as the true, unknown scale of achievement that may be distinct from the observed test scale. The fundamental idea is that W_0 represents the scaling that would render cardinal methods valid in a particular application. I assume that W_0 is totally inaccessible to the researcher.

The only *a priori* restriction I place on W_0 is that it be weakly increasing in s.[9] Weak monotonicity is a natural assumption in this setting because higher test scores should correspond to weakly higher underlying achievement, and life outcomes (income, marriage, etc.) should be causally linked to achievement. I do not assume that W_0 is strictly monotone because I want to allow for the possibility that changes in test scores in some regions do not effect overall welfare, either because the scores themselves are uninformative or because higher achievement does not always lead to better outcomes. Even if the map from test scores to achievement is

[7] For example, consider a test of athletic ability and suppose that we are interested in lifetime labor income. Reasonable preferences on income will likely be concave, but the relationship between athletic ability and income may be highly convex. The increase in income associated with moving from the level of a good college basketball player to the level of LeBron James is so large that it may well swamp any concavity in social welfare.

[8] Life outcomes such as longevity, health, total labor market earnings, marriage quality, and so forth are only fully revealed decades after most achievement test scores are recorded. Estimating even some of these outcomes with the best longitudinal data available is a major econometric challenge. Nielsen[20] carries out such a calculation for lifetime earnings in the National Longitudinal Surveys of Youth (NLSY) data.

[9] That is, $s > s' \implies W_0(s) \geq W_0(s') \; \wedge W(s) > W_0(s') \implies s > s'$. Note that this formulation supposes that test scores are perfectly reliable. A full treatment of test-score measurement error is beyond the scope of this paper. Nonetheless, I argue in section 7 that classical measurement error will tend to make mean-based achievement gap/change estimates appear more robust to scale misspecification than they in fact are.

strictly monotone, either or both of the maps from achievement to life outcomes or from life outcomes to social welfare may have flat regions. Weak monotonicity does not rule out the possibility that W_0 is constant everywhere. However, the worst-case W_0's I derive in subsequent sections will be strictly increasing somewhere in all but the most extreme cases. Unless I explicitly specify otherwise, I will therefore treat generic W_0's in the remaining analysis as having at least two values $s > \tilde{s}$ such that $W_0(s) > W_0(\tilde{s})$.

I make a number of assumptions and normalizations on the observed test-score distributions and true score weighting functions in order to simplify the analysis. These assumptions do not rule out any economically interesting cases.

Definition 3.1. F satisfies (A1) iff:

(i) $F \in \mathcal{F}$, the set of univariate distributions with continuous densities everywhere on their support. Let f denote the probability density function (pdf) associated with F.

(ii) $Support(F) = [0, 1]$

Part (i) of definition 3.1 is convenient for technical reasons. Part (ii) is just a normalization and is without loss of generality because test scores can always be rescaled to fit in [0,1] from whatever cardinal scale the researcher prefers.[10]

Definition 3.2. W_0 satisfies (A2) iff:

(i) W_0 is weakly increasing and right-continuous in s.

(ii) $W_0(s) \in [0, 1]$ for all $s \in Support(F)$.

The assumption that W_0 is weakly increasing was justified previously. The requirement that W_0 be right-continuous in part (i) of definition 3.2 is a technical assumption that guarantees uniqueness of the worst-case weighting functions.[11] Part (ii) of definition 3.2 normalizes $W_0(s)$ to have the same support as F. One can change the units of W_0 without changing anything in the analysis except for the units of the distance restriction and the resulting biases. Figure 1 in appendix A plots several possible W_0's satisfying (A2). The figure shows that convex functions, concave functions, and discontinuous functions can all satisfy (A2).

[10] Suppose a researcher has a candidate cardinal scale such that test scores follow distribution $\tilde{F}$ with $Support(\tilde{F}) = (a, b) \subset (-\infty, \infty)$. Since a and b are finite, an affine transformation will rescale test scores to [0,1] while preserving the purported cardinality of $\tilde{F}$.

[11] In particular, the worst-case W_0's will often have discontinuous jumps somewhere on $Support(F)$. Right-continuity rules out the existence of multiple W_0's that differ only on these (measure-0) regions.

In order to assess how sensitive a given cardinal achievement statistic is to scale misspecification, I must first define a distance measure on test scales. I define the distance between two candidate test scales using the sup norm.

Definition 3.3. Let W and $\tilde{W}$ be test-score weighting functions with support on [0,1]. The distance between W and $\tilde{W}$ is

$$D(W, \tilde{W}) \equiv \sup_{s \in [0,1]} |W(s) - \tilde{W}(s)|.$$

The sup norm gives an intuitive way to assess the degree to which two weighting functions disagree. If $D(W, \tilde{W})$ is very small, then at no point do the weighting functions differ by very much. In contrast, when $D(W, \tilde{W})$ is large, there is at least one test score that the two scales value very differently. The sup norm is not the only way to formalize the notion of distance between weighting functions, but the analysis using alternative distance measures is much less tractable mathematically.[12]

Section 4 shows that the bias created by incorrectly scaled test scores depends on expressions of the form $\int (W_0(s) - s)G(s)ds$. The function G, which depends on the particular bounding application, determines both the magnitude of the bias and the functional form of the worst-case test-score weighting functions. As theorems 4.2 and 4.6 demonstrate, these worst-case weights are generally parametrized by the number and location of the crossing-point zeros of G. Assumption (A3) therefore characterizes functions G according to these zeros.

Definition 3.4. G satisfies (A3) for $N \in \mathbb{N}$ if the following conditions hold:

(i) G is continuous with support on [0,1].

(ii)$\exists s_1^*, s_2^*, \ldots, s_N^*$ with $0 < s_1^* < s_2^* < \ldots < 1$ such that $G(s_i^*) = 0 \; \forall i \in \{1, \ldots, N\}$.

(iii) For each s_i^* there exist $\varepsilon_{i,l} > 0$ and $\varepsilon_{i,h} > 0$ such that $\text{sign}[G(\underline{s}_i - \delta_{i,l})] = -\text{sign}[G(s_i^* + \delta_{i,h})]$ for all $\delta_{i,l} \in (0, \varepsilon_{i,l.})$ and $\delta_{i,h} \in (0, \varepsilon_{i,h})$ where $\underline{s}_i = \sup\{s | s \leq s_i^* \wedge G(s) \neq 0\}$.

(iv) For some $\varepsilon_1 > 0$, $G(s_1^* - \delta_1) \leq 0$ for all $\delta_1 \in (0, \varepsilon_1)$.

(v) If $G(\bar{s}) = 0$ for some $\bar{s} \notin \{s_i^*\}$, then $\bar{s}$ does not satisfy condition (iii).

Assumption (A3) defines a very general class of functions with support on [0,1]. Parts (ii) and (iii) assert that G crosses 0 exactly N times on (0,1), although they allow for the possibility

[12] For instance, one could define $\mathcal{D}(W, \tilde{W}) \equiv \int |W(x) - \tilde{W}(x)|dx$. This alternative definition will assess a large difference in the case that W and $\tilde{W}$ differ by a small amount everywhere on [0,1].

that G is identically equal to 0 for some subintervals of [0,1]. Part (iv) of the definition says that G is weakly negative before the first such crossing point. This assumption is without loss of generality in the applications studied here because it will always be possible to pick a reference group of students such that it holds.[13] Figure 2 in appendix A displays several functions consistent with (A3) for various values of N.

4. Bounding Analysis

This section presents theoretical bounding results for a number of empirically relevant applications. Section 4.1 characterizes the worst-case bounds for mean gap and gap-change estimates. Sections 4.2 analyzes the bias in regression coefficients due to scale miss-specification when the right-hand side variable is continuous. Appendix D extends the analysis to binary regression and mean differences with multiple dimensions of achievement.

4.1. **Mean Gaps/Changes.** Consider measuring the cross-sectional achievement gap between two groups of students, A and B. Letting $F_{A,t}$ and $F_{B,t}$ denote their test-score distributions in period t, the true cross-sectional achievement gap between them is given by

$$\Delta V(W_0, A, B, t) \equiv \mathbb{E}_{F_{A,t}}[W_0] - \mathbb{E}_{F_{B,t}}[W_0] = \int_0^1 W_0(s) \underbrace{[f_{A,t}(s) - f_{B,t}(s)]}_{\equiv \Delta f_t(s)} ds.$$

Analogously, the change in the cross-sectional achievement gap from period t to $t+1$ is[14]

$$\Delta V(W_0, A, B, t, t{+}1) \equiv \Delta V(W_0, A, B, t{+}1) - \Delta V(W_0, A, B, t) = \int_0^1 W_0(s) \underbrace{[\Delta f_{t+1}(s) - \Delta f_t(s)]}_{\equiv \Delta f_{t+1,t}(s)} ds.$$

In both of these cases, the object of interest is an integral of the form $\int_0^1 W_0(s)\Delta f(s)$, where Δf is some sum and difference of the relevant density functions. The specific application (cross-sectional or gap-change) matters only insofar as it alters Δf. Therefore, I will characterize bias in expressions with the general form $\Delta V(W_0, \Delta f) \equiv \int_0^1 W_0(s)\Delta f(s)ds$, while leaving the specific objective in the background.

[13] For example, if (iv) is not satisfied when low-income students are the reference group and high-income students are the comparison group, switching these two group's roles in the analysis will guarantee that (iv) holds.

[14] I will exclusively use language describing gap-changes occurring over time. However, nothing in the analysis requires time to be the dimension along which change is assessed. For instance, one could replace "t" with "urban school district" and "$t+1$" with "suburban school district," and nothing about the mathematics would change.

Suppose that $\mathbb{I}(s) = s$ were used to calculate ΔV instead of W_0. The "pseudo-gap" as measured by $\mathbb{I}$ would then be $\Delta V(\mathbb{I}, \Delta f) = \int_0^1 s\Delta f(s)ds$. The bias created from using $\mathbb{I}$ instead of W_0 is just the difference between the true gap and the pseudo gap. There are two cases to consider for bounding this bias: weights that maximizes the degree to which the true difference is larger than the observed difference and weights that maximizes the degree to which the true difference is smaller than the observed difference. Define

$$\mathcal{B}^+(\mathbb{I}, W_0, \Delta f) \equiv \int_0^1 (W_0(s) - s)\,\Delta f(s)ds, \quad \mathcal{B}^-(\mathbb{I}, W_0, \Delta f) \equiv \int_0^1 (s - W_0(s))\,\Delta f(s)ds.$$

The worst-case W_0's for a given k are just those weighting functions that maximize $\mathcal{B}^+$ and $\mathcal{B}^-$ among all weighting functions that satisfy $D(W, \mathbb{I}) \leq k$.

Definition 4.1. The worst-case W_0's satisfying (A2) and $D(\mathbb{I}, W) \leq k$ for a given distance restriction k are defined by

$$\begin{aligned} W_0^+(s|k, \Delta f) &\equiv \max_{W \in \mathcal{W} \wedge D(\mathbb{I}, W) \leq k} \mathcal{B}^+(\mathbb{I}, W, \Delta f) \\ W_0^-(s|k, \Delta f) &\equiv \max_{W \in \mathcal{W} \wedge D(\mathbb{I}, W) \leq k} \mathcal{B}^-(\mathbb{I}, W, \Delta f). \end{aligned}$$

Let $\bar{\mathcal{B}}^+(k) \equiv \mathcal{B}^+(\mathbb{I}, W_0^+(s|k, \Delta f), \Delta f)$ and $\bar{\mathcal{B}}^-(k) \equiv \mathcal{B}^-(\mathbb{I}, W_0^-(s|k, \Delta f), \Delta f)$ denote the worst-case biases given k. Similarly, let $\Delta V(W_0^+(k)) \equiv \Delta V(\mathbb{I}) + \bar{\mathcal{B}}^+(k)$ and $\Delta V(W_0^-(k)) \equiv \Delta V(\mathbb{I}) - \bar{\mathcal{B}}^-(k)$ be the worst-case true gaps/changes as a function of k.

I now construct closed-form expressions for W_0^+ and W_0^- when Δf satisfies (A3) for some N. Both W_0^+ and W_0^- have relatively simple functional forms under (A3) for any value of $k \in [0, 1]$. In contrast, it will not generally be possible to find closed-form expressions for $\bar{\mathcal{B}}^+(k)$ and $\bar{\mathcal{B}}^-(k)$. Nonetheless, knowing the forms of W_0^+ and W_0^- makes approximating $\bar{\mathcal{B}}^+(k)$ and $\bar{\mathcal{B}}^-(k)$ fairly straightforward in most empirical applications.

The functional forms of W_0^+ and W_0^- depend on whether N is even or odd, as both W_0^+ and W_0^- are parametrized by the values they take at the various crossing points of Δf. In particular, W_0^+ is parametrized by its values at even-indexed crossing points (s_i^* such that i is even), while W_0^- depends on its values at odd-indexed crossing points. Theorem 4.2 below characterizes W_0^+ and W_0^- for arbitrary N. Figures 5 and 6 in appendix A plot possible W_0^+ and W_0^- functions when $N = 2$ or $N = 3$.

Theorem 4.2. *If (A1)-(A3) hold for $N \in \mathbb{N}$, then there exist non-decreasing sequences $0 \leq s_2^+ \leq s_4^+ \leq \ldots \leq 1$ and $0 \leq s_1^- \leq s_3^- \leq \cdots \leq 1$ such that $W_0^+(s_i^*|k) = s_i^+ \in [\max\{s_i^* - k, 0\}, \min\{s_i^* + k, 1\}]$ for even $i \leq N$, $W_0^-(s_i^*|k) = s_i^- \in [\max\{s_i^* - k, 0\}, \min\{s_i^* + k, 1\}]$ for odd $i \leq N$, and such that*

$$W_0^+(s|k) = \begin{cases} \max\{0, s-k\}, & s \leq s_1^* \\ \min\{s+k, s_2^+\}, & s \in (s_1^*, s_2^*] \\ \max\{s-k, s_2^+\}, & s \in (s_2^*, s_3^*] \\ \vdots & \\ \max\{s-k, s_N^+\}, & s \in (s_N^*, 1], \ N \text{ even} \\ \min\{s+k, 1\}, & s \in (s_N^*, 1], \ N \text{ odd.} \end{cases} \qquad W_0^-(s|k) = \begin{cases} \min\{s+k, s_1^-\}, & s \leq s_1^* \\ \max\{s-k, s_1^-\}, & s \in (s_1^*, s_2^*] \\ \min\{s+k, s_3^-\}, & s \in (s_2^*, s_3^*] \\ \vdots & \\ \min\{s+k, 1\}, & s \in (s_N^*, 1], \ N \text{ even} \\ \max\{s-k, s_N^-\}, & s \in (s_N^*, 1], \ N \text{ odd.} \end{cases} \tag{4.1}$$

Proof. In appendix C. □

Theorem 4.2 is somewhat difficult to parse. Therefore, I will discuss the special case $N = 1$ at length, as W_0^+ and W_0^- have particularly simple and intuitive expressions when Δf crosses 0 only once. The forces determining W_0^+ and W_0^- when Δf has more than one zero are identical, but the mathematical expressions are more complicated.

Corollary 4.3. *If (A1)-(A3) hold for $N = 1$, then for some $s_c \in [\max\{s_1^* - k, 0\}, \min\{s_1^* + k, 1\}]$, W_0^- and W_0^+ are given by*[15]

$$W_0^+(s|k) = \begin{cases} \max\{s-k, 0\}, & s \in [0, s_1^*) \\ \min\{s+k, 1\}, & s \in [s_1^*, 1] \end{cases} \quad W_0^-(s|k) \begin{cases} \min\{s_c, s+k\}, & s \in [0, s_1^*) \\ \max\{s_c, s-k\}, & s \in [s_1^*, 1]. \end{cases} \tag{4.2}$$

To understand the expression for W_0^+ in equation (4.2), recall that $\mathcal{B}^+$ is large when $[W_0(s) - s]$ and $\Delta f(s)$ have the same sign. This implies that $\mathcal{B}^+$ will be maximized when W_0^+ is as far as possible below the 45 degree line for values of s less than s_1^* and as far above the diagonal when s is greater than s_1^*. The farthest possible value below s consistent with $D(\mathbb{I}, W_0^+)$ is just $\max\{s-k, 0\}$, which is the expression for W_0^+ on $[0, s_1^*)$, while the farthest possible value above is $\min\{s+k, 1\}$, which defines W_0^+ on $[s_1^*, 1]$. Figure 3 in appendix A plots one such W_0^+.

[15] I will always include s_1^* in the "upper half" of W_0^+ or W_0^-. This choice is arbitrary and unimportant since s_1^* has 0 measure.

The analysis for W_0^- when $N = 1$ is more involved. The complicating factor is that $\mathcal{B}^-$ is large when $[W_0(s) - s]$ and Δf have opposite signs. Therefore, W_0^- would "like" to be as far above the diagonal as possible on $[0, s_1^*)$ and as far below the diagonal as possible on $[s_1^*, 1]$. But W_0^- must be weakly increasing, so the larger $s_c = W_0^-(s_1^*)$ is, the less bias can be created on $[s_1^*, 1]$, and the smaller $s_c = W_0^-(s_1^*)$ is, the less bias can be created on $[0, s_1^*)$. The functional form of W_0^- is straightforward to derive given s_c, and each potential choice of s_c trades off bias creation below and above s_1^* differently. Since (A1)-(A3) imply that this tradeoff is a smooth function of s_c, there must be some value of s_c in the interval $[s_1^* - k, s_1^* + k]$ that maximizes the overall bias. Figure 4 in appendix A plots W_0^- for three different values of s_c.

Both W_0^+ and W_0^- have an intuitive interpretation for cross-sectional achievement gaps in the case that $F_A \succ F_B$. FOSD implies that any weighting scheme will measure a positive gap between A and B. Since the scores in A dominate those in B, type-B students have relatively greater density among scores close to 0 and relatively lower density among scores close to 1 so that Δf satisfies (A3) for $N = 1$. The true gap between A and B will therefore be very large if scores close to 0 are given as little weight as possible while scores close to 1 are weighted quite heavily, which is exactly what W_0^+ does. Symmetrically, the true gap between them will be as small as possible exactly when low scores are given as much as weight as possible relative to high scores, which is just what W_0^- does.

Consider now the case that Δf has more than one interior crossing point ($N > 1$). This modification substantially complicates the determination of W_0^+ and W_0^-, although closed-form expressions still exist for both weighting functions. The source of the complication is again the tension between setting W_0^+ or W_0^- as low (or high) as possible over an interval $[s_i^*, s_{i+1}^*]$ and setting it as high (or low) as possible on $[s_{i+1}^*, s_{i+2}^*]$. The intuition for the exact forms of these worst-case weighting functions is exactly the same as the intuition behind the determination of s_c for W_0^- in the case with only one crossing point. The functional forms of W_0^+ and W_0^- are pinned down by the values they take at the interior cross points of Δf. One then need only search over the range of feasible values for these crossing points to find the true worst-case test scales. Figures 5 and 6 in appendix A plot various potential W_0^- and W_0^+ functions for $N = 2$ and $N = 3$ crossing points.

The robustness of a cardinal gap/change estimate to deviations in scale depends on how rapidly the associated biases $\bar{\mathcal{B}}^+$ and $\bar{\mathcal{B}}^-$ increase as k increases. If these biases increase rapidly with k, then a relatively small k may be sufficient to flip the sign of the gap/change estimate. In contrast, if they increase slowly, a reversal will only be possible when k is quite large. In general, it is not possible to derive closed-form expressions for the derivatives of $\bar{\mathcal{B}}^+$ and $\bar{\mathcal{B}}^-$ with respect to k because these derivatives depend on the particular shape of Δf. Nonetheless, it is still possible to gain some intuition about what features of Δf determine how quickly $\bar{\mathcal{B}}^+$ and $\bar{\mathcal{B}}^-$ increase with increases in k. I only present the analysis for the case that Δf crosses 0 once; the results are qualitatively similar with more crossing points, but the expressions are messier and less intuitive.

Theorem 4.4. *If (A1)-(A3) hold for $N = 1$ and k is sufficiently close to 0, then*[16]

$$\begin{aligned}
\frac{\partial \bar{\mathcal{B}}^+}{\partial k} &= \int_k^{1-k} |\Delta f(s)| ds \\
\frac{\partial \bar{\mathcal{B}}^-}{\partial k} &= \int_{s_c+k}^{1} \Delta f(s) ds - \int_0^{s_c-k} \Delta f(s) ds - \int_{s_c-k}^{s_c+k} \frac{\partial s_c}{\partial k} \Delta f(s) ds.
\end{aligned}$$

Proof. In appendix C. □

Theorem 4.4 says that $\bar{\beta}(k)^+$ will increase rapidly with k when there is a lot of area between Δf and 0 on the central interval $[k, 1-k]$. Furthermore, $\frac{\partial \bar{\mathcal{B}}^+}{\partial k}$ is monotonically decreasing in k and approaches 0 from above as k approaches 0.5. The expression for $\frac{\partial \bar{\mathcal{B}}^-}{\partial k}$ is somewhat harder to interpret because s_c is only defined implicitly. For simplicity, suppose that Δf is symmetric in the sense that $\Delta f(0.5 - x) = -\Delta f(0.5 + x)$ for any $x \in [0, 0.5]$. It is immediate in this case that s_c is equal to 0.5 for all values of k, which implies that $\frac{\partial \bar{\mathcal{B}}^-}{\partial k}$ is larger the larger is the area between 0 and Δf on the tail intervals $[0, 0.5 - k]$ and $[0.5 + k, 1]$. Theorem 4.4 also implies that $\frac{\partial \bar{\mathcal{B}}^+}{\partial k}|_{k=0} = \frac{\partial \bar{\mathcal{B}}^-}{\partial k}|_{k=0} = \int_0^1 |\Delta f(s)| ds$. For values of k very close to 0, $\bar{\mathcal{B}}^+$ and $\bar{\mathcal{B}}^-$ increase symmetrically with k. As k grows larger, the relevant subintervals of [0,1] contributing the most to $\bar{\mathcal{B}}^+$ and $\bar{\mathcal{B}}^-$ become more and more distinct. This divergence, coupled with possible increases or decreases in s_c as k grows larger, means that $\frac{\partial \bar{\mathcal{B}}^+}{\partial k}$ and $\frac{\partial \bar{\mathcal{B}}^-}{\partial k}$ will not generally be equal when k is strictly greater than 0.

Theorem 4.2 shows that both W_0^+ and W_0^- generically consist of regions where increases in scores are not valuable, regions where the true value increases one to one with observed test scores, and discontinuous achievement thresholds where the true value jumps up between adjacent test scores. Although these bias-maximizing scales may look extreme, they are not

[16] In particular, the expression for $\frac{\partial \mathcal{B}^+}{\partial k}$ assumes that $k < \min\{s^*, 1 - s^*\}$, while the expression for $\frac{\partial \mathcal{B}^-}{\partial k}$ requires that $k < \min\{s_c, 1 - s_c\}$.

economically implausible. For example, consider a test score equal to the share of the Russian Cyrillic alphabet that a student knows. This scale is interval in the sense that each score increment of $\frac{1}{33}$ corresponds to a new, identifiable skill: knowing an additional letter of the alphabet. However, if Russian literacy is the ultimate objective, a plausible *economic* weighting of these scores should be mostly flat for scores between 0 and $\frac{32}{33}$ and display a sizable increase between $\frac{32}{33}$ and 1 because knowing the entire alphabet is a prerequisite for reading and writing in the Russian language. Similarly, a job may require a constellation of skills such that the productivity of a worker lacking any one of the skills is close to 0 while the productivity of a worker possessing all of the requisite skills is quite high.[17] Finally, selective institutions may employ admissions thresholds, again creating discontinuities and kinks in the economically-relevant score weighting function. In short, it is not hard to find realistic scenarios where the relevant test scale may not be a smooth function of the observed scores.

4.2. **Regression Bias with a Continuous Covariate.** Suppose we are interested in using linear regression to understand the relationship between some continuous variable x and achievement.[18] For example, one might regress test scores on household income, school expenditures, or parental education. Using test scores on the left-hand side of a regression assumes that the scores are cardinal measures of achievement. This section investigates how robust such regression-based methods are to scale misspecification.

As in previous sections, I assume that the true, cardinal scale of achievement is unobservable. I use Y_0 throughout this section to denote the true scale in order to limit potential confusion between the weights used for bounding regression coefficients and those used for bounding mean differences. I carry through all of the assumptions and definitions from section 4.1. What remains is to specify the properties of the variable x and its relationship with observed test scores.

Definition 4.5. The variable x satisfies (A4) for $N_x \in \mathbb{N}$ with respect to s iff:

(i) x follows distribution $H \in \mathcal{F}$ with mean μ_x and variance σ_x^2. Let h denote the pdf associated with H and $h(\textbf{.}|s)$ the conditional pdf of x given s.

(ii) $Support(H) = [0, 1]$.

[17] An airplane pilot who can take off but not land a plane is useless.

[18] The mean difference bounding methodology can be used to study bias in regression coefficients for binary covariates. Recall that for a binary variable D, the coefficient β in the regression $s = \alpha + \beta D + \varepsilon$ identifies the difference in the mean of s conditional on $D = 1$ versus $D = 0$. Therefore, letting group A be those students with $D = 1$ and group B be those students with $D = 0$, the analysis of bias in β is equivalent to the gap-change case presented in section 4.1. Please refer to appendix D.1 for an elaboration of this point.

(iv) $(\mu_{x|s} - \mu_x)$ satisfies (A3) for N_x, where $\mu_{x|s} \equiv \int_0^1 xh(x|s)dx$. Let $s^*_{x,1} < s^*_{x,2} < \ldots < s^*_{x,N_x}$ denote the interior crossing zeros of $(\mu_{x|s} - \mu_x)$.

In essence, definition 4.5 simply guarantees that $(\mu_{x|s} - \mu_x)$ satisfies the same properties that Δf was assumed to satisfy in the mean difference case. The assumption that $(\mu_{x|s} - \mu_x) < 0$ for s sufficiently close to 0 is without loss of generality because $(-x)$ can always be used in place of x.

Consider a linear regression of s on x and let $\hat{\beta}(\mathbb{I})$ denote the corresponding OLS estimator, where the $\mathbb{I}$ argument indicates that the test scores were scaled by the identity function prior to running the regression.[19] The probability limit (plim) of $\hat{\beta}(\mathbb{I})$ is just $\beta(\mathbb{I}) \equiv \frac{cov(s,x)}{var(x)}$. Similarly, $\beta(Y(s)) \equiv \frac{cov(Y(s),x)}{var(x)}$ gives the plim of the OLS coefficient from the regression of $Y(s)$ on x. Define $\mathcal{R}^+(\mathbb{I}, Y, G) \equiv \beta(Y) - \beta(\mathbb{I})$ and $\mathcal{R}^-(\mathbb{I}, Y, G) \equiv \beta(\mathbb{I}) - \beta(Y)$; the bounding exercise amounts to finding weighting functions $Y_0^+(s|k)$ and $Y_0^-(s|k)$ consistent with $D(\mathbb{I}, Y) \leq k$ such that $\mathcal{R}^+$ and $\mathcal{R}^-$ are maximized.[20]

The conditions imposed by assumption (A4) render problem of selecting Y_0^+ and Y_0^- formally equivalent to the selection of W_0^+ and W_0^- under assumptions (A1), (A2), and (A3). The key assumption driving this equivalence is that $(\mu_{x|s} - \mu_x)$ satisfies (A3) and therefore behaves like Δf from the mean-difference case. Theorem 4.6 shows that Y_0^+ and Y_0^- have the same functional form as W_0^+ and W_0^-, but with the zeros of $(\mu_{x|s} - \mu_x)$ playing the role of the zeros of Δf.

Theorem 4.6. *Suppose that (A1), (A2), and (A4) hold for N_x. Then there exist sequences $0 \leq s^-_{x,1} \leq s^-_{x,3} \leq \ldots \leq 1$ and $0 \leq s^+_{x,2} \leq s^+_{x,4} \leq \ldots \leq 1$ such that Y_0^- has the same functional form as W_0^- and Y_0^+ has the same form as W_0^+ from theorem 4.2 but with the $\{s^-_{x,i}\}$ playing the role of the $\{s^-_i\}$ and the $\{s^+_{x,i}\}$ playing the role of the $\{s^+_i\}$.*

Proof. In appendix C. □

Although $Y_0^+(s|k)$ and $Y_0^-(s|k)$ are formally identical to $W_0^+(s|k)$ and $W_0^-(s|k)$, the distributional features of the data that determine their shapes and their ability to generate bias are quite different. In particular, the analog of Δf in this setting is $\Gamma(s) \equiv \left(\sigma_x^2\right)^{-1} f(s)(\mu_{x|s} - \mu_x)$.[21]

[19] All regressions discussed in this section include constants.

[20] At most one of these regression models can be correctly specified. If s is linear in x, then $Y_0(s)$ will not be linear, and vice versa. I am not focused on model specification in this section. Rather, the goal is to determine when scale deviations will cause one to conclude that two variables are positively or negatively associated with each other, when the opposite is the case.

[21] Please see the proof of theorem 4.6 for a demonstration of this claim.

The test scores at which $\left(\mu_{x|s} - \mu_x\right)$ switches sign determines the functional forms of Y_0^+ and Y_0^-. The rate at which bias increases with k depends on the total area between $\Gamma(s)$ and 0, which in turn depends on the magnitude of the variance of x and the interplay between $|\mu_{x|s} - \mu_x|$ and $f(s)$.

The analysis can be extended very easily to multivariable linear regression. Suppose that we estimate the regression model $s = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_N x_N + \varepsilon$. Denote by $\tilde{s}$ and $\tilde{x}_1$ the residuals from regressions of s and x_1 on $\{x_2, \ldots, x_N\}$. The plim of $\hat{\beta}_1(\mathbb{I})$ from the full multivariable regression is $\beta_1(\mathbb{I}) = \frac{cov(\tilde{s},\tilde{x}_1)}{var(\tilde{x}_1)}$. In other words, β_1 is identified by the relationship between those parts of s and x_1 that are orthogonal to the other x's. If the distributions of $\tilde{s}$ and $\tilde{x}_1$ satisfy (A1), (A2), and (A4), then theorem 4.6 will provide valid worst-case residual test score weights for bias in β_1.

4.3. **IV Regression.** Suppose that we are interested in estimating β in the regression $s = \alpha + \beta x + \varepsilon$ for some continuously distributed x, but we fear that $cov(x, \varepsilon)$ is not 0. In response, suppose that we instrument using some continuous variable z that satisfies the exogeneity and relevance conditions necessary to be a valid instrument. If the test scale is not cardinal, then $plim\ \hat{B}_{IV} \equiv \beta_{IV}$ may not identify the sign of the true causal relationship between x and achievement. The question addressed here is how sensitive IV regression is to cardinal differences between s and the true scale $Y_0(s)$.

The theoretical results for IV regression are almost identical to the univariate regression case studied previously. The main difference is that the function playing the role of $\Gamma(s)$ depends now on the joint distribution of the instrument and the endogenous regressor as well as on the conditional mean of the instrument given s. Interestingly, the bounding analysis exposes yet another problem with weak instruments (instruments where $cov(x, z) \equiv \sigma_{xz}$ is close to 0): all else equal, the magnitude of the bias consistent with $D(Y_0, \mathbb{I}) \leq k$ is proportional to $1/\sigma_{xz}$ so that as $\sigma_{xz} \rightarrow 0$, $Bias \rightarrow \infty$.

Theorem 4.7. *Suppose that (A1), (A2), and (A4) hold for N_z. Then there exist sequences $0 \leq s_{z,1}^- \leq s_{z,3}^- \leq \ldots \leq 1$ and $0 \leq s_{z,2}^+ \leq s_{z,4}^+ \leq \ldots \leq 1$ such that Y_0^- has the same functional form as W_0^- and Y_0^+ has the same form as W_0^+ from theorem 4.2 but with the $\{s_{z,i}^-\}$ playing the role of the $\{s_i^-\}$ and the $\{s_{z,i}^+\}$ playing the role of the $\{s_i^+\}$. Moreover, the function which plays the role of $\Gamma(s)$ from theorem 4.6 is $\Gamma_{IV}(s) \equiv \frac{f(s)}{\sigma_{xz}}\left[\mu_{z|s} - \mu_z\right]$.*

Proof. In appendix C. □

Theorem 4.7 simply says that the bounding analysis for β_{IV} is identical to the bounding analysis for β_{OLS} but with $\Gamma_{IV} = \frac{f(s)}{\sigma_{xz}}\left[\mu_{z|s} - \mu_z\right]$ standing in place of $\Gamma = \frac{f(s)}{\sigma_z^2}(\mu_{x|s} - \mu_x)$. The key interaction is between the distribution of s and the difference $(\mu_{z|s} - \mu_z)$; the distribution of x, the ultimate variable of interest, only enters through the scaling term σ_{xz}^{-1}.

The discussion so far has hidden a subtlety generated in the transition from bounding population correlations (OLS) to bounding causal effects (IV). OLS simply seeks to understand the population association between achievement and x. However, in order for IV to be desirable, it must be that we are interested in the *causal effect* of x on achievement. Since Y_0 will not generally be a linear transformation of s, at most one of the two relevant causal relationships can be described fully by a single parameter. For example, suppose that $s = \alpha + \beta x + \varepsilon$ fully describes the relationship between test scores and x, so that β is the causal effect of x on s. In this case, the relationship between x and true achievement is given by $Y_0(s) = Y_0(\alpha + \beta x + \varepsilon)$. The local causal effect of x on Y_0 (assuming differentiability) is $\beta(s) \equiv \frac{\partial Y_0}{\partial s}(s)\beta$ and the average causal effect over the support of s is $\bar{\beta} = \beta \int_0^1 \frac{\partial Y_0}{\partial s}(s) f(s) ds$. Generally, $\beta(s)$ and $\bar{\beta}$ will not equal β, complicating the discussion of bias. However, since $\frac{\partial Y_0}{\partial s} \geq 0$ always, the sign of $\bar{\beta}$ (and $\beta(s)$ when $\frac{\partial Y_0}{\partial s}(s) > 0$) will be determined by the sign of β. This implies that it will suffice to study sign miss-identification for β_{IV} in (s, x)-space, since this will also lead (weakly) to sign miss-identification in (Y_0, x)-space.

This discussion shows that there can be an interesting trade-off between endogeneity bias and cardinality bias. IV will not suffer from endogeneity bias when s is a cardinal measure (which is why it is so popular in empirical work). On the other hand, if σ_{xz} is much smaller than σ_x^2, it is possible for β_{IV} to be much more sensitive than β_{OLS} to scale misspecification. Whether or not IV is preferable to standard OLS depends on the signs and magnitudes of these two types of bias.

5. Empirical Sensitivity Analysis: Mean Differences

This section uses several common data sets to assess the sensitivity of standard achievement gap/change estimates to scale misspecification. The headline conclusion from this exercise is that cross-sectional gaps are often quite robust to scale misspecification, whereas gap changes are typically much less robust. The values of k that are needed to flip the sign of most cross-sectional estimates are quite large, or even non-existent, while the values of k that are needed to flip the sign of many gap-change estimates are often much smaller.

5.1. **Data and Method.** I employ four commonly used surveys: the NLSY 1979, NLSY 1997, NELS 1988, and the ELS 2002. The two NLSY surveys were designed to be nationally representative and directly comparable to each other, as were the NELS and the ELS. Both pairs of surveys have comparable demographic, income, and achievement data that allow one to estimate income and racial achievement gaps/changes. Please refer to appendix E for a more detailed discussion of these data.

I restrict my analysis to students who were between the ages of 15 and 17 at the time of testing. I make this restriction for two reasons. First, students in this age range are relatively close to completing school, so their test scores should provide a summary measure of the cumulative effects of their initial endowments and investments over time by parents, schools, and the students themselves. Second, estimates using a narrow range of student ages are not sensitive to how test scores are adjusted for student age. This is particularly important for the NLSY comparisons because these surveys had very different test-taker age distributions.

Valid gap change estimates require at a minimum that test scores be ordinally comparable over time.[22] Fortunately, it is possible to scale achievement scores in these surveys such that students from the NELS can be ranked consistently against students from the ELS and students in the NLSY79 can be ranked consistently against students in the NLSY97. Although the exact psychometric details differ somewhat between the pairs of surveys, the basic feature that allows such a scaling is the existence of a group of test takers who answered test questions appearing on both of the relevant achievement tests.

Each pair of surveys collect consistently defined and comparable student demographic and household income variables. The demographic comparisons I make are by race and household income. For the NLSY surveys, I use a comprehensive measure of household income that sums income for all household members from all sources. I use this continuous variable to define high-income youth as those respondents with household income in the top 20% of the year-specific household income distribution and low-income youth as those in the bottom 20%. The NELS and ELS surveys only record income categorically, so I define “high-income” and “low-income” to be the sets of categories that most closely approximate the upper and lower quintiles.

[22] Simply normalizing scores to have a mean of 0 and a standard deviation of 1 within each year/age group is not likely to render the scores ordinally comparable. Researchers should have a positive reason for believing that a score s means the same thing in different years.

I estimate eachΔf by first estimating each component density on a grid using a smoothed kernel estimator. I then re-normalize the densities so that each has support on [0,1] and estimate Δf as the sum or difference in these normalized distributions. Importantly, I use the same normalization for all of the component densities in Δf, which guarantees that the normalized scores will still correctly order students from different surveys. W_0^+ and W_0^- are parametrized by their values at the zeros of Δf. Therefore, I search over a grid of all possible values at these crossing points and select the configuration that maximizes bias given k.

5.2. **Black/White Achievement Gaps/Changes.** Figure 7 plots $\Delta V(\mathbb{I})$, $\Delta V(W_0^-)$, and $\Delta V(W_0^+)$ as functions of k for both cross-sectional and gap-change achievement estimates using the NLSY data. The cross-sectional plots show that as k grows larger, $\Delta V(W_0^+)$ and $\Delta V(W_0^-)$ for both math and reading fan out from their observed values. The $\Delta V(W_0^-)$'s for math never cross 0, while for reading they cross at around $k \approx 0.4$. The observed black/white reading achievement gaps in the NLSY79 and NLSY97 may not correctly identify the sign of the true gaps, while the sign of the cross-sectional math gaps will never be misidentified. Although the $\Delta V(W_0^-)$'s for reading do cross 0, they remain very close to the horizontal axis for large k. The gap-change plots also show $\Delta V(W_0^-)$ and $\Delta V(W_0^+)$ fanning out from their observed values as k increases. However, unlike the cross-sectional ΔV's, the gap-change $\Delta V(W_0^+)$'s cross 0 for relatively small values of k.[23] For values of k greater than about 0.1, the true gap change could be positive for either math or reading, while the corresponding observed gap change would be negative. The NLSY gap-change estimates are substantially less robust to cardinal deviations than the cross-sectional estimates.

Figure 7 is representative of the various achievement gap/change comparisons in both the NLSY and NELS/ELS data. Rather than present similar graphs for each comparison, I summarize the relevant robustness information in tables 4 and 5 in appendix B. These tables show k^* for each gap/change estimate, where k^* is defined as the smallest value of k such that the true and observed estimates have opposite signs.[24]

The qualitative results on black/white achievement inequality are similar using the NELS/ELS data. The cross-sectional estimates for both math and reading in the ELS are not reversible; there is no value of k such that the observed and true gaps have opposite signs. In the NELS,

[23] The cross-sectional gaps are positive, so only the $\Delta V(W_0^-)$'s have a chance at reversing sign. In contrast, the observed gap-change estimates are all negative, so the only candidates for reversal are the $\Delta V(W_0^+)$'s.

[24] Formally, define $k^* = \inf\{k|\Delta V(W_0^-(s|k), \Delta f) < 0\}$ if $\Delta V(\mathbb{I}, \Delta f) > 0$ and $k^* = \inf\{k|\Delta V(W_0^+(s|k), \Delta f) > 0\}$ if $\Delta V(\mathbb{I}, \Delta f) < 0$ in the case the a sign flip is possible.

such a reversal is possible for both math and reading, but only for values of k greater than about 0.33. Both the math and the reading gap-change estimates are reversible for values of k greater than about 0.29, which is a substantially higher value than what was required in the NLSY. It is also worth noting that the observed sign of the gap change in the NELS/ELS data is positive for both math and reading, while both observed gap changes in the NLSY data are negative.

5.3. **High-/Low-Income Achievement Gaps/Changes.** I now repeat the sensitivity analysis for achievement gaps/changes for youth from high- versus low-income households. Tables 4 and 5 generally suggest that income-achievement gaps/changes are less robust than black/white gaps/changes. Furthermore, the gap-change estimates are again substantially less robust than the cross-sectional estimates.

In the NELS/ELS data, neither of the cross-sectional gaps in math are reversible for any k. The reading achievement gap is not reversible in the ELS, while a reversal is only possible for k greater than 0.38 in the NELS. In contrast, all of the cross-sectional gaps are reversible in the NLSY. The math and reading gaps in the NLSY79 can be flipped for k greater than about 0.13, indicating that these estimates are quite sensitive to cardinal deviations. The NLSY97 gaps are more robust, with k^*'s of 0.2 (reading) and 0.33 (math). With the exception of the math gap change in the NLSY ($k^* = 0.27$), all of the gap-change estimates using either data source are very sensitive to scale misspecification. Each of the other gap-change estimates has a k^* less than 0.1. Even minor rescalings may be sufficient to reverse conclusions about trends in achievement inequality by household income.

5.4. **What if Z-Scores Are Used?** The calculations in sections 5.3 and 5.2 deviate from most of the literature on achievement inequality in that they do not use z-scores (scores in standard-deviation units) to estimate achievement differences. Instead, they use scores that enable one to rank students from different surveys against each other.[25] There are strong reasons to prefer such ordinally comparable scores, and there is no reason to think that z-score gap/change estimates will be particularly robust to scale misspecification. Indeed, tables 4 and 5 show that gaps/changes estimated using NELS/ELS z-scores are roughly as fragile as estimates using ordinally comparable scores. The cross-sectional gap estimates are mostly not reversible, or are reversible only for fairly large values of k. The gap-change estimates using

[25] That is, scores such that $s_i > s_j$ implies that student i has more achievement than student j regardless of whether i and j are from the same survey.

z-scores are uniformly quite fragile, with the black/white gap-change estimates substantially more fragile than estimates using ordinally comparable scores.

5.5. **The Magnitude of k.** Some achievement gaps/changes are sign-identified in the NELS/ELS and NLSY data no matter how different the true and observed test scales are. For other achievement gaps/changes, the sign may be misidentified by the observed test scores for sufficiently large values of k. The magnitude of the smallest k for which a sign reversal is possible varies enormously across different comparisons, from a minimum of 0.04 to a maximum of 0.4. Since the bounding analysis is well-defined for any k in [0,1], a value of 0.04 might seem small and 0.4 might seem large. However, it is not entirely clear what the scale of k means. Pinning down the scale of k is a fundamentally hard problem since the relevant units of achievement are not knowable (otherwise there would be no need to go through the bounding analysis). This section explores a number of methods to determine what constitutes a "large" or a "small" value of k.

Education researchers are familiar with test scores normalized to have a mean of 0 and a standard deviation of 1. Although my work here questions whether such z-scores have an interpretable scale, it is still possible to report ΔV^{+}, ΔV^{-}, and k in standard-deviation units. For instance, the math z-scores in the NELS and ELS have a range of -2.2 to 2.4, which implies that $k = 0.04$ corresponds to $0.18 = (2.4 + 2.2) \times 0.04$ standard-deviation units, while $k = 0.4$ corresponds to 1.8 standard-deviation units. Students typically gain about 0.07 standard deviations of achievement per month in primary school, so a difference of 0.18 is neither very large nor very small by this metric, while 1.8 is huge.[26] Cross-sectional black/white and high-/low-income achievement gaps are typically around 0.5 to 0.8 standard deviations, again making $k = 0.04$ seem relatively small and $k = 0.4$ relatively large.[27]

[26] Krueger[15] uses the Tennessee STAR experiment to estimate that smaller class sizes correspond to about 0.22 standard deviations. He argues that this figure corresponds to about 3 months of progress in school. Since most of the literature examining the effects of various inputs on student achievement apply cardinal methods to z-scores, I can compare the "z-score" units of k to virtually any educational effect size I wish. For example, Hanushek and Rivkin [12] review the literature on teacher value-added models and report that a standard deviation in teacher performance is associated with student gains on the order of 0.1 to 0.2 standard deviations.

[27] In my data, the black/white math gap is 0.79 in the NELS and 0.84 in the ELS. Fryer and Levitt[9] estimate black/white achievement gaps for early elementary school students of between 0.4 to 0.7. Reardon[23] estimates the math achievement gap between students from the 90th and 10th percentiles of the household income distribution to be around 1 in the NELS and 1.1 in the ELS, whereas I estimate that the NELS math income-achievement gap is 1.039 and in the ELS it is 0.904.

6. Empirical Sensitivity Analysis: Regression

This section assesses the sensitivity of ordinary least squares regression to scale misspecification when test scores are used as the outcome variable. I estimate simple linear regressions between household income and math, reading, and composite achievement test scores in both NLSY surveys. I then implement the bounding methodology outlined in section 4.2 to assess the robustness of the estimated regression coefficients to order-preserving transformations of the test scores. I also estimate the sensitivity of the estimated change in the coefficients from the NLSY79 to the NLSY97. As with the mean-based estimates, the cross-sectional regression coefficients are very robust to scale misspecification. It is always impossible to flip the sign of the cross-sectional regression coefficients, no matter how different are the true and observed scales. The coefficient difference estimates are much less robust. In all cases, there exist k's such that the true and observed differences have opposite signs, and for some achievement measures even small values of k are sufficient for the true and observed differences to have opposite signs.

6.1. **Data and Method.**

I use the same subset of the NLSY data as in the mean difference analysis. For each achievement measure and each survey, I regress the normalized test scores s on p, the survey-specific income percentiles.[28] Implementing the sensitivity analysis requires the estimation of the conditional mean of p given s ($\mu_{p|s}$) and the density of s ($f(s)$). I use standard kernel methods to estimate the test score densities as in section 5. I use local polynomial regression to estimate the conditional mean of p given s in a flexible, non-parametric way. The resulting sensitivity estimates do not seem to be sensitive to the particular kernels and smoothing parameters used.

The theory developed in section 4.2 does not quite apply to a *difference* in regression coefficients. However, theorem 4.6 can be extended easily to cover this new case. The key new requirement is that the following function satisfy assumption (A3):

$$(\sigma^2_{p,t})^{-1} f_t(s)(\mu_{p|s,t} - \mu_{p,t}) - (\sigma^2_{p,t-1})^{-1} f_{t-1}(s)(\mu_{p|s,t-1} - \mu_{p,t-1}).$$

[28] To be precise, I estimate regressions of the form $s_{i,t} = \alpha_{s,t} + \beta_{s,t} p_{i,t} + \varepsilon_{i,t}$. I use the income percentile, rather than the raw income level normalized to fit in [0,1], for two reasons. First, it is computationally convenient for the income distributions across both surveys to have the same variance. Second, papers such as Reardon[23] estimate similar regressions in order to assess changes in achievement inequality at different relative locations in the income distribution. For example, $\Delta\hat{\beta}(0.9 - 0.1)$ gives the change in the expected test score gap between youth at the 90th versus 10th percentiles of the household income distribution.

Appendix D.3 demonstrates this claim formally.

6.2. **Regression Estimates Using Income.** Table 6 displays the baseline regression coefficients relating income and test scores. Income is strongly correlated with all three measures of achievement in both NLSY surveys. The relationship between income and achievement seems to be slightly weaker in the NLSY97 than in the NLSY79 – the regression coefficients and R^2's are slightly lower for each achievement measure. Based on these regression results, an analyst willing to treat test scores as cardinal measures of achievement would conclude both that substantial achievement differences by income class exist in both surveys and that these differences are smaller in then NLSY97 than in the NLSY79.

Figure 9 and Table 7 display the sensitivity estimates for the cross-sectional regression coefficients and their first-differences. The basic conclusions here are similar to the mean gap/change analysis in section 5. The $\hat{\beta}_{s,t}$ estimates are uniformly robust to cardinal deviations; it is never possible to flip their sign and the lower bounds always remain substantially above 0. In contrast, it is always possible to flip the sign of $\Delta\hat{\beta}$ for sufficiently large values of k. Moreover, the smallest k's needed to flip the sign are often quite small. For example, the sign of the coefficient difference for math achievement may be misidentified for values of k greater than 0.04. Estimated trends in achievement inequality once again appear to be less robust than estimated cross-sectional inequality.

7. Estimation and Measurement Error

The empirical analysis so far has ignored estimation error. The functions that critically determine the sensitivity of the gap/change estimates in sections 5-6 (Δf, Γ, and Γ_{IV}) are themselves estimated. Moreover, given an estimate for one of these functions, I approximate the corresponding worst-case test scales on a finite grid of points and use these approximations to estimate k^*. These multiple layers of estimation error mean that the estimated k^*'s I report are noisy estimates of their true population values.

From one perspective, this concern is secondary to the main thrust of this paper. The estimated functions Δf, Γ, and Γ_{IV} are consistent; as such, they are plausible guesses for the kinds of functions that govern bias in important, applied settings. The estimated k^*'s suggest that for many such functions, sign reversals are possible with even mild rescalings of test scores. Even without knowing the estimation errors associated with my empirical procedure, I have certainly supplied evidence that standard methods applied to test-score data can easily be quite sensitive to scale misspecification.

However, in order to state with confidence that the specific estimates I have identified as being sensitive are in fact sensitive to cardinal deviations, I need some way to account for estimation error. Bootstrapping is difficult to implement in this setting because the forms of the worst-case weights depend on the number and location of the zeros of G, and different bootstrap iterations may result in G's that cross 0 a different number of times. Working out a valid and computationally feasible way to conduct inference in this setting is on the agenda for future research.

This paper also does not tackle the problem of test-score measurement error. I have explicitly assumed that observed test scores perfectly order students according to their true achievement. In practice, however, tests are noisy measures of achievement – student rank-orders change somewhat from test to test. Fully characterizing the effects of test-score measurement error is a project for future work. Nonetheless, it is clear that classical test-score measurement will tend to exaggerate the apparent robustness of mean-based achievement gap/change estimates. The intuition is that measurement error will tend to make the group-level test score distributions less distinct, lowering the total area between Δf and 0, while leaving the estimated mean gap/change unchanged. The closer Δf is to 0, the less "room" there is for scale-misspecification to create bias. Therefore, mean gap/change sensitivity estimates calculated from noisy test score data will tend to understate how sensitive a given mean difference is to cardinal deviations in the scale of achievement.

8. Conclusion and Extensions

This paper develops a method for assessing the sensitivity of standard achievement gap/change estimates to test scale misspecification. The method makes precise the intuitive idea that cardinal methods will provide valid inference on the sign of achievement differences and trends when the true scale and the observed scale are close to each other and incorrect inference when the two scales are very different. The approach is readily interpretable and straightforward to apply in many real-world empirical scenarios.

I use the method to investigate the cardinal sensitivity of standard achievement gap/change estimates in the NLSY and NELS/ELS data. I find that cross-sectional black/white and high-/low-income achievement gaps are usually robust to scale misspecification in these data. In many cases, there is no rescaling of the test scores that would reverse the sign of the estimated gap, while in other cases the true scale would have to be quite different from the observed scale in order for the sign to be misidentified. In contrast, achievement gap-change estimates in these data are much less robust; even small differences between the true and observed scales are often sufficient to reverse the sign of an estimated trend.

The same basic pattern holds empirically for bias in regressions of test scores on income. In both the NLSY79 and NLSY97, income is strongly positively associated with achievement, and there is no way to reverse this conclusion by rescaling the test scores. In contrast, the estimated change in this association from the NLSY79 to the NLSY97 is much more fragile.

Cardinal statistical methods are easy to use and familiar to most researchers. If the observed test scale is close to the true scale, cardinal methods are preferable because they have greater power than ordinal approaches and will not misidentify the sign. This paper has shown that relying on such methods may lead one very far astray if the true scale and the observed scale are sufficiently different from each other. Ultimately, the true scale of achievement is unknowable in most applied work; researchers must use their best judgment about how best to employ test-score data. However, if my sensitivity method shows that a given conclusion using cardinal methods is quite sensitive to the (essentially arbitrary) test scale used, applied researchers may wish to abandon cardinal approaches and instead rely only on the ordinal content of the test scores. If ordinal methods yield ambiguous results, then researchers should invest more effort in crafting test scales that are plausibly cardinal for the application at hand.

Both the theoretical and empirical work presented here are quite preliminary, and each calls out for a number of extensions. The bounding analysis depends on the choice of distance measure. The sup norm yields tractable expressions for the worst-case score weighting functions for both mean differences and regression coefficients. Nonetheless, other distance measures may produce bounds that are easier to interpret. Empirically, it would be worthwhile to extend the sensitivity analysis to other achievement gaps/changes and other data sets. It would also be useful to work out more completely how to conduct valid inference on k^*. Finally, future work should investigate the empirical relevance of the methods presented here to other empirical settings, such as value-added models and poverty indices.

References

[1] Rolf Aaberge, Tarjei Havnes, and Magne Mogstad. A Theory for Ranking Distribution Functions. *IZA Discussion Papers no 7738*, 2013.

[2] Joseph Altonji, Prashant Bharadwaj, and Fabian Lange. Changes in the Characteristics of American Youth: Implications for Adult Outcomes. *Journal of Labor Economics*, 30, 4:783–828, 2011.

[3] Anthony B. Atkinson. On the Measurement of Inequality. *Journal of Economic Theory*, 2:244–263, 1970.

[4] Timothy Bond and Kevin Lang. The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results. *Review of Economics and Statistics*, 95:1468–1479, 2013.

[5] Elizabeth Cascio and Douglas Staiger. Knowledge, Tests, and Fadeout in Education Intervention. *NBER Working Papers*, 18038, 2012.

[6] Charles Clotfelter, Helen Ladd, and Jacob Vigdor. The Academic Achievement Gap in Grades 3-8. *The Review of Economics and Statistics*, 91:398–419, 2009.

[7] Flavio Cunha and James. Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Journal of Human Resources*, 43(4):738–782, 2008.

[8] Greg Duncan and Katherine Magnuson. The Role of Family Socioeconomic Resources in the Black-White Test Score Gap Among Young Children. *Developmental Review*, 87:365–399, 2006.

[9] Roland G. Fryer and Steven D. Levitt. Understanding the Black-White Test Score Gap in the First Two Years of School. *The Review of Economics and Statistics*, 86(2):447–464, 2004.

[10] Roland G. Fryer and Steven D. Levitt. The Black-White Test Score Gap Through Third Grade. *American Law and Economics Review*, 8:249–81, 2006.

[11] Eric Hanushek and Steven Rivkin. School Quality and the Black-White Achievement Gap. *NBER Working Papers*, 12651, 2006.

[12] Eric Hanushek and Steven Rivkin. The Distribution of Teacher Quality and Implications for Policy. *Annual Review of Economics*, 4:131–57, 2012.

[13] Caroline Hoxby. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics*, 115(4):1239–1285, 2000.

[14] Tim Kautz, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success. *OECD Report*, 2014.

[15] Alan Krueger. Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*, 115(2):497–532, 1999.

[16] Kevin Lang. Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member. *Journal of Economic Perspectives*, 24:167–181, 2010.

[17] Frederic Lord. The 'Ability' Scale in Item Characteristics Curve Theory. *Psychometrika*, 40:205–217, 1975.

[18] Derek Neal. *Why Has Black-White Skill Convergence Stopped?*, volume 1, chapter 9, pages 511–576. Elsevier, Amsterdam, 2006.

[19] Eric Nielsen. *The Income-Achievement Gap and Adult Outcome Inequality.* PhD thesis, University of Chicago, 2014.

[20] Eric Nielsen. The Income-Achievement Gap and Adult Outcome Inequality. *Finance and Economics Discussion Series, Board of Governors of the Federal Reserve System (U.S.)*, 041, 2015.

[21] Stephen Raudenbush. What Are Value-Added Model Estimating and What Does This Imply for Statistical Practice? *Journal of Educational and Behavioral Statistics*, 29 (1):121–129, 2004.

[22] Sean Reardon. Thirteen Ways of Looking at the Black-White Test Score Gap. CEPA Working Paper, Stanford University, 2007.

[23] Sean Reardon. *The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations*, chapter 5, pages 91–116. Russell Sage Foundation, New York, July 2011.

[24] Carsten Schroeder and Shlomo Yitzhaki. Revisiting the Evidence for a Cardinal Treatment of Ordinal Variables. *Working Paper*, 2015.

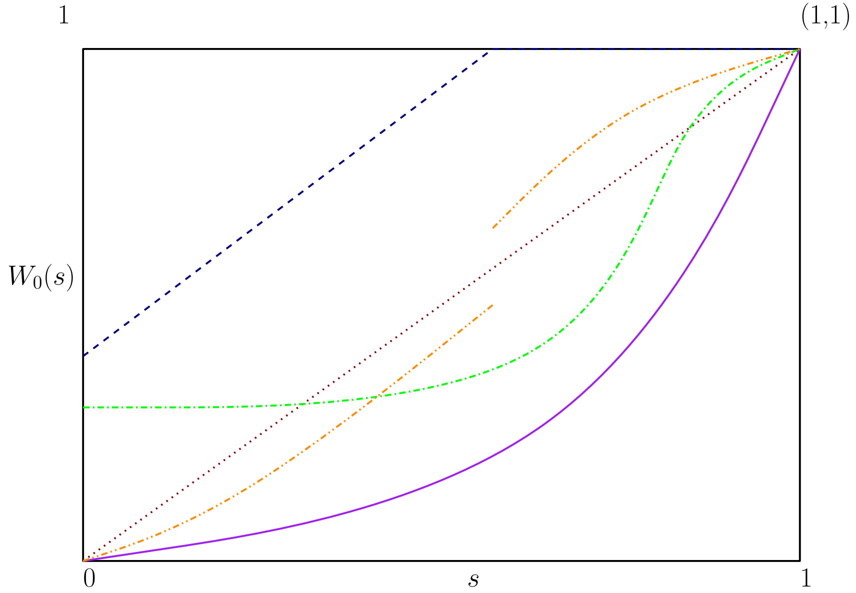
[25] D. Segall. Equating the CAT-ASVAB. In *Computerized Adaptive Testing: From Enquiry to Operation.* American Psychological Association, 1997.

[26] D. Segall. Chapter 18: Equating the CAT-ASVAB with the P&P-ASVAB. (from) CATBOOK, Computerized Adaptive Testing: From Enquiry to Operation. Technical report, United States Army Research Institute for the Behavioral and Social Sciences, 1999.

[27] S. Stevens. On the Theory of Scales of Measurement. *Science*, 103:677–680, 1946.

Appendix A. Figures

Figure 1. Functions Satisfying (A2)



Note: Plot shows five weighting functions consistent with (A2). The red curve is the identity and is the weighting function assumed when achievement gaps/changes are estimated using differences in sample means. The other curves (in purple, green, orange, and blue) demonstrate the W_0 can be convex, concave, discontinuous, and non-differentiable and still satisfy (A2).

Figure 2. Examples of Δf's Satisfying (A3)

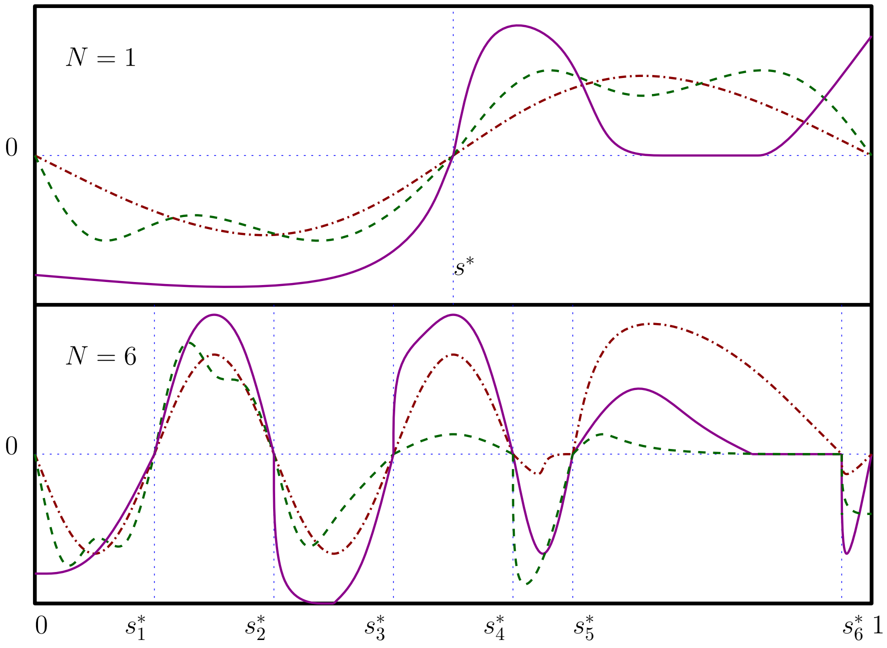
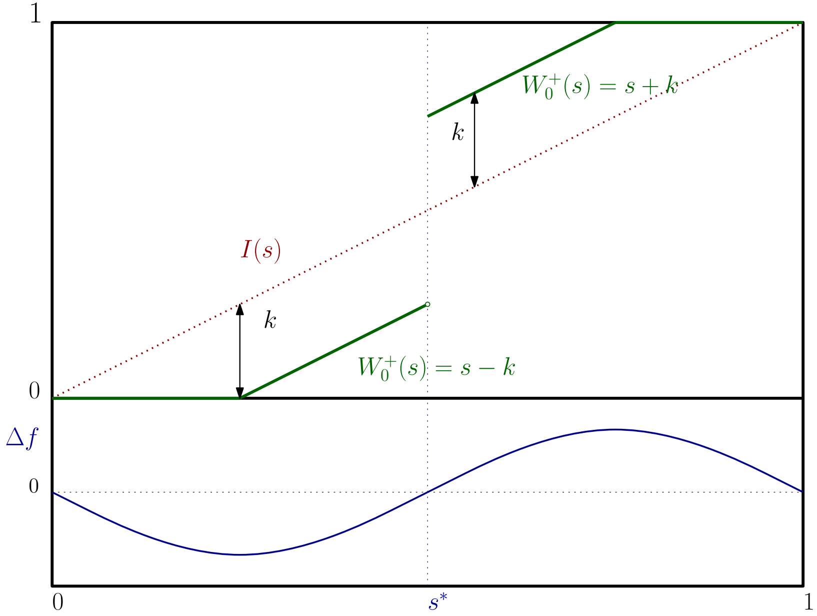
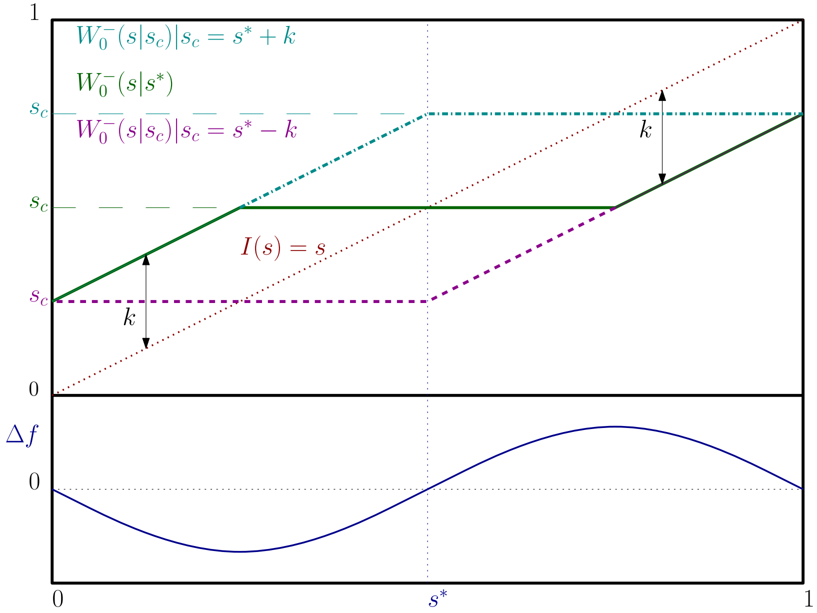


FIGURE 3. $W_0^+(s|k)$ with One Crossing Point



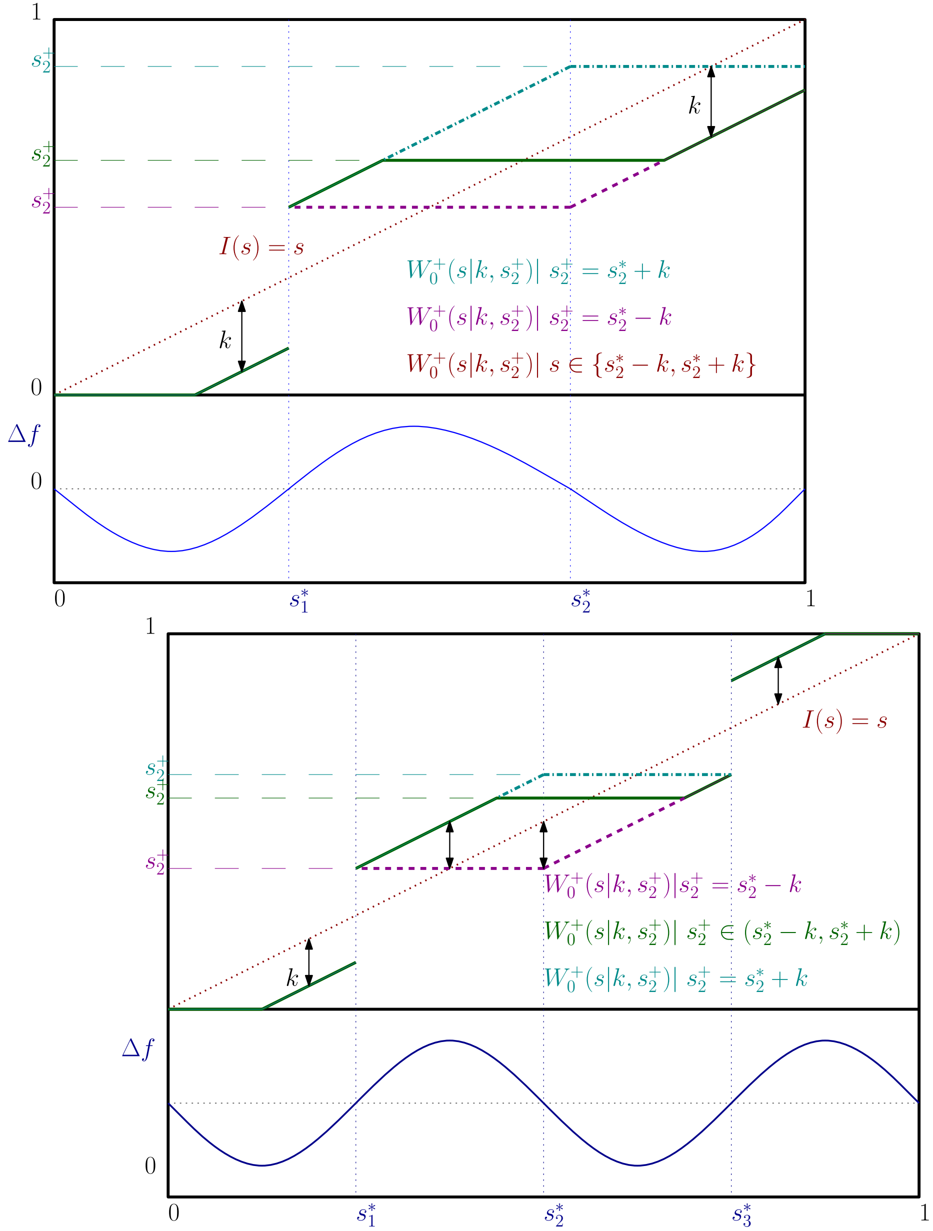
Note: The green curve plots W_0^+ when $k < \min\{s^*, 1 - s^*\}$. For values of s less than k or greater than $1 - k$, W_0^+ is flat. W_0^+ increases 1-1 with s on the interval $[k, 1 - k]$ except for the point $s^* = 0.5$, where W_0^+ jumps by $2k$.

FIGURE 4. Potential $W_0^-(s|k)$'s with One Crossing Point



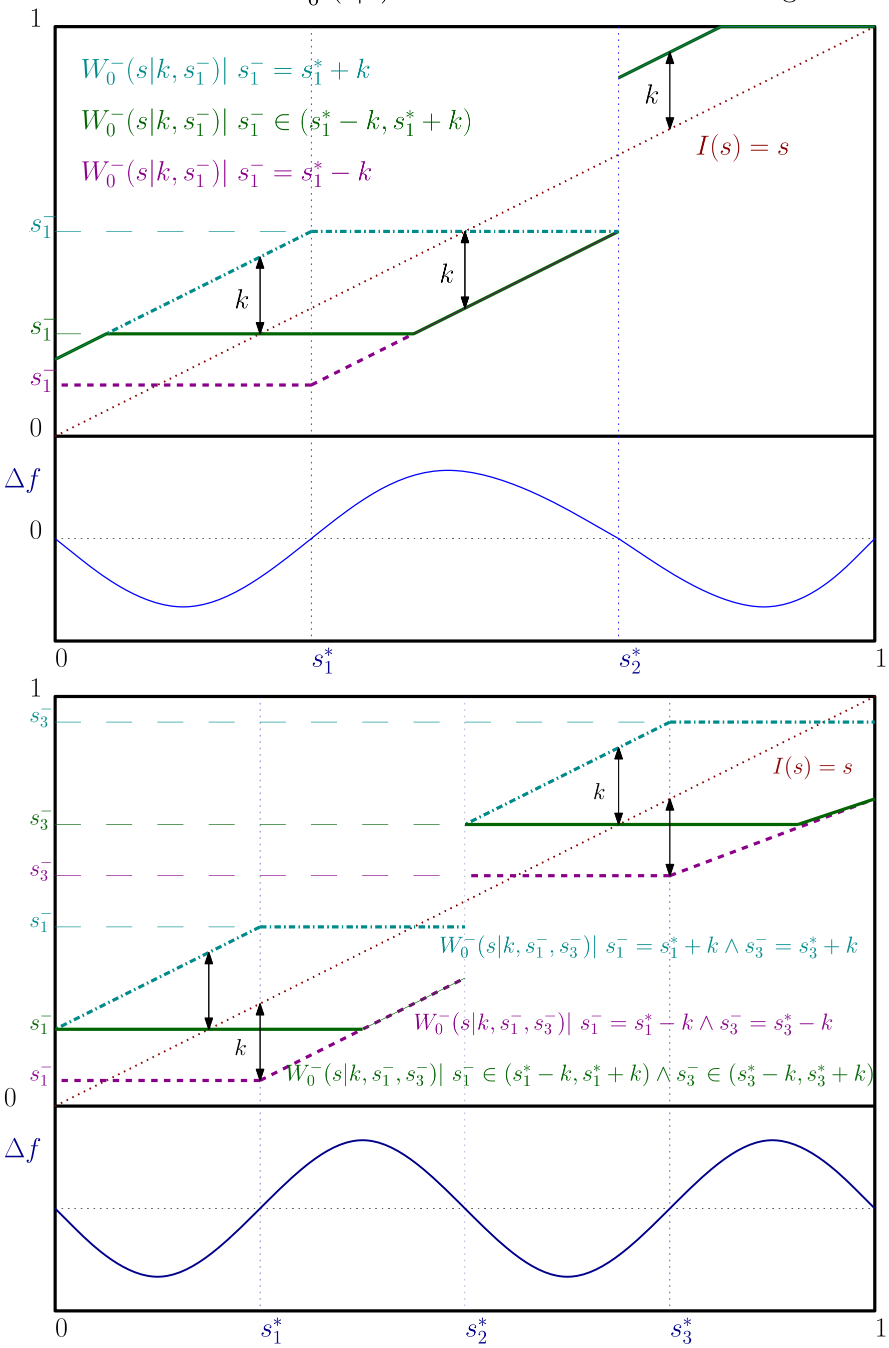
Note: The function in green plots $W_0^-(s|s^*, k)$ when $s^* - k > 0$ and $s^* + k < 1$. In this case, the constraint that $D(\mathbb{I}, W_0^-) \leq k$ binds both above and below s^*. The purple curve shows $W_0^-(s|s_c, k)$ for $s_c = s^* - k$ where k is such that $s_c - k < 0$. In this case, $D(\mathbb{I}, W_0^-)$ only binds above s^*. Symmetrically, the teal curve plots $W_0^-(s|s_c, k)$ when $s_c = s^* + k$ and k is such that $D(\mathbb{I}, W_0^-)$ only binds below s^*.

FIGURE 5. Potential $W_0^+(s|k)$'s with Two and Three Crossing Points


1
s_2^+
s_2^+
s_2^+
k
$I(s) = s$
$W_0^+(s|k, s_2^+)|\ s_2^+ = s_2^* + k$
$W_0^+(s|k, s_2^+)|\ s_2^+ = s_2^* - k$
$W_0^+(s|k, s_2^+)|\ s \in \{s_2^* - k, s_2^* + k\}$
k
0
Δf
0
0
s_1^*
s_2^*
1
1
$I(s) = s$
s_2^+
s_2^+
s_2^+
$W_0^+(s|k, s_2^+)|s_2^+ = s_2^* - k$
$W_0^+(s|k, s_2^+)|\ s_2^+ \in (s_2^* - k, s_2^* + k)$
$W_0^+(s|k, s_2^+)|\ s_2^+ = s_2^* + k$
k
Δf
0
0
s_1^*
s_2^*
s_3^*
1

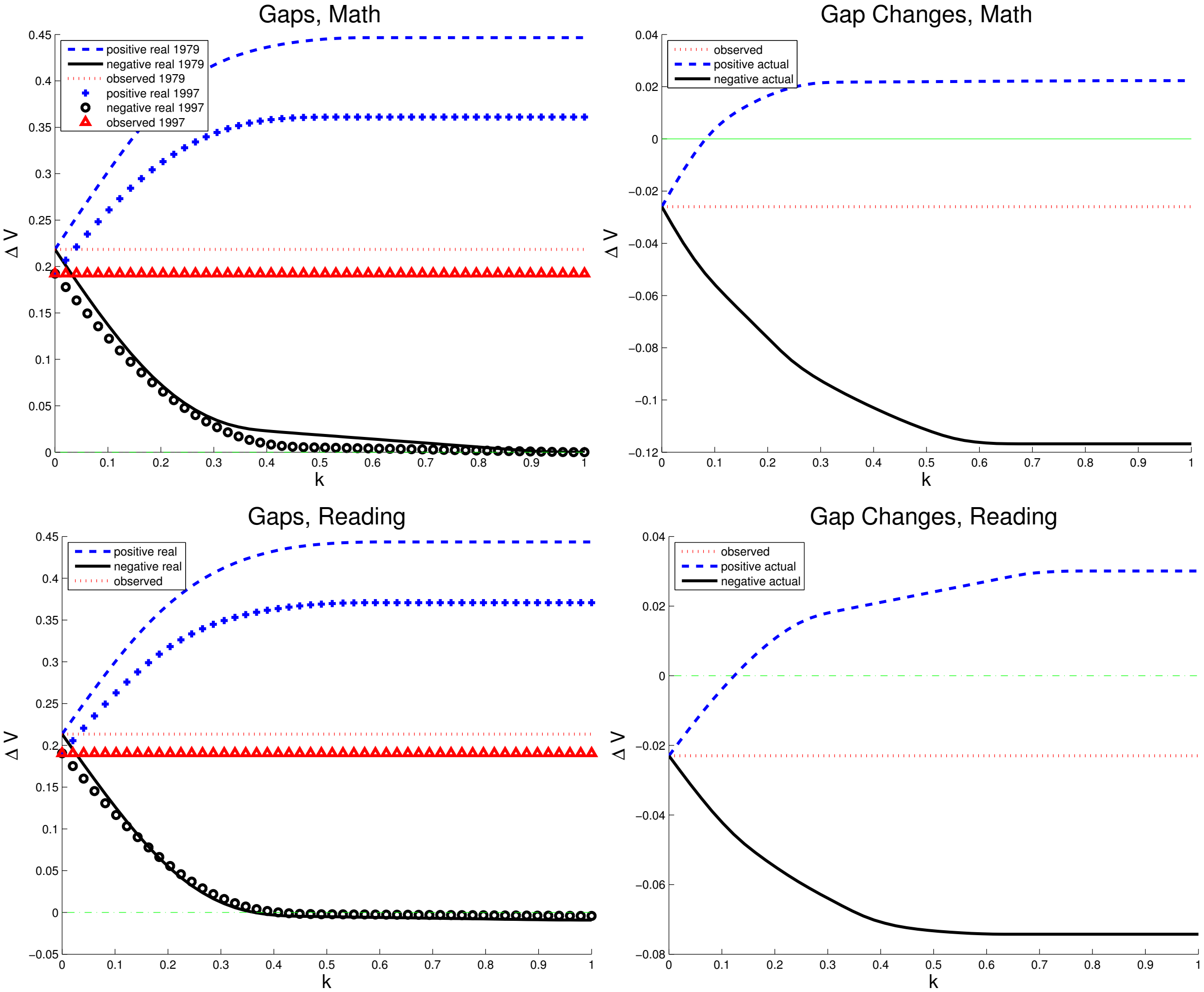
Note: The potential W_0^+'s are indexed by $W_0^+(s_2^*) \equiv s_2^+$. The curves in magenta depict the case that $s_2^+ = s_2^* - k$ while the curves in teal set $s_2^+ = s_2^* + k$. The curves in green show intermediate cases where s_2^+ lies between these two extremes.

Figure 6. Potential $W_0^-(s|k)$'s with Two and Three Crossing Points



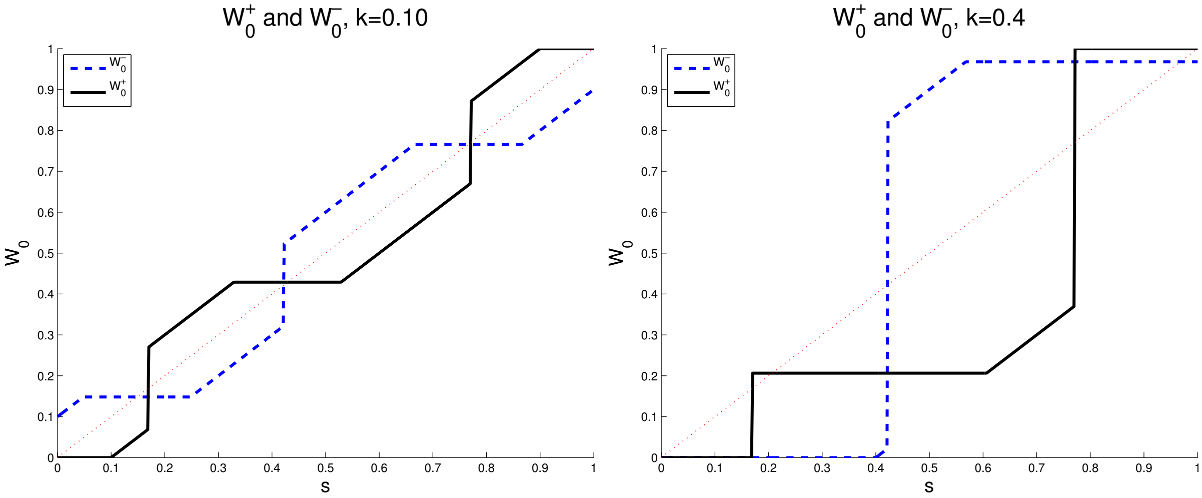
Note: The potential W_0^-'s are indexed by $W_0^-(s_1^*) \equiv s_1^-$ and $W_0^-(s_3^*) \equiv s_3^-$ (for $N = 3$). The curves in magenta depict the case that $s_i^- = s_i^* - k,\ i \in \{1, 3\}$, while the curves in teal set $s_i^- = s_i^* + k$. The curves in green show intermediate cases where both values of s_i^- lie between these two extremes.

Figure 7. Black/White Achievement Gap/Change Bounds, NLSY



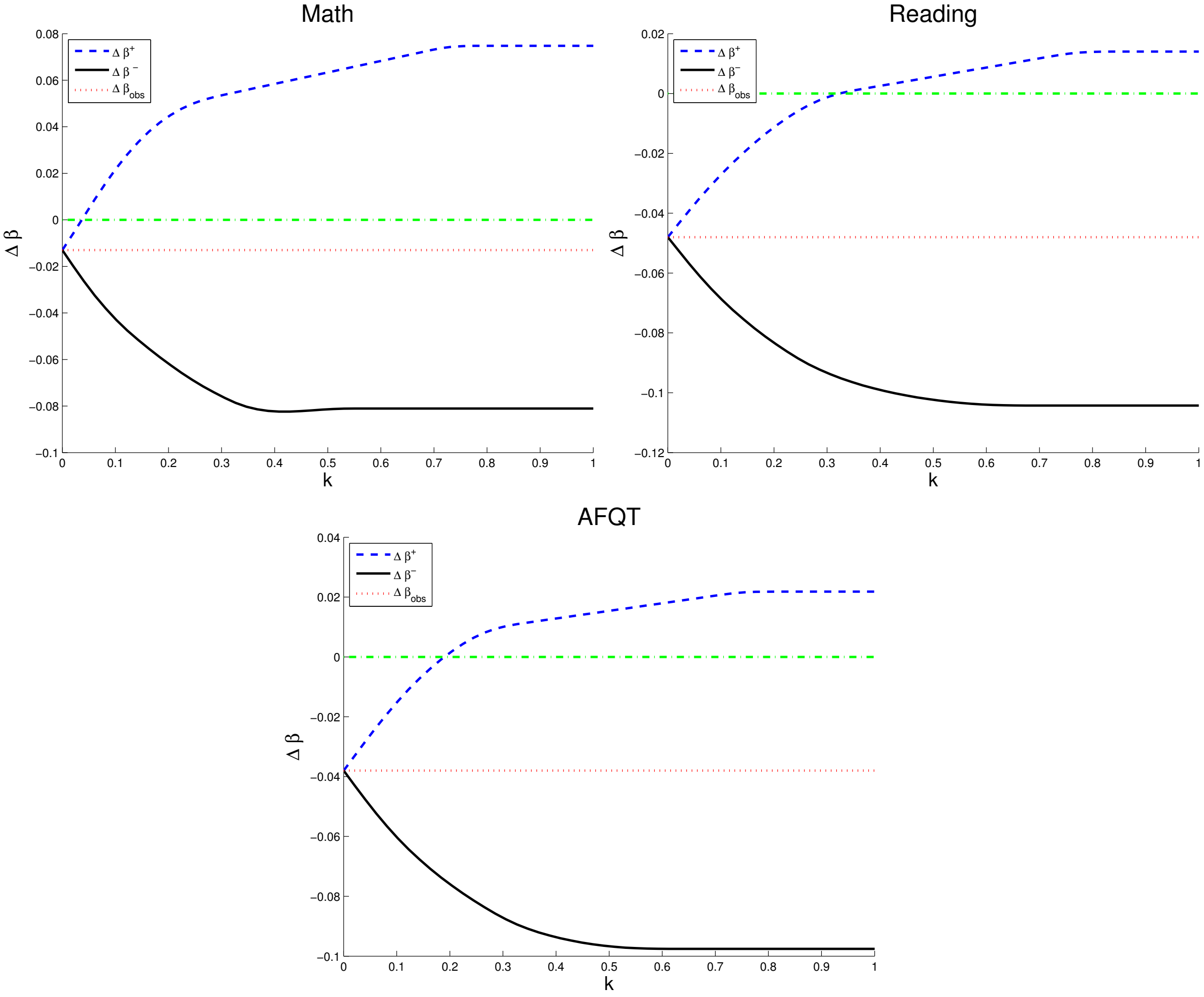
Note: Curves estimated using Δf's calculated on a grid of 5,000 evenly spaced points and 50 evenly spaced values of k. The left-hand panels show the cross-sectional gaps for the NLSY79 (solid) and NLSY97 (dashed) calculated such that the differences in the observed curves (in red) equal the observed gap changes in the right-hand panels. Data cleaned as described in section 5 and appendix E.

FIGURE 8. W_0^+ and W_0^- for Math Income-Achievement Gap Changes, NELS/ELS



Note: Curves estimated using Δf's calculated on a grid of 5,000 evenly spaced points. Data cleaned as described in section 5 and appendix E. For these data, $k = 0.1$ is sufficient for the observed test scores to misidentify the sign of the true gap change. The right panel shows that the worst case weighting functions for $k = 0.1$ do not look particularly extreme. Under both W_0^+ and W_0^-, the observed scores are cardinal for most of [0,1], and neither weighting function ever strays too far from the identity function. In contrast, W_0^+ and W_0^- look very different from the identity when $k = 0.4$; the observed scores are almost never cardinal and the jumps at the achievement thresholds are very large.

Figure 9. Income Regression Coefficient Bounds, NLSY



Note: Curves estimated on a grid of 1,000 evenly spaced points and 50 evenly spaced values of k. Data cleaned as described in section 5 and appendix E.

Appendix B. Tables

Table 1. NLSY Summary Statistics

Variable	Survey	N	Mean	Median	S.D.
math	NLSY79	3,277	96.77	95	18.23
math	NLSY97	2,833	98.74	99	18.82
reading	NLSY79	3,277	94.19	98	19.32
reading	NLSY97	2,833	93.41	98	20.39
AFQT	NLSY79	3,277	142.57	146	26.94
AFQT	NLSY97	2,833	142.88	147.4	28.11
income	NLSY79	3,388	$44,000	$39,800	$28,700
income	NLSY97	3,570	$54,700	$43,100	$49,500
age	NLSY79	3,388	16.08	16	0.78
age	NLSY97	3,570	15.76	16	0.72
black	NLSY79	3,388	0.14	0	0.35
black	NLSY97	3,570	0.15	0	0.36

Note: Respondent ages are restricted to 15-17 as of ASVAB test date. All dollars have been converted to a 1997 basis using the CPI-U. The N shown for a variable is the sample size used in calculations involving that variable. Data cleaned as described in section 5 and appendix E.

Table 2. NELS/ELS Summary Statistics

Variable	Survey	Wave	N	Mean	Median	S.D.	Missing
math	NELS	1990	14,410	44.03	44.31	13.57	777
math	NELS	1992	12,008	49.00	49.53	14.07	2,138
reading	NELS	1990	14,427	30.93	31.38	9.91	760
reading	NELS	1992	11,999	33.33	34.68	10.01	2,147
age	NELS	1990	15,187	16.13	16	0.68	0
age	NELS	1992	14,146	18.14	18	.62	0
black	NELS	1990	15,187	0.12	0	0.32	0
black	NELS	1992	14,146	0.11	0	0.32	0
female	NELS	1990	15,187	0.51	1	0.50	0
female	NELS	1992	14,146	0.50	1	.50	0
math	ELS	2002	14,934	44.62	44.79	13.57	0
math	ELS	2004	13,444	50.22	51.38	14.13	1,148
reading	ELS	2002	14,934	29.29	29.65	9.44	0
reading	ELS	2004	NA	NA	NA	NA	NA
age	ELS	2002	14,934	15.67	16	0.61	0
age	ELS	2004	14,592	17.70	18	0.61	0
black	ELS	2002	14,592	0.14	0	0.35	0
black	ELS	2004	14,934	0.14	0	0.35	0
female	ELS	2002	14,934	0.50	0	0.50	0
female	ELS	2004	14,592	0.50	0	0.50	0

Note: Statistics shown for the NELS first-year follow up (1990) and the ELS base year (2002). Respondent ages restricted to 15-17 as of survey date. Averages shown for non-missing, non-imputed observations using cross-sectional weights. NELS 1990 sample includes “freshened” observations. Data cleaned as described in section 5 and appendix E.

Table 3. NELS/ELS Income Variables

NELS	Percentage	Percentage	ELS	Percentage	Percentage
Income	Full Sample	Analysis Sample	Income	Full Sample	Analysis Sample
none	.26	.27	none	.45	.43
less than $1,000	.49	.48	less than $1,000	1.09	1.14
$1,000-$2,999	1.07	1.13	$1,001-$5,000	1.73	1.78
3,000-$4,999	1.57	1.60	$5,001-$10,000	2.12	2.08
$5,000-$7,499	2.68	2.82	$10,001-$14,000	4.22	4.27
$7,500-$9,999	3.13	3.10	$15,001-$20,000	4.87	4.95
$10,000-$14,999	7.26	7.48	$20,001-$25,000	6.53	6.47
$15,000-$19,999	7.08	7.21	$25,001-$35,000	12.21	12.40
$20,000-$24,999	10.17	10.44	$35,001-$50,000	19.69	19.65
$25,000-$34,999	19.34	19.18	$50,001-$75,000	21.03	20.81
$35,000-$49,999	21.98	21.59	$75,001-$100,000	13.14	13.09
$50,000-$74,999	16.41	16.30	$100,001-$200,000	10.20	10.19
$75,000-$99,999	4.07	4.03	$200,001 or more	2.74	2.75
$100,000-$199,999	3.21	3.16			
$200,000 or more	1.26	1.21			

Note: Dollar ranges shown in survey-specific base-year real dollars (1988 for the NELS and 2002 for the ELS). The full sample columns show the cross-sectionally weighted percentages for the full range of ages in each survey base year. The analysis sample columns show the percentages of youth in the final sample used to construct the various Δf 's. Data cleaned as described in section 5 and appendix E.

Table 4. Cross-Sectional Mean Gap k^*'s

NELS/ELS				
Subject	Year	Comparison	k^*	Crosses?
math	1990	black/white	0.33	Yes
math	2002	black/white	–	No
reading	1990	black/white	0.32	Yes
reading	2002	black/white	–	No
math z	1990	black/white	–	No
math z	2002	black/white	–	No
reading z	1990	black/white	–	No
reading z	2002	black/white	0.17	Yes
math	1990	income	–	No
math	2002	income	–	No
reading	1990	income	0.38	Yes
reading	2002	income	–	No
math z	1990	income	–	No
math z	2002	income	–	No
reading z	1990	income	0.36	Yes
reading z	2002	income	–	No
NLSY				
Subject	Year	Comparison	k^*	Crosses?
math	1979	black/white	–	No
math	1997	black/white	–	No
reading	1979	black/white	0.35	Yes
reading	1997	black/white	0.40	Yes
math	1979	income	0.11	Yes
math	1997	income	0.33	Yes
reading	1979	income	0.13	Yes
reading	1997	income	0.20	Yes

Note: k^*'s estimated using Δf's calculated on an evenly-spaced test-score grid of 5,000 points and k-grid of 1,000 points. Data cleaned as described in section 5 and appendix E. All observed cross-sectional gaps are positive.

Table 5. Mean Gap-Change k^*'s

Survey	Subject	Comparison	k^*	Crosses?	Observed Sign
NELS/ELS	math	black/white	0.29	Yes	pos
NELS/ELS	reading	black/white	0.28	Yes	pos
NELS/ELS	math z	black/white	0.08	Yes	pos
NELS/ELS	reading z	black/white	0.11	Yes	neg
NELS/ELS	math	income	0.08	Yes	neg
NELS/ELS	reading	income	0.04	Yes	pos
NELS/ELS	math z	income	0.14	Yes	neg
NELS/ELS	reading z	income	0.00	Yes	neg
NLSY79/97	math	black/white	0.11	Yes	neg
NLSY79/97	reading	black/white	0.12	Yes	neg
NLSY79/97	math	income	0.27	Yes	neg
NLSY79/97	reading	income	0.05	Yes	neg

Note: k^*'s estimated using Δf's calculated on an evenly-spaced test-score grid of 5,000 points and k-grid of 1,000 points. Data cleaned as described in section 5 and appendix E.

Table 6. NLSY Regression Results

Survey	Subject	Covariate	$\hat{\beta}$ or $\Delta\hat{\beta}$	t-stat	R^2
NLSY79/97	math	income	-0.01	-0.71	0.14
NLSY79/97	reading	income	-0.05	-2.62	0.14
NLSY79/97	AFQT	income	-0.04	-2.16	0.16
NLSY97	math	income	0.29	19.94	0.13
NLSY97	reading	income	0.26	18.43	0.11
NLSY97	AFQT	income	0.27	20.00	0.13
NLSY79	math	income	0.30	24.19	0.15
NLSY79	reading	income	0.31	25.14	0.16
NLSY79	AFQT	income	0.31	26.51	0.18

Note: Data cleaned as described in section 5 and appendix E.

Table 7. NLSY Income Regression Coefficient k^*'s

Survey	Subject	Coefficient	k^*	Crosses?	Observed Sign
NLSY79/97	math	income $\Delta\hat{\beta}$	0.04	Yes	neg
NLSY79/97	reading	income $\Delta\hat{\beta}$	0.33	Yes	neg
NLSY79/97	AFQT	income $\Delta\hat{\beta}$	0.19	Yes	neg
NLSY97	math	income $\hat{\beta}$	–	No	pos
NLSY97	reading	income $\hat{\beta}$	–	No	pos
NLSY97	AFQT	income $\hat{\beta}$	–	No	pos
NLSY79	math	income $\hat{\beta}$	–	No	pos
NLSY79	reading	income $\hat{\beta}$	–	No	pos
NLSY79	AFQT	income $\hat{\beta}$	–	No	pos

Note: k^*'s estimated on an evenly-spaced test-score grid of 1,000 points and k-grid of 1,000 points. Data cleaned as described in section 5 and appendix E.

Appendix C. Proofs

For notational simplicity, define $B^+(W,x,y) \equiv \int_x^y (W(s)-s)\Delta f(s)ds$ and $B^-(W,x,y) \equiv \int_x^y (s - W(s))\Delta f(s)ds$.

C.1. **Proofs of the Main Theorems.**

Proof. (theorem 4.2 and theorem 4.3) Let $\mathcal{W}_k^+$ denote the set of weighting functions satisfying (A2) and $D(\mathbb{I},W) \leq k$ that have the form given by the expression for W_0^+ in equation (4.1). Further, let $\mathcal{M}_k^+$ denote the set of weighting functions satisfying (A2) and $D \leq k$ that differ from any $W_0^+ \in \mathcal{W}_k^+$ on at least one interval with positive measure. Suppose $\exists \tilde{W}_0 \in \mathcal{M}_k^+$ such that $\mathcal{B}^+(\tilde{W}_0) > \mathcal{B}^+(W_0)$ for all $W_0 \in \mathcal{W}_k^+$. There are two cases to consider: N even and N odd. Suppose first that N is even. Let $\{\tilde{s}_2, \tilde{s}_4, \ldots, \tilde{s}_N\}$ be the points satisfying $\tilde{W}_0(s_i^*) = \tilde{s}_i$ for even values of i. Consider $W_0^+(s|k, \tilde{s}_2, \tilde{s}_4, \ldots, \tilde{s}_N) \equiv \tilde{W}_0^+$. I claim that $\mathcal{B}^-(\tilde{W}_0^+) > \mathcal{B}^-(\tilde{W}_0)$. To see that this inequality follows, suppose that $\tilde{W}_0$ deviates somewhere on $[s_{i-1}^*, s_{i+1}^*]$ for i even. Such a deviation implies that $\tilde{W}_0(s) \leq \tilde{W}_0^+(s)$ on $[s_{i-1}^*, s_i^*]$ and $\tilde{W}_0(s) \geq \tilde{W}_0^+(s)$ on $[s_i^*, s_{i+1}^*]$ with at least one of these inequalities strict. Therefore, $B^+(\tilde{W}_0, s_{i-1}^*, s_{i+1}^*) < B^+(\tilde{W}_0^+, s_{i-1}^*, s_{i+1}^*)$, which implies that $\tilde{W}_0^+$ dominates $\tilde{W}_0$ on any interval not $[0, s_1^*]$ such that $\tilde{W}_0$ does not correspond to some $W_0^+ \in \mathcal{W}_k^+$. To finish, consider $[0, s_1^*]$. Note that all $W_0^+ \in \mathcal{W}_k^+$ are identical on $[0, s_1^*]$, so if $\tilde{W}_0$ deviates on this interval it must be that $\tilde{W}_0(s) \neq \max\{s-k, 0\}$ on some $[s_L, s_H] \subseteq [0, s_1^*]$. Because all functions satisfying (A2) and $D(\mathbb{I},W) \leq k$ are bounded from below by the maximum of 0 and $s-k$, $\tilde{W}_0(s) > W_0^+(s)$ for any $W_0^+ \in \mathcal{W}_k^+$on $[\underline{s}, \bar{s}]$, which implies that $B^+(\tilde{W}_0, 0, s_1^*) < B^+(W_0^+, 0, s_1^*)$ for all $W_0^+ \in \mathcal{W}_k^+$, a contradiction. Now consider the case that N is odd and construct $\tilde{W}_0^+$ as before. The argument that $\tilde{W}_0^+$ dominates $\tilde{W}_0$ on $[0, s_{N-1}^*]$ is exactly analogous to the domination argument for N even on $[0,1]$. N being odd implies that $\Delta f > 0$ on $(s_N^*, 1)$. Note that all $W_0^+ \in \mathcal{W}_k^+$ are identical on $[s_N^*, 1]$, so if $\tilde{W}_0$ deviates on this interval it must be that $\tilde{W}_0(s) \neq \min\{s+k, 1\}$ on some $[s_L, s_H] \subseteq [s_N^*, 1]$. Because all functions satisfying (A2) and $D(\mathbb{I},W) \leq k$ are bounded by the minimum of 1 and $s+k$, $\tilde{W}_0(s) < W_0^+(s)$ for any $W_0^+ \in \mathcal{W}_k^+$ on $[s_L, s_H]$, which implies that $B^+(\tilde{W}_0, s_N^*, 1) < B^+(W_0^+, s_N^*, 1)$ for all $W_0^+ \in \mathcal{W}_k^+$, a contradiction. Let $\mathcal{W}_k^-$ denote the set of weighting functions satisfying (A2) and $D(\mathbb{I},W) \leq k$ that can be written as the expression for W_0^- in equation (4.1). Further, let $\mathcal{M}_k^-$ denote the set of weighting functions satisfying (A2) and $D \leq k$ that differ from any $W_0^- \in \mathcal{W}_k^-$ on at least one interval with positive measure. Suppose $\exists \tilde{W}_0 \in \mathcal{M}_k^+$ such that $\mathcal{B}^-(\tilde{W}_0) > \mathcal{B}^-(W_0^-)$ for all $W_0^- \in \mathcal{W}_k^-$. There are two cases to consider: N even and N odd. Suppose first that N is odd. Let $\{\tilde{s}_1, \tilde{s}_3, \ldots, \tilde{s}_N\}$ be the points satisfying $\tilde{W}_0(s_i^*) = \tilde{s}_i$ for i odd. Consider $W_0^-(s|k, \tilde{s}_1, \tilde{s}_3, \ldots, \tilde{s}_N) \equiv \tilde{W}_0^-$. I claim that $\mathcal{B}^-(\tilde{W}_0^-) > \mathcal{B}^-(\tilde{W}_0)$. To see this, suppose that $\tilde{W}_0$ deviates somewhere on $[s_{i-1}^*, s_{i+1}^*]$ for some odd i. This implies that $\tilde{W}_0(s) \geq \tilde{W}_0^-(s)$ on $[s_{i-1}^*, s_i^*]$ and $\tilde{W}_0(s) \leq \tilde{W}_0^-(s)$ on $[s_i^*, s_{i+1}^*]$ with at least one of these inequalities strict. Therefore, $B^-(\tilde{W}_0, s_{i-1}^*, s_{i+1}^*) < B^-(\tilde{W}_0^-, s_{i-1}^*, s_{i+1}^*)$, implying that $\tilde{W}_0^-$ dominates $\tilde{W}_0$ on any interval such that $\tilde{W}_0$ does not correspond to some $W \in \mathcal{W}_k^-$, a contradiction. Now consider the case that N is even and construct $\tilde{W}_0^-$ as before. The argument that $\tilde{W}_0^-$ dominates $\tilde{W}_0$ on $[0, s_{N-1}^*]$ is exactly analogous to the domination argument for N odd on $[0,1]$. N being even implies that $\Delta f < 0$ on $(s_N^*, 1)$. Note that all $W_0^- \in \mathcal{W}_k^-$ are identical on $[s_N^*, 1]$, so if $\tilde{W}_0$ deviates on this interval it must be that $\tilde{W}_0(s) \neq \min\{s+k, 1\}$ on some $[s_L, s_H] \subseteq [s_N^*, 1]$. Because all functions satisfying (A2) and $D(\mathbb{I}, W) \leq k$ are bounded by the minimum of 1 and $s+k$, $\tilde{W}_0(s) < W(s)$ for any $W_0^- \in \mathcal{W}_k^-$on $[s_L, s_H]$, which implies that $B^-(\tilde{W}_0, s_N^*, 1) < B^-(W_0^-, s_N^*, 1)$ for all $W_0^- \in \mathcal{W}_k^-$, a contradiction. □

Proof. (theorem 4.4) Consider $\frac{\partial \bar{\mathcal{B}}^+}{\partial k}$. Suppose that $k < \min\{s^*, 1-s^*\}$ so that W_0^+ has the form given in equation 4.1. In this case, $\bar{\mathcal{B}}^+$ may be written as $\bar{\mathcal{B}}^+ = -\int_0^k s\Delta f(s)ds - \int_k^{s^*} k\Delta f(s)ds + \int_{s^*}^{1-k} k\Delta f(s)ds + \int_{1-k}^1 (1-s)\Delta f(s)ds$. Differentiating each of these integrals with respect to k yields $\frac{\partial \bar{\mathcal{B}}^+}{\partial k} = \int_{s^*}^{1-k} \Delta f(s)ds - \int_k^{s^*} \Delta f(s)ds$. Now consider $\frac{\partial \bar{\mathcal{B}}^-}{\partial k}$ if $s_c > k$ and $s_c + k < 1$. In this case, $\bar{\mathcal{B}}^-$ may be written as $\bar{\mathcal{B}}^- = -\int_0^{s_c-k} k\Delta f(s)ds + \int_{s_c-k}^{s_c+k}(s - s_c)\Delta f(s)ds + \int_{s_c+k}^1 k\Delta f(s)ds$. Taking the derivative while noting that s_c depends on k yields $\frac{\partial \bar{\mathcal{B}}^-}{\partial k} = \int_{s_c+k}^1 \Delta f(s)ds - \int_0^{s_c-k} \Delta f(s)ds - \int_{s_c-k}^{s_c+k} \frac{\partial s_c}{\partial k} \Delta f(s)ds$. □

Proof. (theorem 4.6) Define $\cdot\beta^+ = \tau_x\left[cov(x, Y_0(s)) - cov(x, s)\right]$ where $\tau = 1/\sigma_x^2$ is the precision of x. Use that $cov(x,s) = \mathbb{E}[xs] - \mu_x\mu_s$ and $cov(x, Y_0) = \mathbb{E}[Y_0 x] - \mu_{W_0}\mu_x$ to write $\Delta\beta^+ = \tau_x\left[\mathbb{E}[xY_0] - \mathbb{E}[xs] - \mu_x(\mu_{Y_0} - \mu_s)\right]$. Next, observe that $\mathbb{E}[xY_0] - \mathbb{E}[xs] = \iint_0^1 x[Y_0(s) - s]g(x,s)dxds$ and $\mu_x(\mu_{Y_0} - \mu_s) = \mu_x \int_0^1 [Y_0(s) - s]f(s)ds$ by definition. Using that $g(x,s) = h(x|s)f(s)$, write $\iint_0^1 x[Y_0(s) - s]g(x,s)dxds = \int_0^1 [Y_0(s) - s]\left[\int_0^1 xh(x|s)dx\right] f(s)ds$. But $\int_0^1 xh(x|s)dx$ is just $\mu_{x|s}$, so the term $\mathbb{E}[xY_0] - \mathbb{E}[xs]$ may be written as $\int_0^1 \mu_{x|s}[Y_0(s) - s]f(s)ds$. Therefore, $\Delta\beta^+ = \tau_x \int_0^1 [Y_0(s) - s](\mu_{x|s} - \mu_x)f(s)ds$. Define $\Gamma(s) \equiv \tau_x(\mu_{x|s} - \mu_x)f(s)$. Since $f(s)\tau_x > 0$ everywhere on [0,1], $\Gamma(s)$ satisfies (A3). Therefore, $\Delta\beta^+ = \int_0^1 [Y_0(s) - s]\Gamma(s)ds$, rendering the bounding problem formally equivalent to that considered in theorem 4.2. Similarly, define $\cdot\beta^- = \tau_x\left[cov(x, s) - cov(x, Y_0(s))\right]$. An analogous argument shows that $\cdot\beta^- =$

$\int_0^1 [s - Y_0(s)]\Gamma(s)ds$, rendering the bounding problem again formally equivalent to that considered in theorem 4.2. □

Proof. (theorem 4.7) Standard arguments show that plim $\hat{\beta}_{IV} = \frac{cov(s,z)}{cov(x,z)}$. Define $\cdot\beta_{IV}^{+} = \sigma_{xz}^{-1}\left[cov(z, Y_0(s)) - cov(z, s)\right]$. By an argument exactly analogous to theorem 4.6, $\Delta\beta_{IV}^{+} = \sigma_{xz}^{-1}\int_0^1 [Y_0(s) - s]\left(\mu_{z|s} - \mu_z\right) f(s)ds$. Define $\Gamma_{IV}(s) \equiv \sigma_{xz}^{-1}\left(\mu_{z|s} - \mu_z\right) f(s)$. Since $f(s)\sigma_{xz}^{-1} > 0$ everywhere on [0,1], $\Gamma_{IV}(s)$ satisfies (A3), rendering the bounding problem formally equivalent to that considered in theorem 4.2. Similarly, define $\cdot\beta_{IV}^{-} = \sigma_{xz}^{-1}\left[cov(z, s) - cov(z, Y_0(s))\right]$. An analogous argument shows that $\cdot\beta_{IV}^{-} = \int_0^1 [s - Y_0(s)]\Gamma_{IV}(s)ds$, rendering the bounding problem again formally equivalent to that considered in theorem 4.2. □

Appendix D. Additional Exposition

D.1. **Regression Bias with a Binary Covariate.**

Achievement test scores are commonly used as outcome variables in regressions with binary predictor variables. For instance, value-added models of teacher quality are based on regressions of test scores (or test score changes) on teacher indicators and various control variables. The estimated coefficients on these indicators are then interpreted as measures of teacher quality. Such value-added estimates depend on the test scale used and thus may be sensitive to cardinal deviations. If the true scale of achievement is sufficiently different from the observed scale, then individual value-added estimates may have the wrong sign. Furthermore, rankings based on value-added estimates may also be erroneous; one may conclude that teacher A is superior to teacher B when in fact just the opposite is the case.

The analysis for mean differences presented in section 4.1 can be straightforwardly adapted to study bias in regressions using binary predictor variables. Consider the regression of s on some binary indicator D. The plim of the regression coefficient in this case is $\mathbb{E}[s|D = 1] - \mathbb{E}[s|D = 0]$. If instead we had regressed D on $W_0(s)$, we would have a plim of $\mathbb{E}[W_0(s)|D = 1] - \mathbb{E}[W_0(s)|D = 0]$. Let f_0 denote the distribution of s conditional on $D = 0$, and f_1 the distribution of s conditional on $D = 1$. The difference in these two plims can be written as

$$\int_0^1 (W_0(s) - s)\left[f_1(s) - f_0(s)\right] ds.$$

If $[f_1(s) - f_0(s)]$ satisfies (A3), this is exactly the objective function that yields $W_0^{+}(s|k)$ and $W_0^{-}(s|k)$ as worst-case weights. The analysis in this case is formally unchanged from that in section 4.1.

D.2. **Multidimensional Achievement Under Additive Separability.** The main body of this paper assumes that there is only one dimension of achievement. This assumption is unrealistic: a large and growing body of research suggests that there are multiple types of achievement relevant for labor market outcomes.[29] In this appendix, I extend the mean-difference bounding analysis to multiple dimensions in the special case that each dimension of achievement enters welfare additively separably. Unfortunately, analyzing bias due to scale misspecification for general welfare functions with multiple arguments is quite a hard problem mathematically. The simple techniques I introduce in this paper are not sufficient to characterize worst-case bias when welfare is multidimensional and not additively separable.

Suppose that achievement has two dimensions with ordinally perfect test scores s and q.[30] Let $W_0(s,q)$ denote the true cardinal value of the pair (s,q) and suppose that this value is known to have the form $W_0(s,q) = \psi_0(s) + \omega_0(q)$ for two increasing, unknown functions ψ and ω. Denote by F, F_s, and F_q the generic joint and marginal distributions of s and q, respectively. Additive separability in W_0 implies that value of the joint distribution F is also additively separable in the sense that $\mathbb{E}_F[W_0] = \mathbb{E}_{F_s}[\psi_0] + \mathbb{E}_{F_q}[\omega_0]$.[31] In turn, this implies that joint distribution F_A will be preferred to joint distribution F_B for all W_0 such that ψ and ω are increasing only if $F_{A,s} \succ F_{B,s}$ and $F_{A,q} \succ F_{B,q}$ both hold. The dependence between s and q does not affect welfare in this case. Therefore, the bias from using $\tilde{W}$ instead of W_0 is $\mathcal{B}(W_0, \tilde{W}, F_A, F_B) = |\mathcal{B}(\tilde{\psi}, \psi_0, F_{A,s}, F_{B,s}) + \mathcal{B}(\tilde{\omega}, \omega_0, F_{A,q}, F_{B,q})|$.

Additive separability does not quite imply that the bounding analysis can be carried out separately in each dimension. There are two subtleties that preclude one from treating each margin separately in constructing worst-case bounds. The first is that using the sup norm to define the distance restriction between W_0 and $\tilde{W}$ links the two dimensions of achievement because the magnitude and sign of the difference along one dimension determines the range of feasible differences along the other dimension.[32] I circumvent this difficulty by requiring that the distance restriction hold separately in each dimension.

[29] Kautz, Heckman, et al.[14] provides a good introduction and overview to this body of work.

[30] In empirical work, researchers typically assume that these dimensions are latent factors and that observed test scores depend on some combination of the underlying achievements. I abstract from these considerations here and simply suppose that we can craft tests which ordinally measure achievement along each relevant dimension.

[31] To see this, note that $V(W,F) = \iint_0^1 \psi(s) f(s,q) dq ds + \iint_0^1 \omega(q) f(s,q) dq ds$. But $\iint_0^1 \psi(s) f(s,q) dq ds = \int_0^1 \psi(s) f_s(s) ds = V(\psi, F_s)$ and $\iint_0^1 \omega(q) f(s,q) dq ds = \int_0^1 \omega(q) f_q(q) dq = V(\omega, F_q)$.

[32] To see this, consider the restriction $D(W_0, \tilde{W}) \leq k$ and suppose that $\sup_s[\psi_0(s) - \tilde{\psi}(s)] = \lambda k$ for some $\lambda \in (0,1)$. Then the maximum possible value of $\sup_q[\omega_0(q) - \tilde{\omega}(q)]$ is $(1-\lambda)k$, while the minimum possible value is $-(1+\lambda)k$.

Definition D.1. Suppose that $W(s,q) = \psi(s) + \omega(q)$ and $\tilde{W}(s,q) = \tilde{\psi}(s) + \tilde{\omega}(q)$. Then the pairwise distance between them is defined as

$$D_p(W, \tilde{W}) = \max\left\{ \sup_{s\in[0,1]} |\psi(s) - \tilde{\psi}(s)|, \sup_{q\in[0,1]} |\omega(q) - \tilde{\omega}(q)| \right\}.$$

Under D_p, the possible values of $\tilde{\omega}(q)$ given $D_p \leq k$ consist of the entire interval $[\omega(q) - k, \omega(q) + k]$, regardless of the shape of $\tilde{\psi}$ and ψ.[33]

There is one last wrinkle here compared to the one-dimensional bounds constructed in section 4.1. That analysis assumed that Δf is negative on some initial interval $(0, s_1^*)$ and has a finite number of 0's. Empirically, if $\Delta f > 0$ on $(0, s_1^*)$, I argued that one could simply switch the roles of A and B in the definition of Δf in order to maintain the assumption that $\Delta f < 0$ for test scores close to 0. If Δf_s and Δf_q both satisfy assumption (A3) or if $-\Delta f_s$ and $-\Delta f_q$ both do, then the analysis can proceed as before, separately for s and q. However, if Δf_s and $-\Delta f_q$ or $-\Delta f_s$ and Δf_s both satisfy (A3), there is no way to define A and B such that the analysis can go forward separately as before. Fortunately, in the single-dimensional case, $W_0^+(s|k, \Delta f) = W_0^-(s|k, -\Delta f)$ and $W_0^-(s|k, \Delta f) = W_0^+(s|k, -\Delta f)$. Since A and B may be interchanged freely, there are only two distinct situations: Δf_s and Δf_q both satisfy (A3) or only one of them does.

Theorem D.2. *Suppose that (A1) and (A2) hold. If Δf_s and Δf_q both satisfy (A3) for N_s and N_q, then $W_{0,s}^+$, $W_{0,q}^+$, $W_{0,s}^-$ and $W_{0,q}^-$ are all as in Theorem 4.2. If instead Δf_s and $-\Delta f_q$ satisfy (A3), then the worst-case weights for s are unchanged. In contrast, $W_{0,q}^+$ is given by the expression for W_0^- in Theorem 4.2 and $W_{0,q}^-$ is given by the expression for W_0^+.*

Theorem D.2, coupled with theorem 4.2, gives a general method for constructing worst-case weighting functions in the two dimensional case. This analysis can be extended easily to more than two achievement dimensions, provided that additive separability holds in each dimension.

[33] It is straightforward to verify that D_p is a distance measure. Separation, coincidence, and symmetry are all satisfied trivially. To see that the triangle inequality is satisfied, note that the sup norm is itself a distance measure and must satisfy the triangle inequality for each dimension s and q.

$$\begin{aligned} D(\psi, \tilde{\psi}) \leq D(\psi, \hat{\psi}) + D(\hat{\psi}, \tilde{\psi}) \quad &\wedge \quad D(\omega, \tilde{\omega}) \leq D(\omega, \hat{\omega}) + D(\hat{\omega}, \tilde{\omega}) \implies \\ \max\left\{D(\psi, \tilde{\psi}), D(\omega, \tilde{\omega})\right\} \quad &\leq \quad \max\left\{D(\psi, \hat{\psi}) + D(\hat{\psi}, \tilde{\psi}), D(\omega, \hat{\omega}) + D(\hat{\omega}, \tilde{\omega})\right\} \\ &\leq \quad \max\left\{D(\psi, \hat{\psi}), D(\omega, \hat{\omega})\right\} + \max\left\{D(\hat{\psi}, \tilde{\psi}), D(\hat{\omega}, \tilde{\omega})\right\} \end{aligned}$$

D.3. **Bounding Differences of Regression Coefficients.** This appendix extends theorem 4.6 to cover the sensitivity analysis for differences of regression coefficients. Consider the difference $\Delta\beta(s) \equiv \beta_A(s) - \beta_B(s)$, where A and B denote different samples on which the regression $s = \alpha + \beta x + \varepsilon$ is run. Let $\Delta\beta(Y_0)$ denote the corresponding difference when the weighted scores Y_0 are used in the regressions instead. The difference that we are trying to maximize for $\Delta\beta^+$ is $\Delta\beta(Y_0) - \Delta\beta(s)$, which is equal to $(\beta_A(Y_0) - \beta_A(s)) - (\beta_B(Y_0) - \beta_B(s))$. The proof of theorem 4.6 showed that $(\beta(Y_0) - \beta(s)) = \tau_x \int_0^1 [Y_0(s) - s](\mu_{x|s} - \mu_x) f(s) ds$, which implies that the bounding objective function here can be written as

$$\int_0^1 [Y_0(s) - s] \left(\tau_{x,A}(\mu_{x|s,A} - \mu_{x,A}) f_A(s) - \tau_{x,B}(\mu_{x|s,B} - \mu_{x,B}) f_B(s)\right) ds.$$

Therefore, if $\Gamma(s) \equiv \left(\tau_{x,A}(\mu_{x|s,A} - \mu_{x,A}) f_A(s) - \tau_{x,B}(\mu_{x|s,B} - \mu_{x,B}) f_B(s)\right)$ satisfies (A3), the bounding analysis will be formally equivalent to theorem 4.2. The analysis for $\Delta\beta^-$ simply replaces $[Y_0(s) - s]$ with $[s - Y_0(s)]$.

Appendix E. Data

The NELS first surveyed a nationally representative sample of eighth graders in the spring of 1988 with follow-up surveys in 1990, 1992, and 2002. I make use of the 1990 wave in order to keep the comparison groups consistent with my prior work on the income-achievement gap. The NELS wave consists mostly of 10th graders who were between the ages of 15 and 17 at the survey date. The ELS first surveyed a nationally representative sample of 10th graders in 2002, so all of my calculations compare this initial ELS wave to the first follow-up wave in the NELS.

Both the NELS and ELS contain data on household income, demographics, and achievement. Respondents in both surveys took comparable achievement tests in each survey wave. These tests covered similar content and followed a similar stratified design. Both assessments included some items in common, and both surveys report three parameter logistic item response theory (IRT) scores in the 1988 base-year scale estimated using these items. If the IRT model is correctly specified, these base-year scale scores should be ordinally comparable between the two surveys. That is, if student i has a higher score than student j, then student i should have higher underlying achievement regardless of whether i and j were drawn from the same or different surveys.

The initial waves of the NELS and ELS collected data on household income. Unfortunately, these data are categorical, significantly complicating the construction of directly comparable income groups from both surveys. For this paper, these details are relatively unimportant, and I simply use one plausible definition out of many for "high-income" and "low-income." I define high-income youth as those from the top 20% of the household income distribution and low-income youth as those from the bottom 20%. I approximate these quintiles by selecting the ranges of income buckets such that the masses of the high and low buckets are as close as possible to 0.2.[34] Unlike the NELS, the ELS imputes test scores, family income, and demographic variables. I drop imputed observations from the ELS sample. Nielsen[20] documents that the inclusion or exclusion of these observations has relatively little effect on the estimated achievement gap changes.

The NLSY79 and NLSY97 are high-quality, nationally representative surveys that contain ordinally comparable achievement data along with detailed student demographic information. Almost all respondents near the start of each survey took the Armed Services Vocational Aptitude Battery (ASVAB). Following an extensive literature in economics using these data, I study the math and reading subscores of the Armed Forces Qualifying Test (AFQT), which itself is a subset of the ASVAB.[35] The ASVAB test format changed from pencil-and-paper to a computer aided design between the NLSY79 and NLSY97. The military commissioned a study to determine how to compare scores from the new and old test formats. Segall[25] constructs a score crosswalk by equating percentiles on the two tests for a sample of military recruits who were randomly assigned to one version of the test or the other.[36] I use these crosswalked scores, as they should be ordinally comparable in the sense previously defined.

Both NLSY surveys collect extensive longitudinal data on each respondent's family, income, health, education, and employment history. I do not use the longitudinal component of these surveys here. I define high- and low-income respondents as those in the top and bottom

[34] For example, suppose there are 8 ordered income categories with equal numbers of respondents in each bucket. Then, the high-income group would simply be the top two income buckets (containing the top 25% of the sample) and the low-income groups would likewise be the bottom two buckets. In this case, both categories are somewhat larger than the target comparison groups.

[35] The ASVAB components feeding in to the AFQT changed in 1989. Throughout, I will use the current definition that sets the math subscore to be the sum of the arithmetic reasoning and math knowledge ASVAB component scores. The definition for reading did not change in 1989.

[36] The crosswalk is available courtesy of Altonji, Bhadarwaj, and Lange[2] and is available at the following url: `http://www.econ.yale.edu/~fl88/data.html`. The crosswalk contain percentile-mapped scores for each component score of the ASVAB. Simply adding these scores together is not strictly valid because it ignores the covariance of the different ASVAB components. Fortunately, Segall[26] reports that summing the crosswalked scores or crosswalking the summed scores leads to virtually identical results.

quintiles of the base-year household income distribution. The household income variable sums together all sources of income (wage, investment, business, etc.) for all household members. Since the youth I study are all younger than 18 years old, their total contribution to household income is typically negligible. Although I have not specifically assessed the robustness of my estimates to these data choices, I found in Nielsen[19] that ordinal income-achievement estimates using these data are not sensitive to plausible alternative income definitions.