Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Achievement Gap Estimates and Deviations from Cardinal Comparability

Eric R. Nielsen

2015-040

Please cite this paper as:
Nielsen, Eric R. (2015). "Achievement Gap Estimates and Deviations from Cardinal Comparability," Finance and Economics Discussion Series 2015-040. Washington: Board of Governors of the Federal Reserve System, http://dx.doi.org/10.17016/FEDS.2015.040.

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 fixed distance from each other. The worst-case weighting functions have simple, closed-form expressions consisting of achievement thresholds, flat regions in which test scores are uninformative, and regions in which the observed test scores are actually cardinally comparable. The paper next estimates these worst-case weighting functions for black/white and high-/low-income achievement gaps and gap changes using data from 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. Standard empirical methods may not robustly identify the sign of the trend in achievement inequality between students from different racial groups and income classes. Furthermore, ordinal methods may be more powerful and will continue to have the correct size when the test scale has been misspecified. 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 of the test scores renders them 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, ages, or time periods.[1] However, neither of these assumptions are well motivated by either economic or psychometric theory. If either fails, standard estimates

Date: May, 12 2015.

Preliminary and incomplete. Please do not cite or circulate without explicit permission of the author. Rick Ogden provided excellent research assistance for this project. 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 earns a 600 on the math section in 1980 should have the same achievement as a student who earns 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.

of achievement gaps and achievement-gap changes ("gaps/changes") can 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 between different groups of students using only the ordinal content of achievement test scores.[2] I also show that focusing on the cardinal/ordinal distinction is not mere methodological pedantry; standard, cardinal methods suggest that the gap in achievement between youth from high- and low-income households widened in recent decades, whereas more-robust ordinal methods strongly suggest the opposite.

The necessary conditions for ordinal statistics to unambiguously identify achievement gaps/changes are quite demanding. Two main conditions are needed, and each is likely to fail in many applied settings. 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] Although these dominance conditions are satisfied in some instances, for many economically interesting achievement comparisons they are not met.

The stringency of the necessary conditions for valid ordinal inference means that many achievement comparisons are inherently ambiguous or scale dependent. There are many situations in which we really cannot determine with certainty how achievement inequality has changed, as much as we would like to and as strongly as standard cardinal methods suggest that we can. Should researchers then simply plead ignorance when ordinal estimates are inconclusive or infeasible?

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 truly is closer to the 1510 student than she is to the 1000 student, even if the ratio of the score differences in the true cardinal scale is not exactly 1/50. Eschewing cardinality completely may be throwing away a lot of useful information and, thus, unnecessarily decreasing one's power to detect achievement differences. If some known test rescaling is truly cardinal, then cardinal statistical tests applied to this scale will have greater power to detect achievement-gap changes than will entirely ordinal approaches.[5]

[2]For an up-to-date draft of that paper, please see the top link at `https://sites.google.com/site/ericnielsenecon/research`.

[3]Many standardized tests are renormed every year, violating the common-scale assumption. If there are common items across the different tests, or if a group of students were randomly assigned to each different test, then it is possible to construct a common scale against which all test-takers from any survey can be coherently ordered. I abstract from this problem in the theory sections of this paper, and I take great care put scores in "equivalent units" in the empirical sections.

[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.

[5]Section 7 demonstrates the greater power of cardinal methods in the case that test scores are normally distributed and cardinally comparable.

Intuitively, if a known test scale is "almost" cardinal, cardinal statistical tests may correctly identify the sign of an achievement gap/change in the limit and have greater power than ordinal tests in finite samples. Of course, if the test scale used is actually very far from the true cardinal scale, then cardinal methods may misidentify achievement gaps/changes in the limit and will definitely have incorrect size and power in finite samples. In order to operationalize this intuitive tradeoff, it is necessary to formalize what it means for the true test scale and the observed scale to be "far" from each other.

Therefore, I study the failure of cardinal comparability as a specification problem. In particular, I introduce a distance measure that allows me to quantify how far apart two candidate test scales are. Next, I suppose that nothing is known about the true cardinal test scale other than that it lies within a fixed distance of the observed test scale. I then search for the unobserved true scale satisfying the hypothesized distance restriction that maximizes the difference between the observed and true achievement gaps/changes. By studying the worst-case bias as a function of the hypothesized distance between the true and observed scales, I can test the sensitivity of standard methods to deviations in the cardinality of test scales.

On the theoretical side, I derive closed-form expressions for the test scales that maximize positive and negative bias relative to the observed scale. Under fairly general conditions, these weighting functions depend only on the distance restriction imposed and a finite vector of statistics of the component test-score distributions being compared. The worst-case weighting functions are all piecewise-linear, with both 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 contain discontinuous jumps, or achievement thresholds, 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 scale sensitivities for black/white and high/low-income achievement gaps/changes in the National Longitudinal Surveys of Youth (NLSY) and the National Education Longitudinal Surveys (NELS/ELS). The cross-sectional achievement gap estimates are quite robust in these data. It is often not possible to find a rescaling of the test scores that flips the sign of a given estimate regardless of 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 observed scores. In contrast, gap-change estimates are typically much more sensitive to scale deviations. It is possible to pick a large enough distance restriction to flip the sign of the estimate for every gap-change estimate I examine. Furthermore, the size of the deviations 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 support, the sign of the income-achievement gap change for reading may be misidentified in the NELS/ELS.

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-gap changes estimated using very widely used data sources are not robust to minor changes in the test scale used. Since there are not good reasons to prefer the observed test scale to any other, estimates of changes in achievement inequality over time using this scale are not credible. Researchers assessing changes in achievement inequality over time should be much more circumspect in their deployment of standard cardinal methods or should eschew scale-dependent techniques entirely.

This paper is not entirely negative, because I also develop a set of tools that allow a researcher to assess whether a particular achievement gap/change 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 gap/change calculations. With my toolkit in hand, empirical researchers can proceed using standard methods and simply check whether or not their particular conclusions are overly scale dependent before switching to less familiar and less powerful ordinal approaches.

The rest of the paper is as follows. Section 2 reviews the relevant literature on achievement gaps, test score cardinality, and the relationship between stochastic dominance and social welfare. Section 3 lays out the notation, defines the necessary mathematical objects, and justifies the normalizations and simplifications I employ. Section 4 derives the worst-case weighting functions for a general class of achievement gap/change estimates. Section 5 outlines a number of empirically relevant extensions to the theoretical bounding analysis. Section 6 assesses the sensitivity of a number of achievement gap/change estimates to cardinal deviations using the NLSY and NELS/ELS data. Section 7 investigates the power and size of cardinal and ordinal tests in the presence of cardinal deviations. Section 8 concludes. Appendices A through D contain figures, point estimates, additional background, and technical discussion.

2. Literature Review

The economics literature using cardinal methods to assess group differences in achievement is vast. Fryer and Levitt[7, 8], Clotfelter, Ladd, and Vigdor[5], Duncan and Magnuson[6], Hanushek and Rivkin[9], and Neal[15], among many others, use cardinal methods to assess changes in black/white achievement inequality in the United States.[6] Reardon[19] employs cardinal methods to argue that

[6]Neal[15] does recognize, however, that "[a]chievement has not natural units," and so he also analyzes the percentile rankings of black versus white test takers.

the gap in achievement between high- and low-income youth has widened tremendously over the past several decades. Finally, research assessing school and teacher 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.[7]

This paper is not the first in either economics or psychometrics to argue that normalized test scores are not cardinally comparable. In psychometrics, Stevens[23] and Lord[14] argue that most psychometric test scores are inherently ordinal. In economics, Lang[13], Bond and Lang[3], Cascio and Staiger[4], Reardon[18], and Nielsen[16] all discuss the sensitivity of standard achievement gap/change estimates to order-preserving transformations of the test scores. The analysis in Bond and Lang[3] 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. Their worst-case transformations typically consist of a set of achievement thresholds with mostly flat regions between sharp jumps. Interestingly, their functional forms are quite similar to those I derive theoretically in this paper.

Ultimately, economists and policymakers are not interested in the test scores themselves, but rather 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, it has been shown 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 second-order dominance (SOSD) identically.[8] Aaberge, Havnes, and Mogstad[20] note that first- and second-order dominance often fail to hold in empirical applications ranking income distributions. In response, they derive economically interpretable preference functions that allow unambiguous ranking of distribution functions under dominance of any order. In principle, their approach could also be used to rank test-score distributions when FOSD and SOSD fail to hold. However, doing so would require imposing conditions on the social welfare function that are less plausible when applied to test scores than when applied to income. For example, concavity can be justified for income by appealing to diminishing marginal utility. However, concavity may not make sense for test scores because the relationship between scores and life outcomes may be

[7]For example, Krueger[12] and Hoxby[11] both use test scores cardinally to estimate the effect of class size on student achievement gains. Value-added methodologies such as those expounded in Raudenbush [17] and elsewhere also suppose that (normalized) test scores are cardinally comparable.
[8]Indeed, in this paper, and in Nielsen[16], I make extensive use of this fact.

quite convex. Even if the social welfare function is concave in life outcomes, it may not be concave in test scores.[9]

3. Formal Setting and Assumptions

Suppose a population of students have test scores s distributed according to cumulative density function (cdf) F. Furthermore, suppose that the test scores are weakly ordinally perfect in the sense that true achievement a corresponding to test score s is given by $a = \psi(s)$ for some weakly increasing function ψ.[10]

Let $W_0(s)$ be the true value of the underlying achievement corresponding to test score s. W_0 is the composition of several conceptually distinct maps: the map ψ from test scores to true achievement, the map from true achievement to economically relevant life outcomes, and the map from life outcomes to social welfare. Even assuming that the choice of the social welfare function is uncontroversial, the first of these maps is not knowable and the second is very difficult to estimate even with the richest data.[11] Therefore, I will assume throughout this paper that W_0 is inaccessible to the researcher.

The only a priori restriction I place on W_0 is that it is weakly increasing in s: $s > s' \implies W_0(s) \geq W_0(s') \ \wedge W(s) > W_0(s') \implies s > s'$. Weak monotonicity is a natural assumption in this setting because higher test scores must correspond to weakly higher underlying achievement, and positive life outcomes should be causally linked to higher true achievement. I do not assume W_0 is strictly monotone because I want to allow for the possibility that changes in test scores in some regions do not change overall welfare, either because the scores themselves are uninformative or because higher achievement does not always lead to better outcomes.[12] Even if the map from test scores to achievement is 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. The worst-case W_0's may actually be constant when the true scale is allowed to be very different than the observed scale. However, the worst-case weighting functions will be strictly increasing somewhere in all but the most extreme cases. Unless I explicitly specify otherwise, I will

[9]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.

[10]This implies that for two students i and j with test scores $s_i > s_j$, a_i should be weakly greater than a_j. Whether ψ is weakly or strictly monotone is not crucial for the analysis. The advantage of maintaining only weak monotonicity conceptually is that is allows test scores to be uninformative in some regions. Of course, ψ must be strictly increasing somewhere if the test is to be useful at all in differentiating students by achievement.

[11]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[16] carries out such a calculation for lifetime earnings in the National Longitudinal Surveys of Youth (NLSY) data.

[12]Real-world institutions often treat test scores in some ranges as being uninformative; for example, graduate economics departments typically do not distinguish between GRE scores of 165-170.

treat generic weighting functions W_0 in the remaining analysis as having at least two values $s > \tilde{s}$ such that $W_0(s) > W_0(\tilde{s})$.

Consider the problem of comparing two distinct test-score distributions F and $\tilde{F}$ given that W_0 is unknown. The total value of F depends on W_0 because $V(W_0, F) = \mathbb{E}_F[W_0(s)] = \int W_0(s)dF(s)$. It is straightforward to show that $V(W_0, F) > V(W_0, \tilde{F})$ for any increasing W_0 if and only if $F \succ \tilde{F}$, where $\succ$ denotes strict FOSD. If FOSD does not hold, there is no unambiguous way to compare F and $\tilde{F}$ in that there must exist distinct increasing functions W and $\tilde{W}$ such that $V(W, F) > V(W, \tilde{F})$ and $V(\tilde{W}, F) < V(\tilde{W}, \tilde{F})$. In contrast, misspecifications of W_0 cannot lead to erroneous conclusions about the sign of the achievement gap if $F \succ \tilde{F}$, although the relative magnitudes of the true and observed achievement gaps may be very different.

I make a number of technical 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 and permit much cleaner statements and proofs of the main results.

Definition 3.1. F satisfies (A1) iff:

(i) $F \in \mathcal{F}$, the space 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 and does not rule out any interesting cases. Part (ii) is just a normalization and is also without loss of generality since test scores can always be rescaled to fit in [0,1] from whatever cardinal scale the researcher prefers.[13]

Definition 3.2. W_0 satisfies (A2) iff:

(i) W_0 is integrable with respect to any F satisfying (A1).

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

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

Part (i) of definition 3.2 is again a technical assumption and does not rule out any interesting cases. The weakly increasing assumption in part (ii) was justified previously. The requirement in part (ii) that W_0 be right-continuous is another technical assumption that guarantees uniqueness of the "worst-case" weighting functions.[14] Part (iii) normalizes $W_0(s)$ to have the same support as F. This normalization

[13]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 $\tilde{F}$.

[14]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) jumps.

is without loss of generality because welfare is bounded and can only ever be identified up to affine transformations. 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. For the remainder of the paper, I will always suppose that (A1) and (A2) hold. Figure A.1 plots several possible W_0's when (A2) hold. The figure shows that W_0 may be convex, concave, linear, and discontinuous while still satisfying (A2).

In order to assess how sensitive a given achievement gap/change estimation method is to scale deviations, I must first define a distance measure on test scales. Given two candidate test scales, I define the distance between them using the sup norm.

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

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

D is a well-defined distance function on the space of weakly increasing functions with domain and range on [0,1].[15]

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 on [0,1] do W and $\tilde{W}$ differ by very much. In contrast, when k is large, there are regions where W and $\tilde{W}$ weigh scores very differently. Definition 3.3 is not the only way to formalize the notion of distance between weighting functions. For instance, one could define $\mathcal{D}(W, \tilde{W}) \equiv \int |W(x) - \tilde{W}(x)| dx$. This alternative definition has the advantage that it will assess a large difference in the case that W and $\tilde{W}$ differ by a small amount everywhere on [0,1]. Using $\mathcal{D}$ instead of D substantially complicates the analysis and is therefore left for future work.

Consider measuring the cross-sectional achievement gap between two groups of students as well as the changes in the cross-sectional gap over time. Labeling the groups A and B, and 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

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

Similarly, the change in the achievement gap between A and B from t to $t+1$ is[16]

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

[15] That is, for any three such functions W, X, and Y, the following hold: (i) $D(W, X) \geq 0$, (ii) $D(W, X) = 0$ if and only $W = X$, (iii) $D(W, X) = D(X, W)$, and (iv) $D(X, W) \leq D(X, Y) + D(Y, W)$.

[16] I will exclusively use language describing gap-changes 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.

In both of these cases, the object of interest consists of an integral from 0 to 1 of the function $W_0(s)\Delta f(s)$, where Δf is some sum and difference of density functions across the relevant comparison groups. The specific context 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 exact objective (cross-sectional or gap-change) in the background.

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 given by $\Delta V(\mathbb{I}, \Delta f) = \int_0^1 s\Delta f(s)ds$. The bias created from using $\mathbb{I}$ instead of W_0 is just the difference between these two ΔV's. There are two cases to consider, one that maximizes the degree to which the true difference is larger than the observed difference, one that maximizes the degree to which the observed difference overestimates the true difference.

$$\mathcal{B}^+(\mathbb{I}, W_0, \Delta f) = \int_0^1 (W_0(s) - s)\,\Delta f(s)ds \tag{3.1}$$

$$\mathcal{B}^-(\mathbb{I}, W_0, \Delta f) = \int_0^1 (s - W_0(s))\,\Delta f(s)ds. \tag{3.2}$$

$\mathcal{B}^+$ will be large when $\Delta f(s)$ and $(W_0(s) - s)$ have the same sign, while $\mathcal{B}^-$ will be large when the opposite is true. 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 3.4. Suppose that all component test-score distributions in Δf satisfy (A1). The worst-case W_0's satisfying (A2) and $D(\mathbb{I}, W) \leq k$ for a given distance restriction k are then given 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 $\bar{\mathcal{B}}^+(k) = \mathcal{B}^+(\mathbb{I}, W_0^+(s|k, \Delta f), \Delta f)$ and $\bar{\mathcal{B}}^-(k) = \mathcal{B}^-(\mathbb{I}, W_0^-(s|k, \Delta f), \Delta f)$ denote the values of the worst-case biases given k.

Although W_0^+ and W_0^- both depend on k and Δf, I will often omit these arguments for brevity when their specific identities are not important. Unless certain symmetry conditions hold on Δf, $\bar{\mathcal{B}}^+(k) \neq \bar{\mathcal{B}}^-(k)$ for most values of k greater than 0. Both biases are 0 when $k = 1$ as W_0^+ and W_0^- are both identically equal to $\mathbb{I}$ in this case.

It is very difficult to make precise statements about these biases if the various component test-score densities are unrestricted other than the conditions imposed by (A1). Therefore, I will consider a number of special cases that encompass many realistic empirical scenarios.

Definition 3.5. Δf satisfies (A3) iff all of its component densities satisfy (A1) and if $\exists! s^* \in (0,1)$ such that $\Delta f(s^*) = 0$, $\Delta f(s) < 0, \forall s \in (0, s^*)$ and $\Delta f(s) > 0, \forall s \in (s^*, 1)$.

Assumption (A3) simply says that Δf is negative for low values of s, positive for high values of s, and crosses 0 only once on (0,1).[17] Although (A3) might appear to be very narrow, it actually encompasses a number of empirically relevant cases. For example, suppose that $\Delta f(s) = f_{A,t}(s) - f_{B,t}(s)$. If the raw distributions of A and B are both unimodal and symmetric with similar variances and if A has a higher mean than B, Δf will typically satisfy (A3) after normalization.[18] Whenever $F_A \succ F_B$, $\Delta f(s) = f_A(s) - f_B(s)$ will satisfy (A3). The reverse implication ((A3) implying FOSD) does not generally hold. However, even in cases where FOSD does not hold, achievement gap/change estimates will typically be quite robust under (A3).

In many interesting applications, particularly those involving achievement gap changes, Δf will cross 0 more than once on (0,1). Definition 3.6 extends definition 3.5 to allow for multiple crossing points.

Definition 3.6. Δf satisfies (A4) for $N > 1$ if the following conditions hold:

(i) $\exists s_0^* s_1^*, s_2^*, \ldots, s_N^*, s_{N+1}^*$ with $s_0^* \equiv 0 < s_1^* < s_2^* < \ldots < 1 \equiv s_{N+1}^*$ such that $\Delta f(s_i^*) = 0 \forall i \in \{1, \ldots, N\}$.

(ii) $\Delta f(s) \neq 0$ if $s \notin \{0, s_1^*, \ldots, s_N^*, 1\}$

(iii) $\Delta f(s) < 0$ for $s \in (0, s_1^*)$ and $\text{sign}[\Delta f(s)] = -\text{sign}[\Delta f(s')]$ whenever $s \in (s_{i-1}^*, s_i^*)$ and $s' \in (s_i^*, s_{i+1}^*),\ i \in \{1, \ldots, N\}$.

Definition 3.6 says that there are exactly N points on (0,1) where Δf is 0 and that at none of these points does $\frac{d\Delta f(s)}{ds}$ equal 0. Furthermore, the definition says that Δf is negative before the first interior 0. This means that if N is odd $\Delta f(s) > 0$ on $(s_N^*, 1)$ and if N is even $\Delta f(s)$ on this interval. Figure A.2 in appendix A displays three Δf's consistent with (A3) and two Δf's consistent with (A4) for $N = 6$.

Assumption (A4) defines a very general class of functions. Since (iii) can always be guaranteed by choosing which distributions to label A and which to label B, the only substantive restrictions placed on Δf by (A4) are that it only cross 0 a finite number of times, that there be no intervals with positive measure on which Δf is 0, and that there be no 0's at which $\frac{d\Delta f(s)}{ds} = 0$. (A4) will be satisfied generically for virtually any finite sum or difference of densities from any commonly used distributional families.

[17] Note that (A3) only restricts $\Delta f(0)$ to be less than or equal to 0 and $\Delta f(1)$ to be greater than or equal to 0.

[18] For example, if $A \sim N(\mu_A, \sigma)$ and $B \sim N(\mu_B, \sigma)$, then Δf will satisfy (A3) once the normalizations in (A1) are imposed.

4. Bounding Analysis Using the Sup Norm

I now construct closed-form expressions for W_0^+ and W_0^- when either (A3) or (A4) hold. Under either assumption, both W_0^+ and W_0^- have relatively simple functional forms for any value of $k \in [0,1]$. Unfortunately, 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 simulating $\bar{\mathcal{B}}^+(k)$ and $\bar{\mathcal{B}}^-(k)$ relatively straightforward.

The worst-case weighting functions under (A4) nest the worst-case functions under (A3) as special cases. Even though doing so is technically redundant, I will present results for (A3) separately, as W_0^+ and W_0^- have particularly simple and intuitive interpretations in this case. Therefore, suppose first that (A1)-(A3) hold. Theorem 4.1 below shows that the only influence that Δf has on W_0^+ is through s^*.

Theorem 4.1. *If (A1)-(A3) hold, then W_0^+ has the form*[19]

$$W_0^+(s|k,s^*) = \begin{cases} \max\{s-k,0\}, & s \in [0,s^*) \\ \min\{s+k,1\}, & s \in [s^*,1] \end{cases} \tag{4.1}$$

Proof. In appendix C. □

Although equation (4.1) in theorem 4.1 is somewhat difficult to parse, the intuition behind it is quite simple. Recall that $\mathcal{B}^+$ is large when $[W_0(s) - s]$ and $\Delta f(s)$ have the same sign, implying 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^* and as far above the diagonal when s is greater than s^*. 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 $[0,s^*)$, while the farthest possible value above is $\min\{s+k,1\}$, which defines W_0^+ on $[s^*,1]$.

In order to understand the definition in more detail, it is helpful to examine a number of cases determined by the size of s^*, $1-s^*$, and k. If $k \geq \max\{s^*, 1-s^*\}$, then the $D \leq k$ restriction never binds and W_0^+ is just a step function given by $W_0^+(s|k,s^*) = 0$ for $s < s^*$ and $W_0^+(s|k,s^*) = 1$ for $s \geq s^*$. If $k < \min\{s^*, 1-s^*\}$, the constraint that $D \leq k$ binds on both intervals $[0,s^*)$ and $[s^*,1]$ and

[19] I will always include s^* in the "upper half" of W_0^+ or W_0^-. This choice is arbitrary and unimportant since s^* has 0 measure. Therefore, $W_0^+(s|k,s^*)$ below could just as well be defined by $\max\{s-k,0\}$ on $[0,s^*]$ and $\min\{s+k,1\}$ on $(s^*,1]$.

$W_0^+(s|k,s^*)$ becomes

$$W_0^+(s|k,s^*) = \begin{cases} 0, & s \leq k \\ s-k, & s \in (k,s^*) \\ s+k, & s \in [s^*, 1-k) \\ 1, & s \geq 1-k. \end{cases} \tag{4.2}$$

Figure A.3 in appendix A plots equation (4.2).

The analysis for W_0^- under (A1)-(A3) is substantially more involved than the analysis for W_0^+. 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 $[0, s^*)$ and as far below the diagonal as possible on $[s^*, 1]$. But W_0^- must be weakly increasing, so the larger $W_0^-(s^*)$ is, the smaller the possible bias contribution is on $[s^*, 1]$. W_0^- must trade off these competing forces.

Equation (4.3) in theorem 4.2 defines $W_0^-(s|k,s^*,s_c)$, where $s_c = W_0^-(s^*)$. The functional form of W_0^- is straightforward to derive given s_c. W_0^- must be as far above the diagonal as possible on $[0, s^*)$ consistent with both $D(\mathbb{I}, W_0^-) \leq k$ and $W_0^-(s^*) = s_c$, while W_0^- must be as far below the diagonal for values of s greater than s^*. Each potential choice of s_c trades off bias creation below and above s^*differently. Since (A1)-(A3) imply that this tradeoff is a smooth function of s_c, there must be some value of s_c that maximizes $\bar{\mathcal{B}}^-(k)$.

Theorem 4.2. *If (A1)-(A3) hold, then for some* $s_c \in [\max\{s^* - k, 0\}, \min\{s^* + k, 1\}]$, W_0^- *is given by*

$$W_0^-(s|k,s^*,s_c) = \begin{cases} \min\{s+k, s_c\}, & s \in [0, s^*) \\ \max\{s-k, s_c\}, & s \in [s^*, 1]. \end{cases} \tag{4.3}$$

Proof. In appendix C. □

Equation (4.3) is much easier to understand if one considers several special cases. For example, if $s_c > k$ and $s_c + k < 1$, then equation (4.3) simplifies to

$$W_0^-(s|k,s^*,s_c) = \begin{cases} s+k, & s < s_c - k \\ s_c, & s \in [s_c - k, s_c + k] \\ s-k, & s > s_c - k. \end{cases} \tag{4.4}$$

Equation (4.3) in some sense defines the most general form for W_0^-. The other possible forms are just those cases where one or both of the kink points $s_c - k$ and $s_c + k$ lie on the boundary of $[0, 1]$. If k is

large enough that $s_c - k \leq 0$ and $s_c + k \geq 1$, W_0^- will simply equal s_c everywhere on [0,1]. If $s_c - k > 0$ and $s_c + k \geq 1$ then only the lower kink point is still present. Similarly, if $s_c - k \leq 0$ and $s_c + k < 1$ then only the upper kink point is present. [20] Figure A.4 illustrates these possibilities by plotting W_0^- for three different values of s_c.

Theorem 4.2 does not fully characterize W_0^- because it does not pin down s_c. Since s_c and k jointly determine the form of W_0^-, for a fixed k, s_c indexes all of the possible W_0^-'s consistent with $D(\mathbb{I}, W_0^{-1}) \leq k$. Each candidate s_c yields a different negative bias $\mathcal{B}^-(\mathbb{I}, W_0^-(\centerdot|s_c), \Delta f)$ and the worst-case s_c is just the point in $[s^* - k, s^* + k]$ that maximizes $\mathcal{B}^-(\mathbb{I}, W_0^-(\centerdot|s_c), \Delta f)$. In practice, calculating this worst-case s_c explicitly is very tedious and fairly uninformative.[21] One exception is the special case that $\Delta f(s^* - x) = -\Delta f(s^* + x),\ \forall x \in [0, \frac{1}{2}]$, which implies $s^* = s_c = 0.5$.

Both W_0^+ and W_0^- under (A1)-(A3) 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 achievement gap between A and B. The maximum possible true gap between A and B is given by $\Delta V(W_0^+(s|k, s^*), \Delta f)$. 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. 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, again, is just what W_0^- does.

The robustness of a cardinal gap/change estimate to deviations in scale depends on how rapidly the associated biases $\mathcal{B}^+$ and $\mathcal{B}^-$ increase as k increases. If these biases increase rapidly with k, then relatively small cardinal deviations may be sufficient to flip the sign of the gap/change estimate. In contrast, if they increase slowly, such a reversal will only be possible when k is quite large. In general, it is not possible to derive closed-form expressions for $\frac{\partial \mathcal{B}^+}{\partial k}$ and $\frac{\partial \mathcal{B}^-}{\partial k}$ because these derivatives depend on the particular shape of Δf. Nonetheless, in the case that Δf satisfies (A3) or (A4), it is still possible to gain some intuition about what features of Δf determine how quickly the positive-side and negative-side biases increase with increases in k. I only present the analysis for the case that (A3) holds; the results are qualitatively similar under (A4), but the exposition is messier and less intuitive.

[20] That is, if $s_c - k > 0$ and $s_c + k \geq 1$, equation (4.3) becomes

$$W_0^-(s) = \begin{cases} s + k, & s < s_c - k \\ s_c, & s \in [s_c - k, 1]. \end{cases}$$

If $s_c - k \leq 0$ and $s_c + k < 1$ then equation (4.3) simplifies to

$$W_0^-(s) = \begin{cases} s_c, & s < s_c + k \\ s - k, & s \geq s_c + k. \end{cases}$$

[21] The difficulty is that the integral $\int_0^1 (s - W_0^-(s|s_c))\Delta f(s) ds$ does not generally have a closed form.

Theorem 4.3. *If (A1)-(A3) hold and k is sufficiently close to 0, then*

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

Proof. In appendix C. □

Theorem 4.3 characterizes $\frac{\partial \mathcal{B}^+}{\partial k}$ and $\frac{\partial \mathcal{B}^-}{\partial k}$ for values of k relatively close to 0.[22] The theorem shows that $\frac{\partial \mathcal{B}^+}{\partial k}$ depends on the total area (both positive and negative) between Δf and 0 on the interval $[k, 1-k]$. If Δf is mostly far away from 0 in this central subinterval, then the positive-side bias will increase rapidly with k. Furthermore, $\frac{\partial \mathcal{B}^+}{\partial k}$ is monotonically decreasing in k and approaches 0 from above as k approaches 0.5. The expression for $\frac{\partial \mathcal{B}^-}{\partial k}$ is somewhat harder to interpret because s_c is only defined implicitly. For simplicity, suppose that Δf satisfies $\Delta f(0.5 - x) = -\Delta f(0.5 + x)$ 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 $\frac{\partial \mathcal{B}^-}{\partial k}$ depends only on the total area (positive and negative) between 0 and Δf on the intervals $[0, 0.5-k]$ and $[0.5+k, 1]$, that is, on the "tails" of [0,1]. $\mathcal{B}^-$ will generally be more sensitive than $\mathcal{B}^+$ to the properties of Δf near the endpoints of [0,1], even in the typical case that s_c depends on k. Theorem 4.3 also implies that $\frac{\partial \mathcal{B}^+}{\partial k}|_{k=0} = \frac{\partial \mathcal{B}^-}{\partial k}|_{k=0} = \int_0^1 |\Delta f(s)| ds$. For values of k very close to 0, $\mathcal{B}^+$ and $\mathcal{B}^-$ increase mostly symmetrically with k. As k grows larger, the relevant subintervals of [0,1] contributing the most to $\mathcal{B}^+$ and $\mathcal{B}^-$ become more and more different. This divergence, coupled with possible increases or decreases in s_c as k grows larger, means that $\frac{\partial \mathcal{B}^+}{\partial k}$ and $\frac{\partial \mathcal{B}^-}{\partial k}$ will not be equal generically when k is greater than 0.

I now relax the single-crossing assumption in favor of (A4). 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 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 $[s_{i+1}^*, s_{i+2}^*]$. For example, consider W_0^+ in the case that $N = 2$. Since $\mathcal{B}^+$ is large when $[W_0^-(s) - s]$ and Δf have the same sign, the contribution to $\mathcal{B}^+$ on $[s_1^*, s_2^*]$ is maximized when $W_0^+(s) = s + k$. However, $W_0^+(s)$ on $[s_2^*, 1]$ cannot be less than $W_0^+(s_2^*)$, but bias in this region is made larger the more negative $W_0^{-1}(s) - s$ is. Therefore, maximizing the bias contribution on $[s_1^*, s_2^*]$ minimizes the bias contribution on $[s_2^*, 1]$. Finding W_0^+ requires that one balance these competing forces, and the strength of these forces depends solely on the particular shape of Δf and the value of k.

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

The functional forms of W_0^+ and W_0^- under assumption (A4) also depend on whether N is even or odd. As with W_0^- under (A3), both W_0^+ and W_0^- are parametrized by the values they take at the various Δf crossing points. 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 the odd-indexed crossing points. Theorem 4.4 below characterizes W_0^+ when (A4) holds for an arbitrary N. Figure A.5 plots potential worst-case weighting functions W_0^+ for the cases $N = 2$ and $N = 3$.

Theorem 4.4. *If (A1), (A2), and (A4) hold for $N \in \mathbb{N}$, then there exists a non-decreasing sequence $0 \leq s_2^+ \leq s_4^+ \leq \ldots \leq 1$ such that $W_0^+(s_i^*|k) = s_i^+ \in [\max\{s_i^* - k, 0\}, \min\{s_i^* + k, 1\}]$ for even $i \leq N$ and such that*

$$W_0^+(s|k) = \begin{cases} \max\{s-k, 0\}, & s \leq s_1^* \\ \min\{s+k, s_2^+\}, & s \in (s_1^*, s_2^*] \\ \max\{s-k, s_2^+\}, & s \in (s_2^*, s_3^*] \\ \vdots & \\ \max\{s-k, s_N^+\}, & s \in (s_N^*, 1] \wedge N \text{ even} \\ \min\{s+k, 1\}, & s \in (s_N^*, 1] \wedge N \text{ odd.} \end{cases} \tag{4.5}$$

Proof. In appendix C. □

Theorem 4.5 below characterizes W_0^- for an arbitrary N. Figure A.6 plots potential worst-case weighting functions W_0^- when $N = 2$ or $N = 3$.

Theorem 4.5. *If (A1), (A2), and (A4) hold for $N \in \mathbb{N}$, then there exists a non-decreasing sequence $0 \leq s_1^- \leq s_3^- \leq \cdots \leq 1$ such that $W_0^-(s_i^*|k) = s_i^- \in [\max\{s_i^* - k, 0\}, \min\{s_i^* + k, 1\}]$ for odd $i \leq N$ and such that*

$$W_0^-(s|k) = \begin{cases} \min\{s+k, s_1^-\}, & s \leq s_1^* \\ \max\{s-k, s_1^-\}, & s \in (s_1^*, s_2^*] \\ \min\{s+k, s_3^-\}, & s \in (s_2^*, s_3^*] \\ \vdots & \\ \min\{s+k, 1\}, & s \in (s_N^*, 1] \wedge N \text{ even} \\ \max\{s-k, s_N^-\}, & s \in (s_N^*, 1] \wedge N \text{ odd.} \end{cases} \tag{4.6}$$

Proof. In appendix C. □

Theorems 4.1 through 4.5 show that bias is maximized when the true weighting function has achievement thresholds, flat regions, cardinal regions, and kinks. Both W_0^+ and W_0^- generically consist of regions where increases in scores are not valuable, regions where the true value increases 1-1 with observed test scores, and discontinuous achievement thresholds where the true value jumps up by a large amount. Although these worst-case W_0's may look extreme compared with most test scales, they are not economically implausible. For example, consider a test score equal to the share of the Russian Cyrillic alphabet that a student knows. This test scale is interval in the sense that each score increment of $\frac{1}{33}$ corresponds to a new, identifiable skill: knowing a letter of the alphabet. However, a plausible economic weighting should be mostly flat for scores between 0 and $\frac{32}{33}$ and display a sizable jump up between $\frac{32}{33}$ and 1 because knowing the whole 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 0 while the productivity of a worker possessing all of the requisite skills is quite high. Finally, selective institutions may employ admissions thresholds, again creating discontinuities and kinks in the economically-relevant score weighting function.

5. Extensions

The approach presented in section 4 is substantially more general than it might at first appear. In particular, similar methods can be applied to bound both the bias in regressions using test scores as outcome variables and the bias in mean difference calculations when achievement has multiple dimensions. A complete, formal analysis of these extensions is beyond the scope of the present paper. In this section, I sketch out results for two special cases. First, I demonstrate that theorems 4.4 and 4.6 can be straightforwardly applied to bound the bias in regression coefficients of test scores on binary predictor variables. Second, I show that these same theorems can be used to bound mean differences when there are multiple dimensions of achievement that enter W_0 additively separably.[23]

Consider the ordinary least squares (OLS) regression of s on some binary indicator D. The goal is to characterize the worst-case bias in the resulting regression coefficient on D due to misspecification in the scale of s. The probability limit (plim) of the OLS estimator in this baseline regression is $\beta(\mathbb{I}) = \mathbb{E}[s|D=1] - \mathbb{E}[s|D=0]$. If instead we had regressed D on $W_0(s)$, the plim of the resulting regression coefficient would be $\beta(W_0) = \mathbb{E}[W_0(s)|D=1] - \mathbb{E}[W_0(s)|D=0]$. The difference in these

[23]The techniques from section 4 can also be adapted to study regressions of test scores on continuously distributed covariates. A rigorous analysis of this extension is the subject of ongoing work that should appear as a separate working paper in the coming months. In contrast, analyzing multiple dimensions of achievement when W_0 is not additively separable presents substantial technical problems and is an area of active research.

plims is

$$\Delta\beta \equiv (\mathbb{E}[W_0(s)|D=1] - \mathbb{E}[W_0(s)|D=0]) - (\mathbb{E}[s|D=1] - \mathbb{E}[s|D=0])\,.$$

Let f_0 denote the pdf of s conditional on $D=0$, and f_1 the pdf conditional on $D=1$. $\Delta\beta$ can then be written as $\Delta\beta = \int_0^1 (W_0(s) - s)\,[f_1(s) - f_0(s)]ds$. This is exactly the same objective function that yields $W_0^+(s|k)$ and $W_0^-(s|k)$ as worst-case weights under the restriction that $D(W, \mathbb{I}) \leq k$ assuming that $\Delta f \equiv f_1(s) - f_0(s)$ satisfies either (A3) or (A4).[24]

The assumption maintained thus far that achievement has only one dimension is unrealistic: a large and growing body of research suggests that there are multiple types of achievement relevant for labor market outcomes.[25] The mean-difference bounding analysis can be easily extended to the special case that there are multiple types of achievement that enter welfare additively separably. In particular, suppose that achievement has two dimensions with corresponding ordinally perfect test scores x and y.[26] Let $W_0(x, y)$ denote the true cardinal value of the test-score pair (x, y), and suppose that this function is known to be additively separable: $W(x, y) = H(x) + G(y)$ for two increasing functions H and G. Denote by F, F_x, and F_y the joint and marginal distributions of x and y, respectively.

Additive separability in W implies that $V(W, F)$ can be decomposed into the sum $V(H, F_x) + V(G, F_y)$.[27] In turn, this implies that F_A will be preferred to F_B for all increasing functions G and H only if $F_{A,x} \succeq F_{B,x}$ and $F_{A,y} \succeq F_{B,y}$ both hold. The dependence between x and y does not matter here; all joint distributions F with equal marginals will be ranked equally by any additively separable W. Additive separability in W does not imply that the bounding analysis can be carried out separately for each dimension of achievement. There are two subtleties that prevent one from considering each margin separately in constructing worst-case bounds.

The first subtlety is that using the sup norm to operationalize the distance restriction between W_0 and $\mathbb{I}$ links the two dimensions of achievement because the magnitude and sign of the difference in one dimension determines the range of feasible differences along the other dimension.[28] A minor tweak to the definition of D stating that the sup norm distance restriction must hold separately in each dimension is sufficient to remove this dependence. Formally, define the new distance measure as follows:

---

[24]The assumption that Δf satisfies (A3) or (A4) in this context is again quite general and will be satisfied in many economically relevant settings. D can always be defined such that Δf, and not $-\Delta f$, satisfies (A3) or (A4).

[25]Kautz, Heckman, et al.[24] provides a good introduction to and overview of this literature.

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

[27]To see this, note that $V(W, F) = \iint_0^1 H(x)f(x, y)dydx + \iint_0^1 G(y)f(x, y)dydx$. But $\iint_0^1 H(x)f(x, y)dydx = \int_0^1 H(x)f_x(x)dx = V(H, F_x)$ and $\iint_0^1 G(y)f(x, y)dydx = \int_0^1 G(y)f_y(y)dy = V(G, F_y)$.

[28]To see this, consider the restriction $D(W_0, W) \leq k$ and suppose that $\sup_x[H_0(x) - H(x)] = \lambda k$ for some $\lambda \in (0, 1)$. Then the maximum possible value of $\sup_y[G_0(y) - G(y)]$ is $(1 - \lambda)k$, while the minimum possible value is $-(1 + \lambda)k$.

Definition 5.1. Suppose that $W(x,y) = H(x) + G(y)$ and $\tilde{W}(x,y) = \tilde{H}(x) + \tilde{G}(y)$. The pairwise distance between W and $\tilde{W}$ is defined as

$$D_p(W, \tilde{W}) = \max \left\{ \sup_{x \in [0,1]} |H(x) - \tilde{H}(x)|, \sup_{y \in [0,1]} |G(y) - \tilde{G}(y)| \right\}.$$

It is straightforward to verify that D_p is a valid distance measure. Under D_p, the possible values of $\tilde{G}(y)$ consist of the entire interval $[G(y) - k, G(y) + k]$ for any functions $\tilde{H}$ and H.

The second subtlety is that it may not be possible to define A and B such that Δf_x and Δf_y simultaneously satisfy (A3) or (A4). For example, if Δf_x and $-\Delta f_y$ both satisfy (A4), then no reshuffling of labels can bring both Δf's into alignment. Since A and B may be interchanged freely, there are only two distinct situations to consider: Δf_x and Δf_y both satisfy (A4) or only one of them does. These cases can be handled by noting that in the single-dimensional case $\mathcal{W}^+(s|k, \Delta f) = \mathcal{W}^-(s|k, -\Delta f)$ and $\mathcal{W}^-(s|k, \Delta f) = \mathcal{W}^+(s|k, -\Delta f)$ always hold.

Theorem 5.2. *Suppose that (A1) and (A2) hold and that D_p is used as the measure of distance between weighting functions. If Δf_x and Δf_y both satisfy (A4) for N_x and N_y, $\mathcal{W}_x^+$ and $\mathcal{W}_y^+$ are given by equation 4.5 while $\mathcal{W}_x^-$ and $\mathcal{W}_y^-$ are given by equation 4.6. If instead Δf_x and $-\Delta f_y$ satisfy (A4), then the worst-case weights for x are unchanged. In contrast $\mathcal{W}_y^+$ is given by equation 4.6 and $\mathcal{W}_y^-$ is given by equation 4.5.*

Theorems 5.2, 4.4, and 4.5 give a general method for constructing worst-case weighting functions in the two dimensional case. This analysis can be generalized easily to more than two dimensions, provided that W_0 is additively separable in all dimensions.

6. Empirical Sensitivity Analysis

This section assesses the sensitivity to cardinal scale misspecifications of standard achievement gap/change estimates derived from several commonly used data sets. My basic approach is to use empirical test-score distributions to estimate the Δf associated with some achievement gap/change of interest. Given an estimate for Δf, I then numerically approximate $\bar{\mathcal{B}}^+(k)$ and $\bar{\mathcal{B}}^-(k)$ for various values of k. The headline conclusion from this exercise is that cross-sectional gaps are often quite robust to cardinal deviations, 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 non existent in the commonly occurring case that FOSD holds), while the values of k that are needed to flip the sign of many gap change estimates are much smaller.

6.1. **Data and Method.** I employ four commonly-used surveys in this paper: the NLSY 1979 and 1997, the 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. All four surveys have comparable demographic, income, and achievement data that allow me to estimate both income and racial achievement gaps/changes. Please refer to appendix D for a more detailed discussion of these data.

I always 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 of the cumulative effects of 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. I do not age adjust the test scores in my baseline specifications. However, using age-adjusted scores yields similar conclusions about the sensitivity of achievement gaps to scale misspecification. Because of the timing of the surveys, I use the first follow-up survey from the NELS, collected in 1990. I use base-year data for the remaining three surveys.

Valid gap change estimates require that test scores have a constant interpretation over time.[29] 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, sex, and household income. The only subtleties involve the use of 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. The ELS employs imputation to fill in missing values of income and other demographics. I drop the imputed values, and I also drop missing observations and invalid responses for all variables in all four surveys. At present, my analysis does not adjust for selection into the final sample.[30]

[29]In many data sets, test scores are renormed each year, invalidating this assumption. Simply normalizing scores to have a mean of 0 and a standard deviation of 1 within each year/age group is not likely to be an adequate response.
[30]In Nielsen[16] and follow-up work using the NELS/ELS, I find that neither ordinal nor cardinal income-achievement gap/change estimates are sensitive to these choices. This does not automatically imply, however, that the estimated

I approximate Δf, W_0^+, W_0^-, $\Delta V(W_0^+)$, and $\Delta V(W_0^-)$ numerically. I estimate the various Δf's 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 by their underlying achievement. 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 of these crossing points and select the configuration that maximizes bias given k. The results are not very sensitive to the fineness of the grid I employ.

6.2. **Black/White Achievement Inequality.** The Δf functions relevant for assessing black/white achievement inequality all satisfy either (A3) or (A4) for $N = 2$ or $N = 3$. Both Δf_{1990} and Δf_{2002} satisfy (A3) in the NELS/ELS data; white achievement is much higher than black achievement in both surveys. Furthermore, $-\Delta f_{t+1,t}$ satisfies (A4) for $N = 3$ for both math and reading.[31] All of the cross-sectional Δf_t's again satisfy (A3) in the NLSY data, while the gap-change $\Delta f_{t+1,t}$'s satisfy (A4) for either $N = 3$ (math) or $N = 2$ (reading). Figures A.7-A.8 plot these Δf functions.

Figure A.9 plots $\Delta V(\mathbb{I}, \Delta f)$, $\Delta V(W_0^-, \Delta f)$, and $\Delta V(W_0^+, \Delta f)$ as functions of k for both math and reading achievement in the NELS/ELS data. The qualitative results are the same for both achievement measures, so I will discuss only the math estimates. The observed math gap in the ELS is somewhat larger than the observed gap in the NEL90. Standard methods would therefore conclude that achievement inequality increased between the two surveys.[32] As k grows larger, $\Delta V(W_0^+, \Delta f)$ and $\Delta V(W_0^-, \Delta f)$ diverge from the observed cross-sectional gaps in each survey. $\Delta V(W_0^-, \Delta f)$ crosses 0 and turns negative for at $k \approx 0.34$ in the NELS; the observed black/white achievement gap in the NELS may not even correctly identify the sign of the true gap. In contrast, $\Delta V(W_0^-, \Delta f)$ never crosses 0 in the ELS data; misspecified test scales will never misidentify the sign of the black/white achievement gap in this survey. The observed black/white achievement gap change between the NELS and ELS is slightly greater than 0. As before, both $\Delta V(W_0^+, \Delta f)$ and $\Delta V(W_0^-, \Delta f)$ fan out from

---

sensitivity to cardinal deviations will be similarly unaffected. I will check the robustness of my results to these data choices in future work.

[31](A4) with $N = 3$ only holds for math achievement after the difference in the kernel-smoothed density estimates is smoothed one more time. For low values of s, the "raw" density difference bounces around close to 0, barely crossing 0 a number of times. Technically, then, I should compute bias in this case using (A4) and $N = 5$. I smooth a second time because removing these wiggles results in substantial improvements in computational speed and code simplicity. Furthermore, since the initial smoothed density estimates are only approximations, and since regions where Δf is close to 0 cannot contribute much to total bias, the conclusions derived using the twice-smoothed data should be almost identical to those using the unsmoothed density difference.

[32]The thought experiment here is that these observed gaps and Δf estimates are the population values as the group sample sizes tend to infinity.

the observed gap as k increases. $\Delta V(W_0^-, \Delta f)$ crosses 0 at $k \approx 0.29$. The change in the black/white achievement gap is relatively robust to cardinal deviations in these data.

Figure A.10 plots $\Delta V(\mathbb{I}, \Delta f)$, $\Delta V(W_0^-, \Delta f)$, and $\Delta V(W_0^+, \Delta f)$ as functions of k for the NLSY data. The cross-sectional achievement gap estimates are somewhat less sensitive to k than the gaps in the NELS/ELS data. The sign of the math gap will never be misidentified in either survey. For $k > 0.39$, the reading gaps using W_0^- turn negative, but they remain very close to 0. With slightly different smoothing settings on the kernel estimation, these asymptotes also remain above 0.[33] In contrast to the NELS/ELS data, the observed mean difference in scores suggests that black/white inequality decreased moderately between these two surveys. However, these gap change estimates are much more sensitive to changes in k. $\Delta V(W_0^+)$ crosses 0 and becomes positive at $k \approx 0.1$.

6.3. **High-/Low-Income Achievement Gaps/Changes.** I repeat the sensitivity analysis in the NELS/ELS and NLSY for achievement gaps/changes between youth from high- and low-income households. Generally, the cardinal sensitivity is more pronounced for income-achievement gaps/changes than for black/white estimates. Figure A.12 shows that the cross-sectional Δf's for math and reading in the NELS/ELS data satisfy (A3), while the gap-change Δf's satisfy (A4) for $N = 3$ (math) or $N = 2$ (reading). Figure A.13 plots the cross-sectional and gap-change ΔV's for different values of k. The observed cross-sectional gaps are positive and quite large.[34] For math achievement, the observed gap in the NELS is slightly larger than the observed gap in the ELS, while for reading achievement the situation is reversed. In neither survey does $\Delta V(W_0^-)$ for math ever drop below 0. For reading achievement, $\Delta V(W_0^-)$ barely dips below 0 for $k > 0.4$ in the NELS and never crosses 0 in the ELS.

In contrast, the income-achievement gap change estimates are not at all robust. The observed gap changes are fairly close to 0, so that relatively small values of k are sufficient to flip the sign of the observed versus the true gap change. In the NELS/ELS data, $\Delta V(W_0^+)$ for math goes from negative to positive at $k \approx 0.1$, while $\Delta V(W_0^-)$ for reading flips from positive to negative at $k \approx 0.04$. Cardinal methods applied to almost any test scale would correctly identify a large positive income achievement gap in any cross section, but cardinal methods applied to misspecified scales could quite easily misidentify the sign of the gap change in the NELS/ELS data.

6.4. **What if Z-Scores Are Used?** The calculations in section 6.3 deviate from most of the literature on achievement inequality in that they do not use cohort/year/age z-scores to estimate achievement differences. Instead, they use equivalent scores that enable one to rank students from different surveys against each other consistently. There are strong reasons to prefer equivalent scores, and there is no

[33] I plan to develop valid inferential procedures for this setting in future work.

[34] In many data sets covering recent decades, the top vs. bottom quintile achievement gap measured in standard-deviation units is roughly equal to the black/white achievement gap.

reason to think that z-score gap/change estimates will be more robust to cardinal deviations. Indeed, I demonstrate in this section that estimates computed using z-scores are similarly, if not more, sensitive to cardinal deviations than estimates using equivalent scores.

Figures A.11 and A.14 reproduce figures A.9 and A.13 for black/white achievement gaps/changes in the NELS/ELS data using survey/age z-scores instead of equivalent scores. The difference in robustness between cross-sectional and gap-change estimates is even starker using z-scores. Neither cross-sectional black/white math gap ever falls below 0, while the W_0^--measured gap change flips sign at $k \approx 0.06$. This is a much lower critical value than the $k \approx 0.3$ needed to flip the sign using equivalent scores. The z-score gap/change estimates for reading achievement likewise do not suggest greater robustness to cardinal deviations. The differences in cross-sectional income-achievement gap sensitivity are less dramatic. The income-achievement gap change estimates are substantially more sensitive to cardinal deviations than are the cross-sectional estimates; the observed gap change using reading z-scores is very close to 0, so that the sign of the estimate flips at $k \approx 0$.

6.5. **The Magnitude of k.** The empirical estimates using the NELS and ELS cohorts showed that some achievement gaps/changes are identified up to sign 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 actually 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 (remember that I simply normalized both s and $W_0(s)$ to be in [0,1]). 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 and in other papers questions whether such z-scores have an interpretable scale, it is still possible for me 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.[35] Cross-sectional black/white and high-/low-income mean achievement gaps are

[35]Krueger[12] 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 [10] review the

typically around 0.5 to 0.8 standard deviations, again making $k = 0.04$ seem relatively small and $k = 0.4$ relatively large.[36]

Figure A.15 plots W_0^+ and W_0^- for $k = 0.1$ and $k = 0.4$ using the income-achievement math Δf estimated from the NELS/ELS data. 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 of figure A.15 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 A.16 plots $W_0^-(s|k = 0.04)$ and $W_0^+(s|k = 0.04)$ for the case that Δf is symmetric and satisfies (A3). Since Δf is symmetric, all of the weighting functions are symmetric as well. Visual inspection suggests that $k = 0.04$ is not much of a deviation, while $k = 0.40$ marks a substantial departure from cardinality.

6.6. **Estimation Error.** The analysis so far has ignored estimation error in calculating the values k^* for which the W_0^+ or W_0^--weighted gaps/changes flip sign relative to their observed counterparts.[37] The Δf's that critically determine the sensitivity of the gap/change estimates to cardinal deviations are themselves estimated from the data. The true Δf's might differ substantially from their sample analogues, which implies that the estimated k^*'s may differ from their population values.

From one perspective, this concern is secondary to the main thrust of the paper. The estimated Δf's are consistent estimates of the population Δf's, and, as such, they are plausible guesses for Δf's that govern bias in important, applied settings. The empirical results show that for most of these Δf's, it is possible to flip the sign of the gap/change estimate for sufficiently large values of k. Furthermore, the results show that k^* is often quite small. Even without knowing the estimation errors associated with my empirical procedure, I have certainly supplied ample evidence that cardinal methods applied to test-score data are quite likely to be sensitive to scale misspecification.

However, in order to state with confidence that the specific gaps/changes I have identified as being sensitive to cardinal deviations are in fact sensitive to cardinal deviations, I need some way to account for the effect of estimation error on ΔV^+ and ΔV^-. Bootstrapping is difficult to implement in this

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.

[36]In my data, the black/white math gap is 0.79 in the NELS and 0.84 in the ELS. Fryer and Levitt[7] estimate black/white achievement gaps for early elementary school students of between 0.4 to 0.7. Reardon[19] 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. In contrast, I estimate that the NELS math income-achievement gap is 1.039 and in the ELS it is 0.904.

[37]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. If there is no such k, then set $k^* = 1$.

setting because the forms of W_0^+ and W_0^- depend on the number of times Δf crosses 0, and different bootstrap iterations may result in Δf's that cross 0 a different number of times. This problem is most acute for the empirical estimates of gap changes; the cross-sectional Δf's essentially never cross 0 more than once on the interior of [0,1]. An additional difficulty is that the bootstrap has not been formally justified in this setting. Working out these theoretical and empirical challenges is on the agenda for future research.

7. Power and Size Calculations

I have shown that for sufficiently large values of k, cardinal methods using observed test scores may misidentify the sign of an achievement gap/change in the limit as the group sample sizes tend to infinity. For small values of k, the incorrectly-specified scale will correctly identify the sign of the achievement gap/change, although the relative magnitudes of the true and observed gaps may be quite far off. In this case, is there any advantage to using the comparatively simple, cardinal approaches familiar to most researchers? It turns out that there is: statistical power. In a loose sense, cardinal methods use more of the information contained in the test-score distribution. If that information mostly preserves the relevant cardinal differences in the true test scale, then such methods may be more likely to reject false null hypotheses at a given level.

7.1. **Theoretical Discussion, Cross-Sectional Achievement Gaps.**

Consider the problem of assessing which of two test-score distributions, F_A or F_B, represents greater overall achievement given independent, random samples of sizes N_A and N_B from each population. Suppose that $F_A \succ F_B$, so that any reasonable method for assessing achievement differences should asymptotically reject with probability 1 the null hypothesis that group B has more achievement. Given that $F_A \succ F_B$ is true, the power of a given testing procedure is just the probability that the false null $F_B \succeq F_A$ is rejected.

I use the procedure developed in Barrett and Donald [2] to test for stochastic dominance. This method allows one to test the null $H_0 : F_B(s) \leq F_A(s) \forall s$ against the alternative $H_1 : \exists \tilde{s} | F_B(\tilde{s}) > F_A(\tilde{s})$ using a test statistic, $\hat{BD}$, that is a modified form of the well-known Kolmogorov-Smirnoff statistic.[38] Since the null of this test is exactly the false null that we wish to reject when $F_A \succ F_B$, the relevant power is just the probability that this null is rejected. To my knowledge, there is no analytic formula for the power of this test. Therefore, I use simulation in the next section (7.2) to compare the power of the Barrett and Donald testing procedure to the power of (cardinal) z-tests of the difference in group means when $F_A \succ F_B$.

[38]Formally, they define $\widehat{BD} \equiv \sqrt{\frac{N_A N_B}{N_A + N_B}} \sup_s \left(\hat{F}_A(s) - \hat{F}_B(s) \right)$, $\hat{F}_G(s) = \sum_i^{N_G} \mathbb{I}(s_i < s)$ and show that $Pr(\widehat{BD} > c) \to \exp\left(-2c^2\right)$ when $(N_A, N_B) \to (\infty, \infty)$ such that $\frac{N_A}{N_B + N_A} \to \lambda > 0$. This implies that the level-α critical value c_α is given by $c_\alpha = \sqrt{-\frac{1}{2}\ln(\alpha)}$.

Suppose it is known that the test-score distributions of groups A and B have the same shape but that F_A is shifted to the right relative to F_B.[39] Let σ_G and μ_G represent the standard deviation and mean respectively of the test scores in groups $G \in \{A, B\}$. Since F_A is simply F_B shifted to the right, $\sigma_A = \sigma_B$ and $\mu_A > \mu_B$. In this case, the null and alternative hypotheses that correspond to the $\hat{BD}$ test of FOSD are $H_0 : \Delta\mu \leq 0$ and $H_1 : \Delta\mu > 0$ where $\Delta\mu \equiv \mu_A - \mu_B$. The statistic $\widehat{Z_{\Delta\mu}} \equiv \frac{\widehat{\Delta\mu}}{\sigma\sqrt{N_A^{-1}+N_B^{-1}}}$ is asymptotically a standard normal random variable that can be used for hypothesis testing. It is straightforward to show that the power function for this test at level α is $\pi(\Delta\mu) = 1 - \Phi\left(z_\alpha - \frac{\Delta\mu}{\sigma\sqrt{N_A^{-1}+N_B^{-1}}}\right)$.[40]

Section 4 showed that for a fixed k, W_0^+ and W_0^- have flat regions and/or discontinuous jumps. These features have the potential to affect the power of both cardinal and ordinal tests of achievement gaps. To see why, consider ordinal testing in the case that $F_A \succ F_B$ such that $F_A(s) = F_B(s) \forall s \notin [\underline{s}, \bar{s}]$ and $W_0(s) = c, \forall s \in [\underline{s}, \bar{s}]$. Under these assumptions, the observed test score distribution for group A dominates the score distribution from group B, but the economically relevant score distributions of the two groups, H_A and H_B, are equal. In this case, FOSD tests will always reject the null that $F_A \preceq F_B$ as the group sample sizes jointly tend to infinity. However, the economically relevant null is not whether F_B dominates F_A but whether H_B dominates H_A. Since $H_B = H_A$ by construction, FOSD tests of the correctly weighted score distributions will never reject the null for arbitrarily large samples. This situation will also cause z-tests on the observed scores to lead researchers to the wrong conclusion; the observed difference in means will be positive while the true difference in means is 0.

The example in the previous paragraph is quite extreme. In all of the simulations and empirical estimates I have presented, $\hat{F}_A(s) \neq \hat{F}_B(s)$ almost everywhere. Furthermore, W_0^+ and W_0^- will typically have non-flat regions precisely where $\hat{F}_A$ and $\hat{F}_B$ are most different. Nonetheless, stochastic dominance tests and z-tests (or t-tests) of mean differences will typically have different rejection rates depending on whether the observed scores or the true scores are used.

Ordinal FOSD tests using W_0^+ and W_0^- reject the null at different rates than tests using the observed scores only because W_0^+ and W_0^- are not strictly monotone functions of the observed scores. However, there is an interpretation of W_0^+ and W_0^- that avoids this problem. Consider an amendment to assumption (A2) stating that W_0 be strictly increasing everywhere on [0,1] with derivative never less than $\varepsilon > 0$. Under this alternative version of (A2), it is straightforward to show that $W_0^-(s|k)$ and $W_0^+(s|k)$ as defined in theorems 4.1 to 4.5 are just the limits of $W_0^-(s|k, \varepsilon)$ and $W_0^+(s|k, \varepsilon)$ as

[39] That is, $f_A(s) = f_B(s - \delta) \forall s$.

[40] If the variances are estimated from the data, then t-tests should be used instead of z-tests. In practice, for group sample sizes larger than 50, t-tests and z-tests provide virtually identical power for a given level α. All of these formulas hold exactly in the limit as N_A and N_B jointly go to ∞, or in the case that the score distributions are jointly normal and independent. However, the formulas will be very close approximations in even moderately sized samples.

$\varepsilon \to 0$. Since $\mathcal{B}^+$ and $\mathcal{B}^-$ are smooth functions of W_0^+ and W_0^-, the upper and lower bounds for ΔV can be thought of as the limits of the bounds using $W_0^-(s|k,\varepsilon)$ and $W_0^+(s|k,\varepsilon)$ as $\varepsilon \to 0$. For a very small value of ε, these bounds will be indistinguishable from each other. As long as $\varepsilon > 0$, the power of ordinal tests will be unchanged for any k. Put differently, there is a discontinuity in the power function of the ordinal tests when ε hits 0. Please refer to appendix C for a formal demonstration of these various claims. Figure C.1 in that appendix plots $W_0^-(s|k,\varepsilon)$ and $W_0^+(s|k,\varepsilon)$ in the case that Δf satisfies (A3).

Unlike ordinal tests, the power of z-tests using the correctly weighted sample means is a smooth function of ε for any k. This fact has several important implications. First, it implies that the power of the z-test will generally be a function of k for any $\varepsilon \geq 0$. Therefore, z-tests using the observed score distributions will either be too likely or too unlikely to reject the relevant null compared with the same test applied to the true test scale. Second, it implies that for $\varepsilon \approx 0$ and k small, z-tests using the observed scores will have greater power than ordinal FOSD tests. However, as k increases, the power of the z-tests applied to the true scores will decrease (or increase) depending on the sign of the observed gap and whether one looks at $\mathcal{B}^+$ or $\mathcal{B}^-$. At some point, the power of the cardinal tests for either W_0^+ or W_0^- may fall below the power of the FOSD tests. Furthermore, as k grows large, the difference between the power of the z-test applied to the observed scores and the power of the z-test applied to the true scores will widen. In contrast, the power of the ordinal test does not depend on k.

The application of the $\widehat{BD}$ statistic to testing achievement gap changes is only slightly more involved. In a parallel working paper, I show that there are two conditions necessary to infer that the achievement gap between groups A and B narrowed unambiguously between periods t and $t+1$: $F_{A,t} \succeq F_{A,t+1}$ and $F_{B,t+1} \succeq F_{B,t}$. In other words, group A's achievement needs to have declined unambiguously, while group B's achievement needs to have increased. If at least one of these stochastic dominance relationships is strict, then any increasing set of weights W would assess a smaller achievement gap in $t+1$ than in t.[41] These stochastic dominance relationships can be tested using the same $\hat{BD}$ statistics that I used to test cross-sectional gaps.

The cardinal analysis for gap-changes involves only a slight modification of the cross-sectional approach. Suppose now that $F_{A,t}$, $F_{B,t}$, $F_{A,t+1}$, and $F_{B,t+1}$ are identical except for location. Let $\Delta\mu_t$ denote the difference means between group A and group B in time t and suppose $\mu_{A,t} \geq \mu_{A,t+1}$ and $\mu_{B,t} \leq \mu_{B,t+1}$ hold with at least one of the inequalities strict. These assumptions imply that the

[41] As formulated, it is possible for either $F_{B,t} \succ F_{A,t}$ or $F_{B,t+1} \succ F_{A,t+1}$. I will usually study empirical settings where the "high" group A dominates the "low" group B in each cross section, but if this is not the case, nothing of importance changes. Instead of the gap narrowing, one would just say that B gained relative to A unambiguously.

achievement gap unambiguously decreased between t and $t+1$. As before, an appropriately chosen z-test is adequate to test the null that the gap increased against the alternative that it decreased.

7.2. **Simulation Results.** I simulate cross-sectional achievement gaps when $F_A = N(\mu_A, \sigma^2)$ and $F_B = N(\mu_B, \sigma^2)$ and $\mu_A \geq \mu_B$. If μ_A is strictly greater than μ_B, then F_A first-order stochastically dominates F_B, which implies that cardinal methods will correctly identify the sign of the achievement gap for any $k < 0.5$ and will never identify a negative gap for any k.[42] Since cardinal and ordinal methods will agree in the limit for any k, it is sufficient in this case to compare cardinal tests of $\mu_A \geq \mu_B$ against ordinal tests of FOSD. I use simulated data to estimate the power of the BD test and compare it to the theoretical z-test power.

Figure A.17 shows the simulated power of the BD test against the theoretical power of the z-test. The left panel shows that both of these powers increase as the sample sizes increase, holding $\Delta\mu$ fixed. The right panel plots both powers as a function of $\Delta\mu$ holding N fixed at 500. For small $\Delta\mu$, neither test is very powerful, and both powers increase monotonically as $\Delta\mu$ increases. Strikingly, for a given pair $(N = N_A = N_B, \Delta\mu)$, the power of the z-test lies always above the power of the BD test.[43] When the observed test scores are cardinally comparable, cardinal methods are always more powerful. The figure also shows the power curves using test scores rescaled according to $W_0^-(\cdot|k = 0.1)$. The basic patterns are largely unchanged, but the tests using W_0^- are uniformly less powerful than those using the raw scores. This is intuitive, as by construction W_0^- narrows the true mean gap as much as possible given k. It is interesting to note that drop-off in power as k increases is much more dramatic for the BD tests than for the z-tests.

Figure A.18 compares the power of the z-test applied to $W_0^-(s|k)$ for different values of k to the power of the BD test applied to the original test scores when $\Delta\mu = 0.25$ and $N = 200$. Applying the BD test to the raw scores is motivated by the re-conceptualization of $W_0^*(s|k)$ as $\lim_{\varepsilon\to 0} W_0^*(s|k)$. The power of the BD test does not depend on k, while the power of the z-test applied to $W_0^-(s|k)$ decreases monotonically in k. For small values of k, the power of the z-test is very close to its power applied to the raw scores and is strictly above the power of the BD test. As k increases from 0, these two powers get closer to each other, eventually crossing. This means that for values of k close to 0, z-tests applied to the true scores will be more powerful than ordinal tests of FOSD. However, when the k is large, ordinal tests will actually be more powerful.

The simulation results for achievement gap changes yield essentially the same conclusions. If the observed test scores are truly cardinal, then cardinal tests will have greater power. Cardinal tests will continue to be superior for small values of k, but as k grows, cardinal tests lose power. Ordinal tests

[42] When $k \geq 0.5$, W_0^- is 0 on [0,0.5) and 1 on [0.5,1], resulting in a gap estimate of 0.
[43] Except for $\Delta\mu = 0$, in which case both have power equal to α.

will be more powerful in most cases given a sufficiently large k provided that one adopts the "small ε" interpretation of W_0^+ and W_0^-.

The alert reader may have noticed something peculiar about this discussion. My claim is that when k is large, the correctly weighted test score gaps/changes may be quite close to 0. For a given sample size, this means that as k increases, the power of cardinal tests applied to the true scores to determine the sign of the achievement/gap change decreases. At the same time, under the small-ε interpretation of W_0^+ and W_0^-, ordinal tests are unchanged for any k such that the true and the observed gap/change have the same sign. But if the true mean difference in the scores is very close to 0, shouldn't cardinal tests on these scores accurately measure this difference? Why is it desirable for ordinal statistics to identify an arbitrarily small gap/change? The solution to this conundrum consists of two observations. First, the difference in power is driven by the fact that ordinal statistics only attempt to determine the sign of a given gap/change, while cardinal methods attempt to determine both the sign and magnitude of the gap/change. Second, $W_0 \in [0, 1]$ is just a normalization. The economic scale of W_0 might be huge. For example, consider W_0 denominated in units of lifetime income. For such a weighting function, even a very small difference in the normalized scale might correspond to an economically significant difference in the un-normalized scale.

8. Conclusion and Extensions

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

I use my proposed 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 cardinal deviations 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 of the estimate to be misidentified. In contrast, achievement gap change estimates in these data are much less robust; even small deviations in the cardinality of the true scale relative to the observed scale are often sufficient to reverse the sign of the estimate. Not only might standard methods misidentify the sign of an achievement gap/change in the limit as the sample sizes tend to infinity, they will also have

incorrect size and lower power than ordinal methods in finite samples if the test scale is incorrectly specified.

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. This paper has shown that relying on such methods may lead one very far astray if the true scale and the observed test scale are sufficiently different from each other. Ultimately, the true scale of achievement is unknowable in most applied work. The researcher must use her own judgment about how to use 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 scale-independent, ordinal content of the test scores.

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 is a plausible distance measure to use, and it yields tractable expressions for the worst-case score weighting functions. Nonetheless, other distance measures, such as the Wasserstein distance, 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 applicability of the methods presented here to non mean-based cardinal uses of test scores.

References

[1] 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.

[2] Garry Barrett and Stephen Donald. Consistent Tests for Stochastic Dominance. *Econometrica*, 71:71–104, 2003.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

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

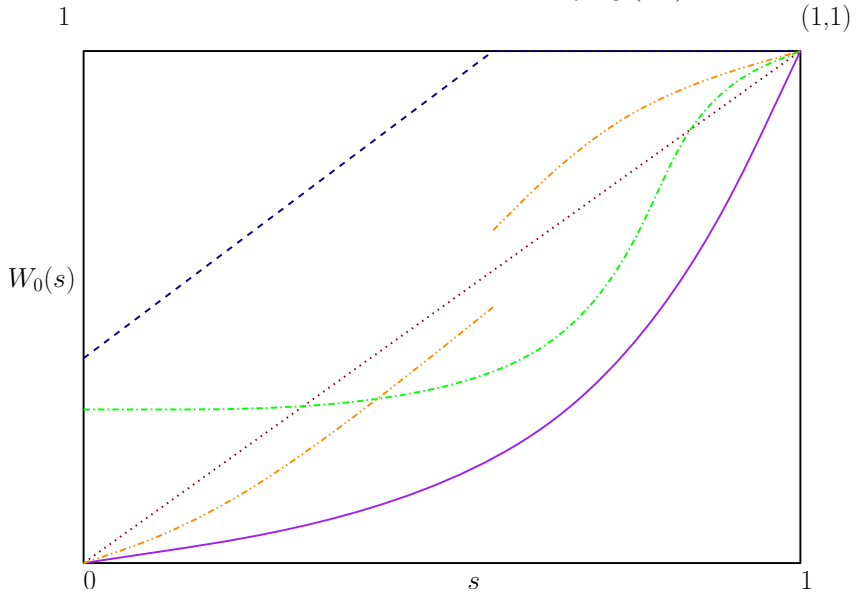
[12] Alan Krueger. Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*, 115(2):497–532, 1999.

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

[14] Frederic Lord. The 'Ability' Scale in Item Characteristics Curve Theory. *Psychometrika*, 40:205–217, 1975.
[15] Derek Neal. *Why Has Black-White Skill Convergence Stopped?*, volume 1, chapter 9, pages 511–576. Elsevier, Amsterdam, 2006.
[16] Eric Nielsen. *The Income-Achievement Gap and Adult Outcome Inequality.* PhD thesis, University of Chicago, 2014.
[17] 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.
[18] Sean Reardon. Thirteen Ways of Looking at the Black-White Test Score Gap. CEPA Working Paper, Stanford University, 2007.
[19] 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.
[20] Tarjei Havnes Rolf Aaberge and Magne Mogstad. A Theory for Ranking Distribution Functions. *IZA Discussion Papers no 7738*, 2013.
[21] D. Segall. Equating the CAT-ASVAB. In *Computerized Adaptive Testing: From Enquiry to Operation.* American Psychological Association, 1997.
[22] 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.
[23] S. Stevens. On the Theory of Scales of Measurement. *Science*, 103:677–680, 1946.
[24] Ron Diris Bas ter Weel Tim Kautz, James J. Heckman and Lex Borghans. Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success. *OECD Report*, 2014.

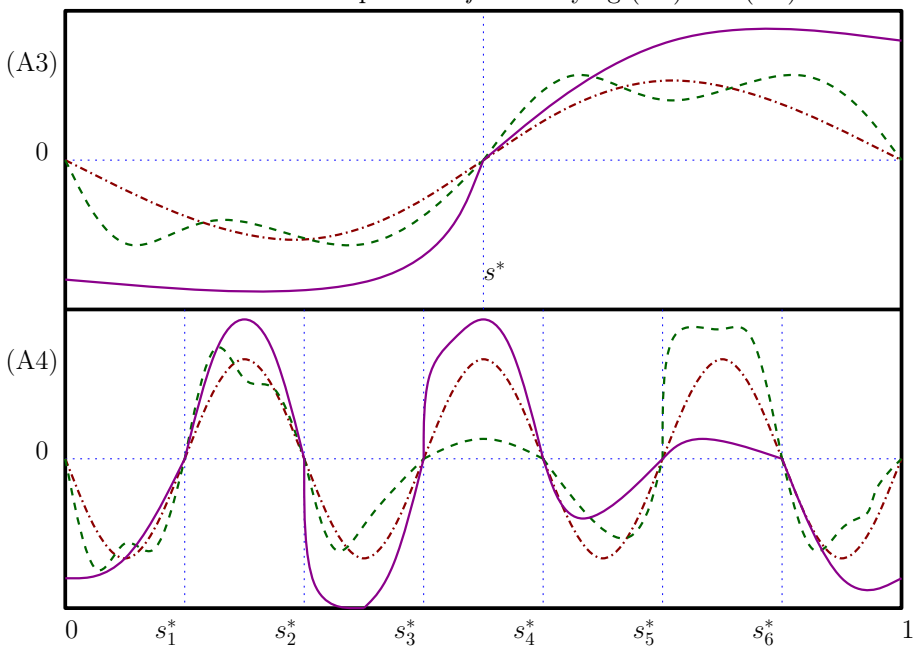
Appendix A. Figures

Figure A.1. W_0 Functions Satisfying (A2)



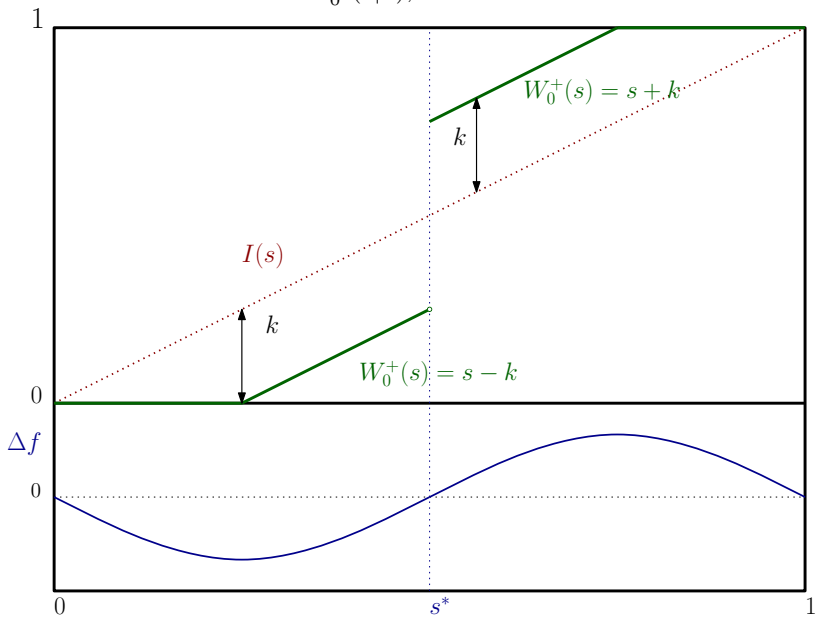
Note: Plot shows five weighting functions consistent with (A2). The red dotted curve is the identity and is the weighting function assumed when achievement gap/changes are estimated using differences in sample means. The other curves (purple solid, green dash-dot, orange dash-dot-dot, and blue dashed)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

FIGURE A.2. Examples of Δf's Satisfying (A3) and (A4)



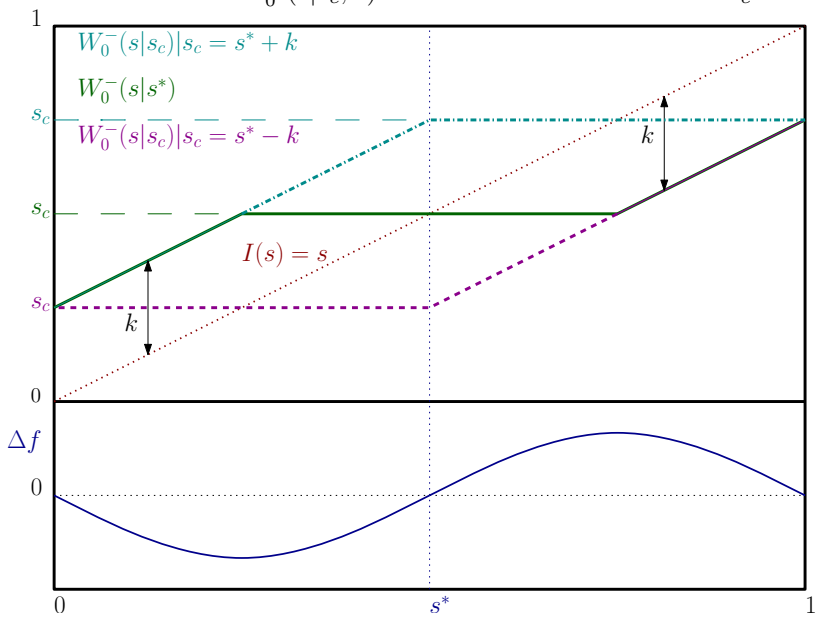
Note: (A3) and (A4) do not require there to be a single point between consecutive zeros at which $\frac{\partial \Delta f}{\partial s} = 0$. This condition does hold for the Δf's drawn as red dash-dot lines but not for those drawn as dashed green lines. Furthermore, as the solid magenta curves demonstrate, neither (A3) nor (A4) require Δf to be 0 at $s = 0$ and $s = 1$. (A3) and (A4) also do not require that the 0's be evenly spaced on [0,1], as depicted above.

FIGURE A.3. $W_0^+(s|k)$, $s^* > k$ and $1 - s^* > k$



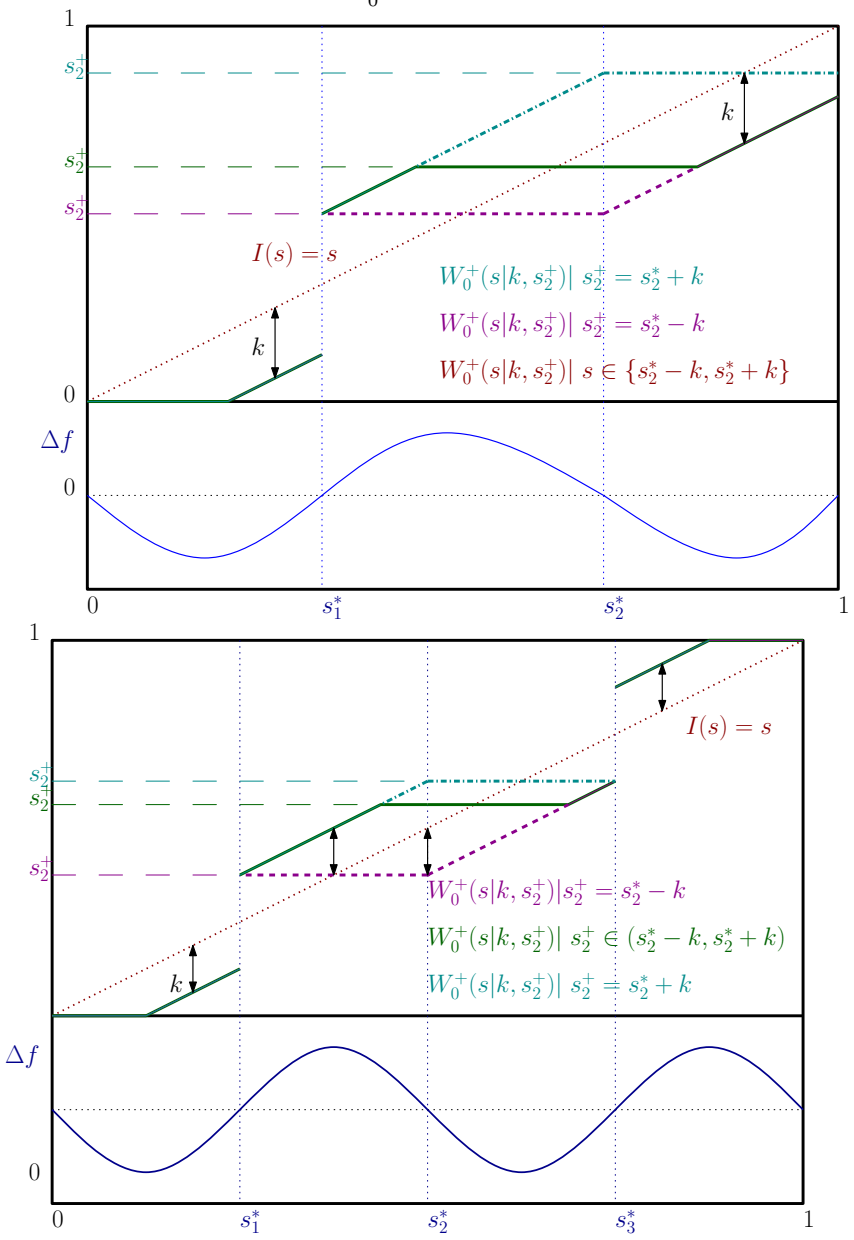
Note: The red dotted line represents the naïve weighting function. The green curve plots W_0^+ when $k < \min\{s^*, 1 - s^*\}$. For values of s less than k or greater than $1 - k$, W_0^+ is flat. W_0^+ increases 1-1 with s on the interval $[k, 1 - k]$ except for the point $s^* = 0.5$, where W_0^+ jumps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

FIGURE A.4. $W_0^-(s|s_c,k)$ for Three Different Values of s_c



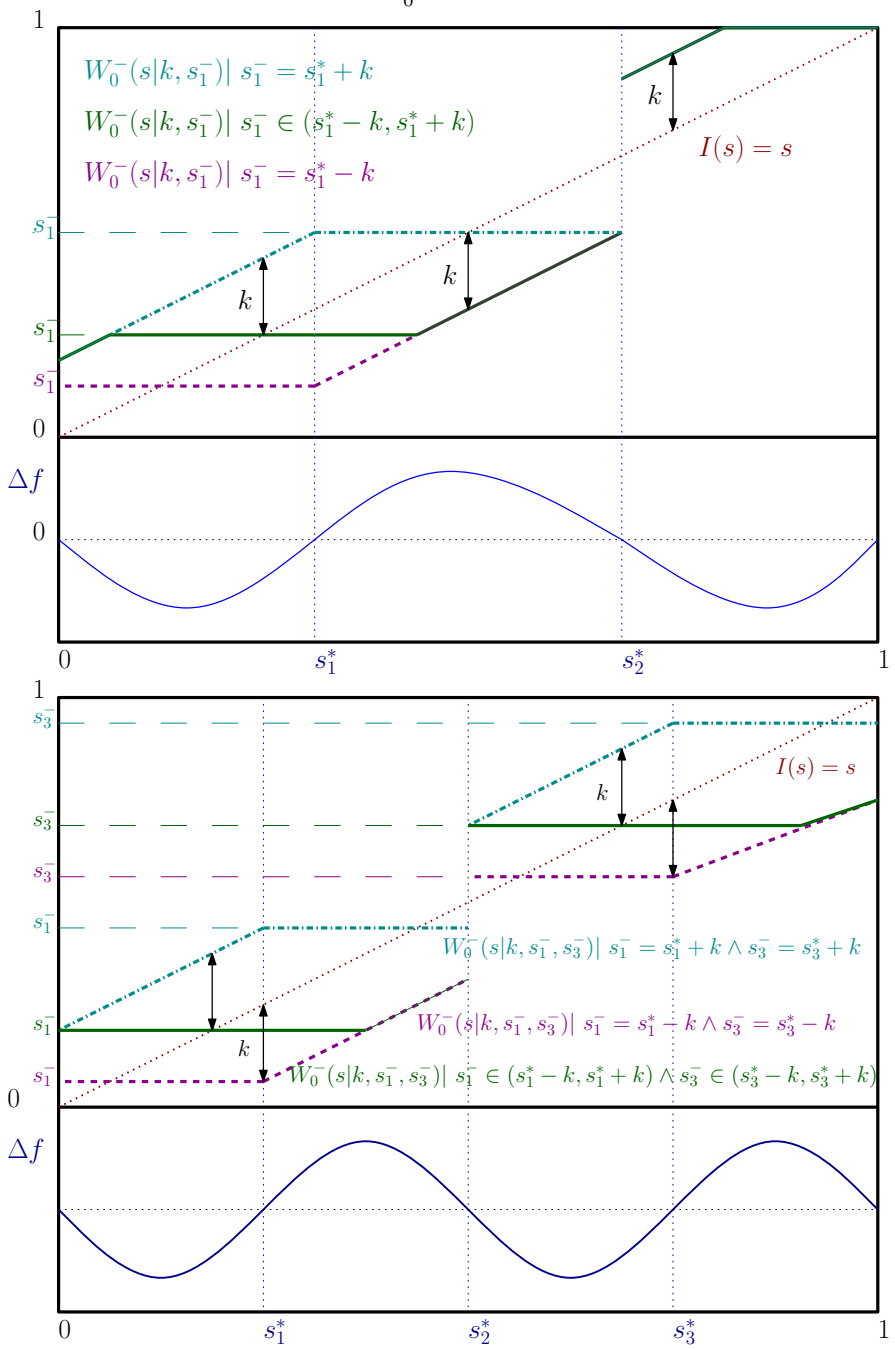
Note: The function in green plots $W_0^-(s|s^*,k)$ when $s^* - k > 0$ and $s^* + k < 1$. In this case, the constraint that $D(\mathbb{I}, W_0^-) \leq k$ binds both above and below s^*. The purple dashed curve shows $W_0^-(s|s_c,k)$ for $s_c = s^* - k$ where k is such that $s_c - k < 0$. In this case, $D(\mathbb{I}, W_0^-)$ only binds above s^*. Symmetrically, the teal dash-dot 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 A.5. W_0^+ for $N = 2$ and $N = 3$



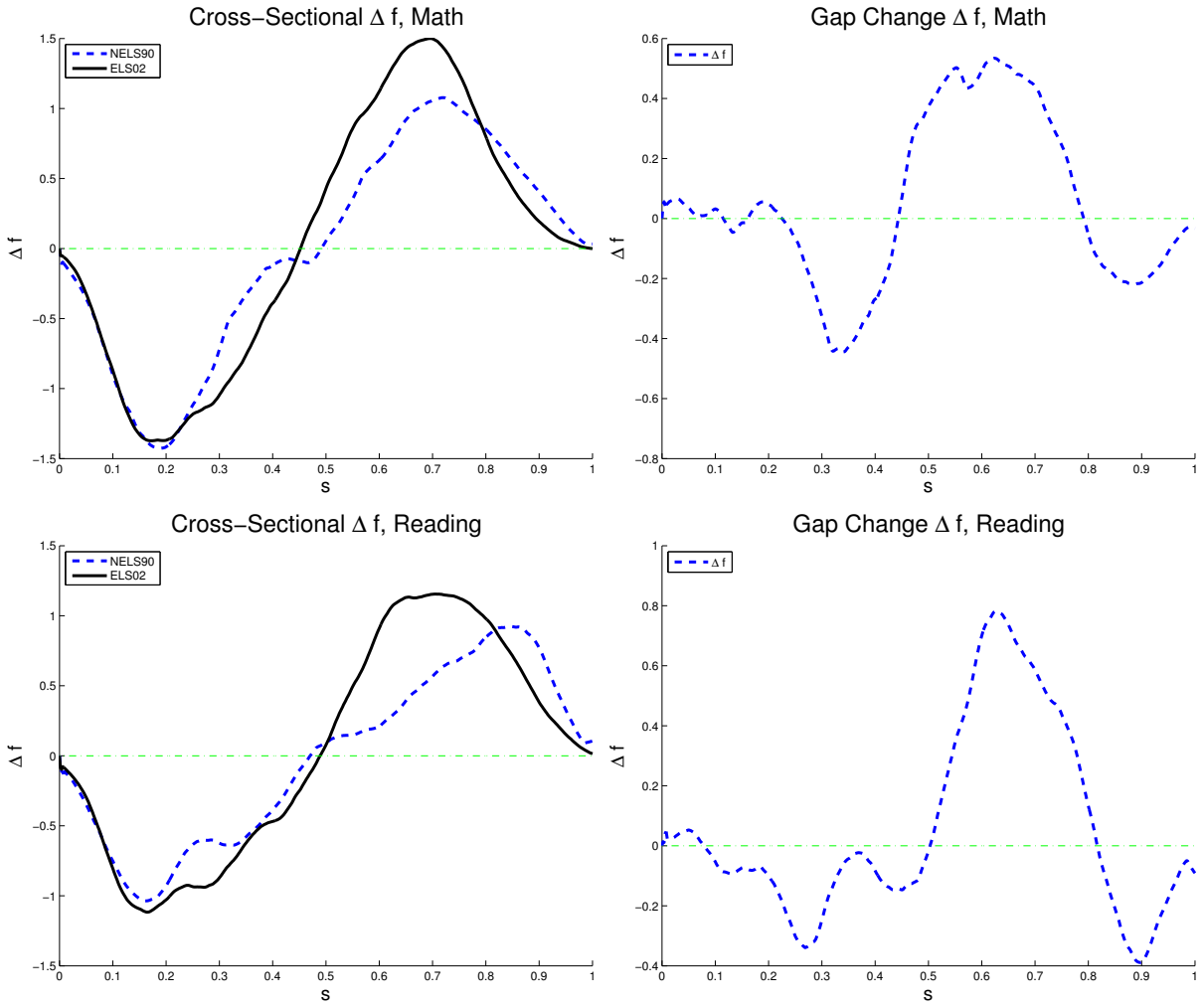
Note: The potential W_0^+'s are indexed by $W_0^+(s_2^*) \equiv s_2^+$. The dashed magenta curves depict the case that $s_2^+ = s_2^* - k$ while the teal dash-dot curves assume $s_2^+ = s_2^* + k$. The solid green curves show intermediate cases where s_2^+ lies between these two extremes.

FIGURE A.6. W_0^- For $N = 2$ and $N = 3$



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 magenta dashed curves depict the case that $s_i^- = s_i^* - k,\ i \in \{1, 3\}$, while the teal dash-dot curves set $s_i^- = s_i^* + k$. The solid green curves show intermediate cases where both values of s_i^- lie between these two extremes.

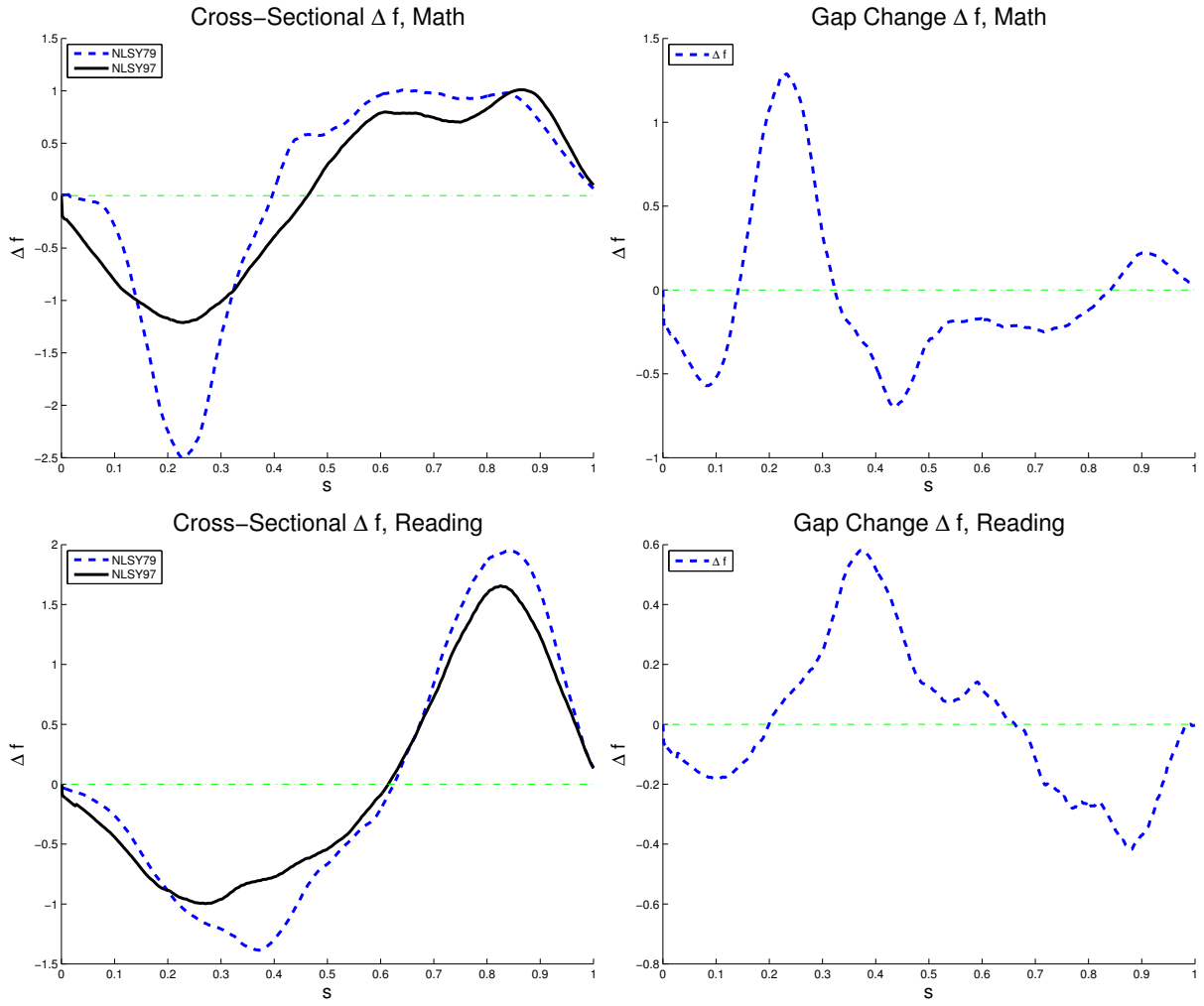
FIGURE A.7. Black/White Achievement Δf's, NELS/ELS



Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

Note: Curves estimated using Epanechnikov smoothing kernels on a grid of 5,000 points. Data cleaned as described in section 6 and appendix D. The math gap-change Δf "wiggles" around 0 for low values of s. These wiggles complicate the use of Δf in the bounding analysis, so I smooth the curve again prior to estimating W_0^+ and W_0^-. This additional layer of smoothing alters the final sensitivity estimates negligibly and greatly speeds up the computation.

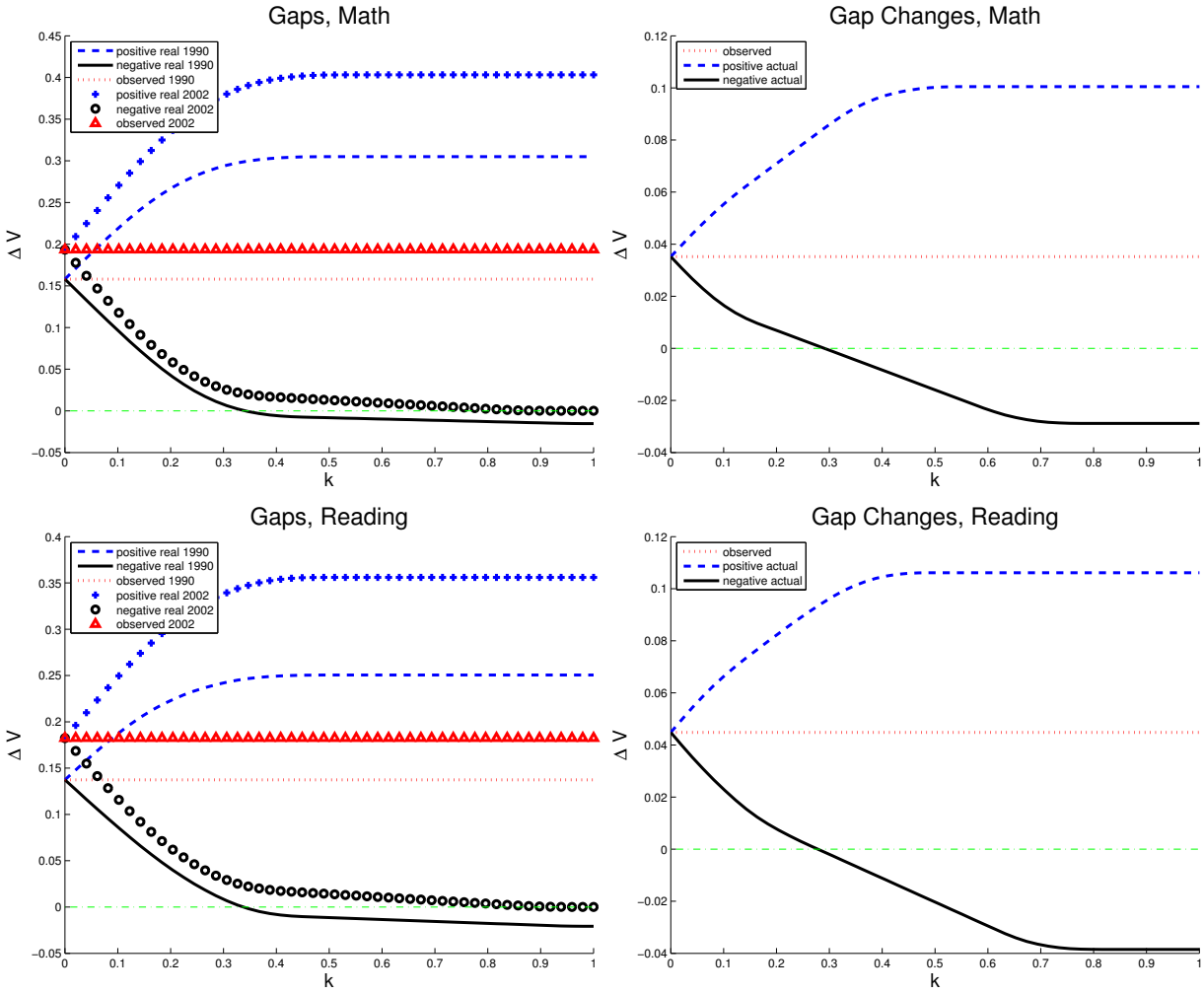
FIGURE A.8. Black/White Achievement Δf's, NLSY79 and NLSY97



Sources: Bureau of Labor Statistics, National Longitudinal Surveys of Youth, NLSY79 and NLSY97, `www.bls.gov/nls/nlsy79.htm`, and `www.bls.gov/nls/nlsy97.htm`

Note: Curves estimated using Epanechnikov smoothing kernels on a grid of 5,000 points. Data cleaned as described in section 6 and appendix D.

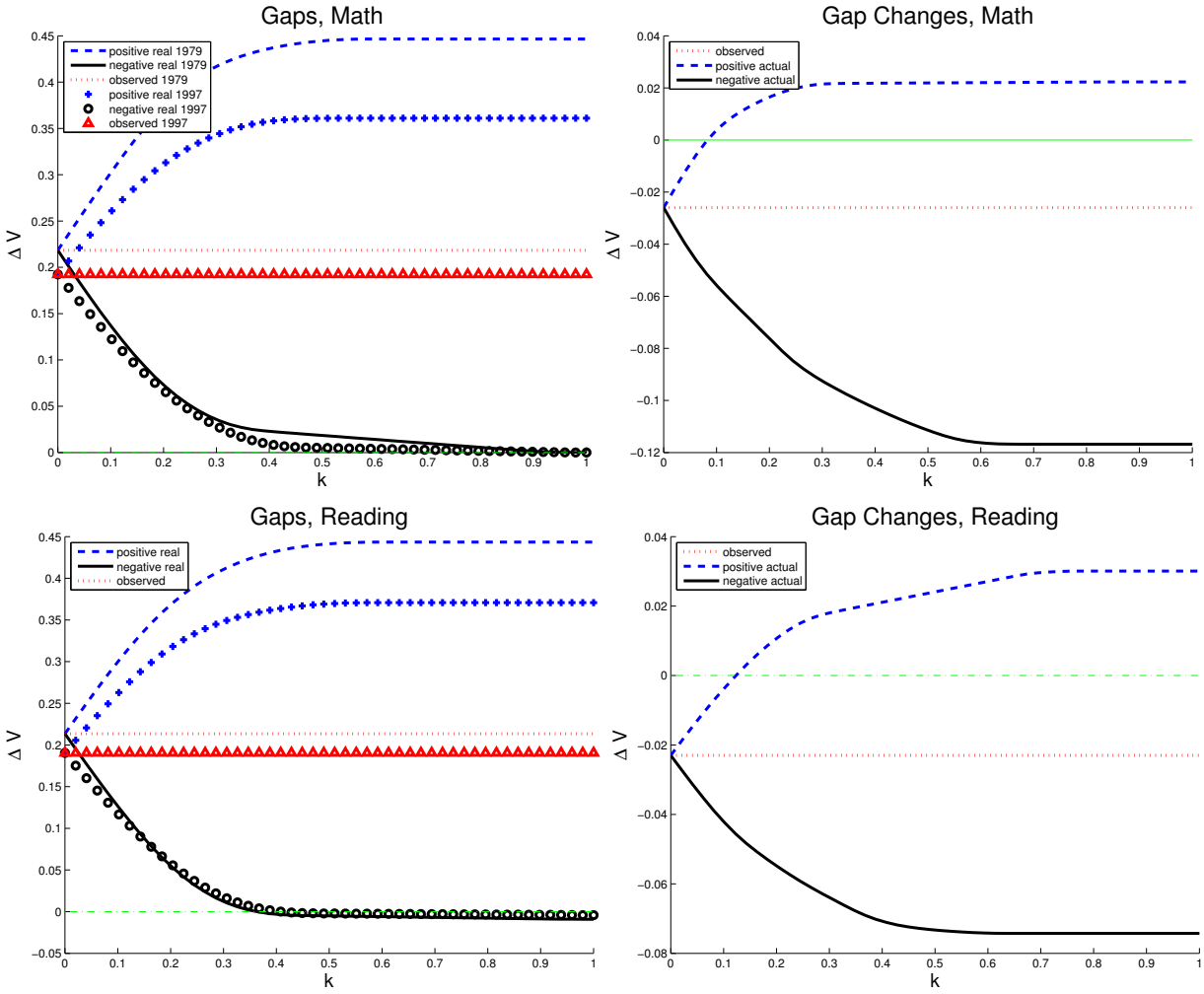
FIGURE A.9. Black/White Achievement Gap/Change Bounds, NELS/ELS



Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

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 NELS and ELS calculated such that the differences in the observed curves (perfectly horizontal) equal the observed gap changes in the right-hand panels. Data cleaned as described in section 6 and appendix D.

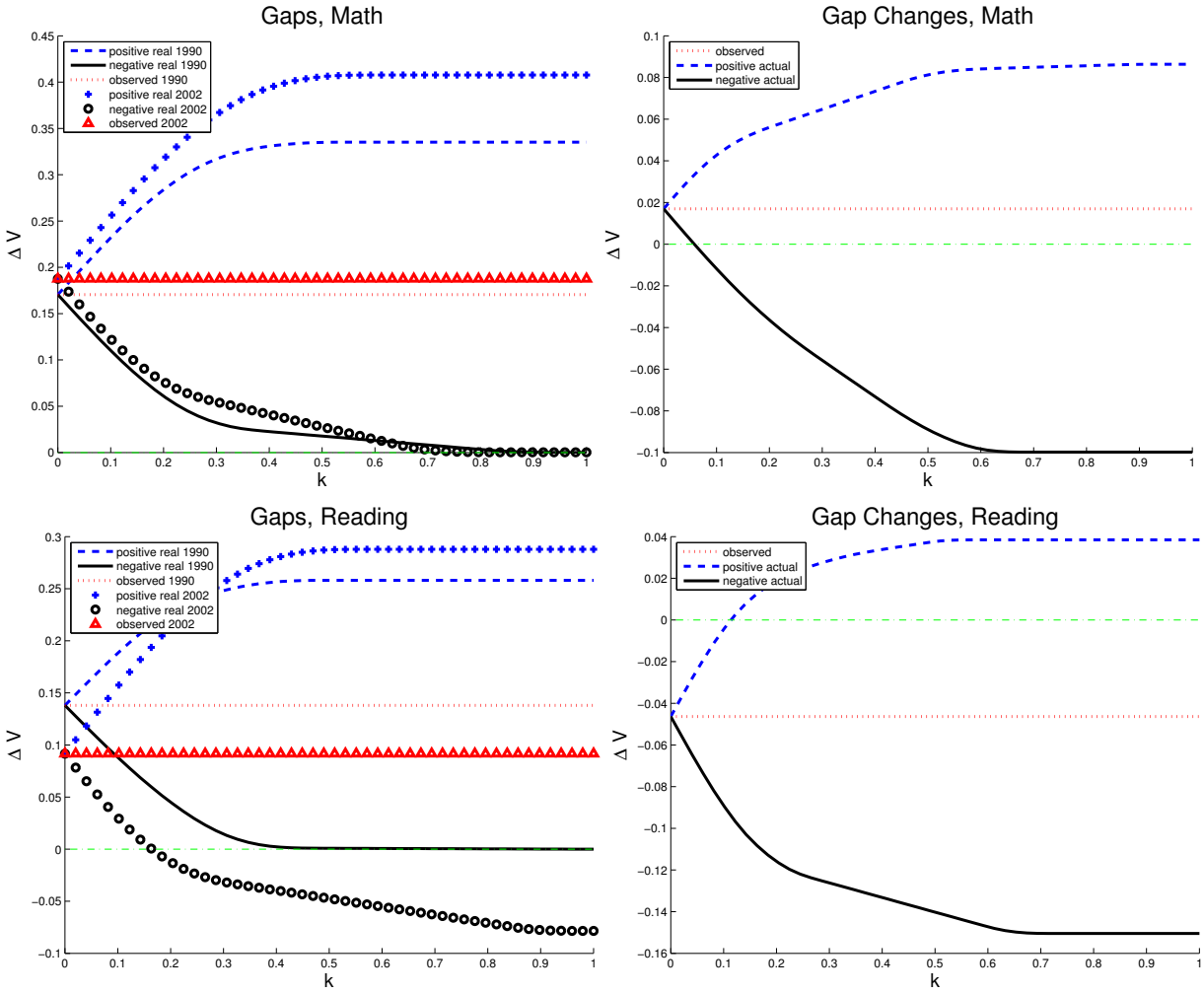
FIGURE A.10. Black/White Achievement Gap/Change Bounds, NLSY79 and NLSY97



Sources: Bureau of Labor Statistics, National Longitudinal Surveys of Youth, NLSY79 and NLSY97, `www.bls.gov/nls/nlsy79.htm`, and `www.bls.gov/nls/nlsy97.htm`

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 and NLSY97 calculated such that the differences in the observed curves (perfectly horizontal) equal the observed gap changes in the right-hand panels. Data cleaned as described in section 6 and appendix D.

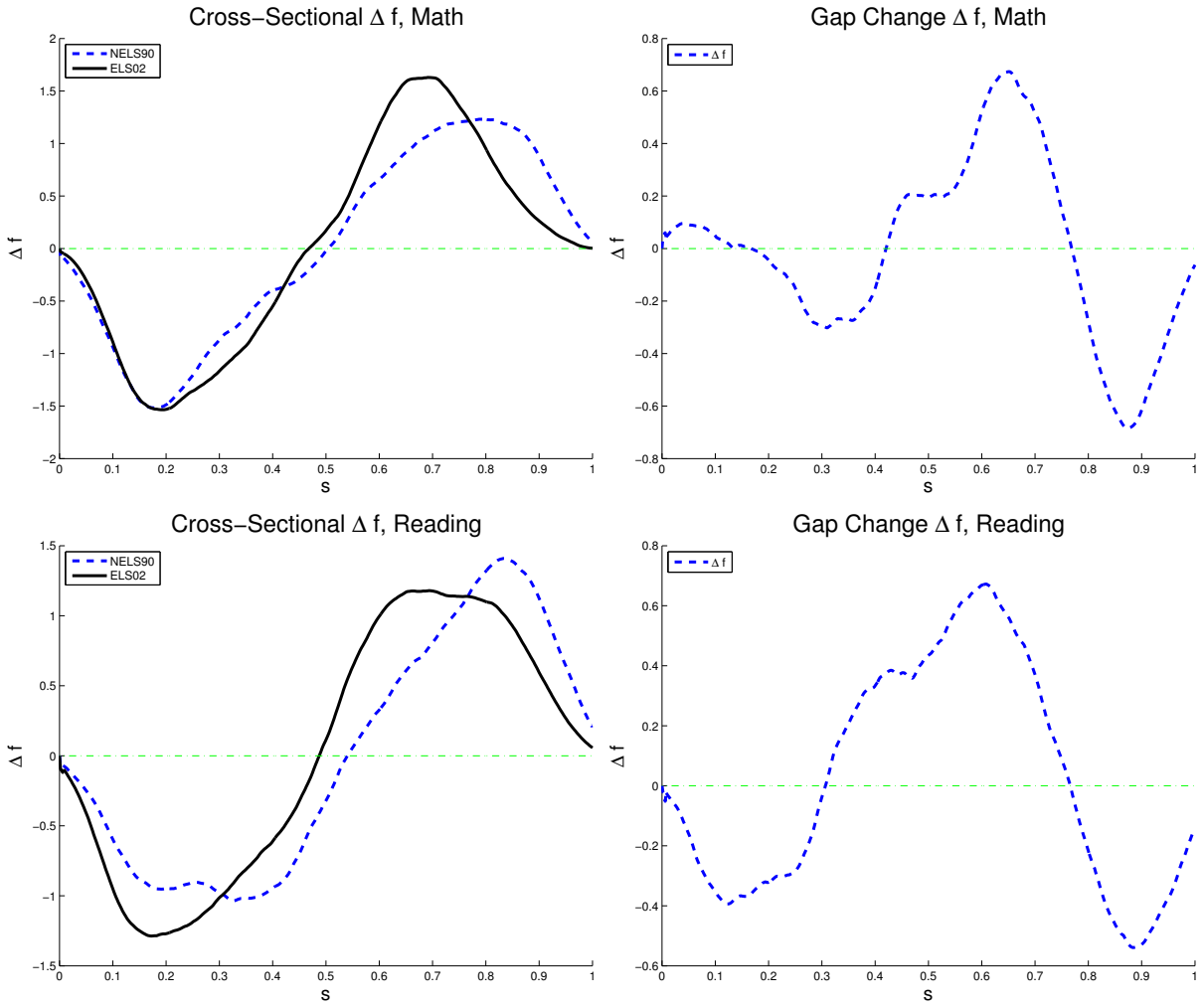
FIGURE A.11. Black/White Achievement Gap/Change Bounds Using Z-Scores, NELS/ELS



Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

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 NELS and ELS calculated such that the differences in the observed curves (perfectly horizontal) equal the observed gap changes in the right-hand panels. Data cleaned as described in section 6 and appendix D.

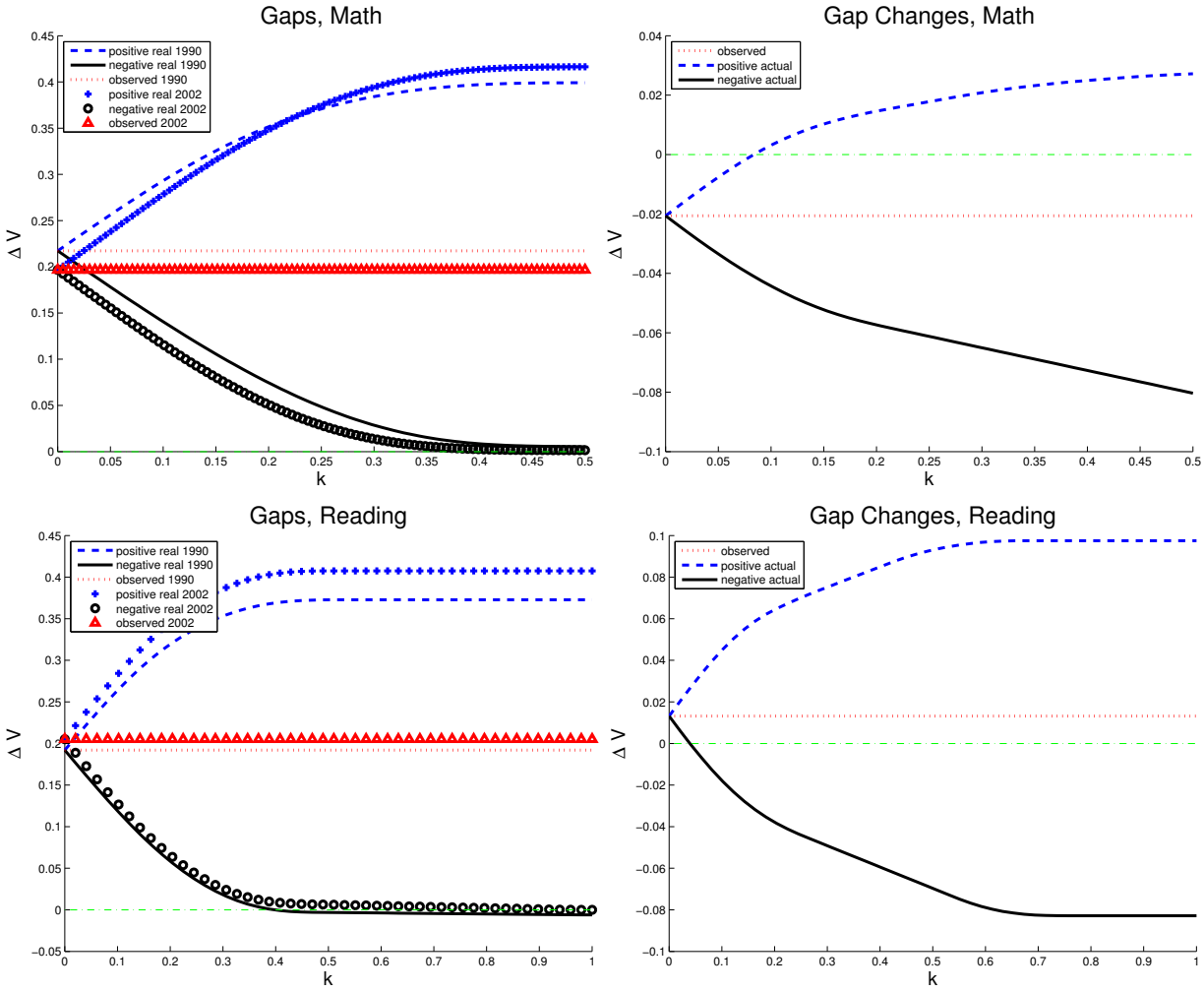
FIGURE A.12. High-/Low-Income Achievement Δf's, NELS/ELS



Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

Note: Curves estimated using Epanechnikov smoothing kernels on a grid of 5,000 points. Data cleaned as described in section 6 and appendix D.

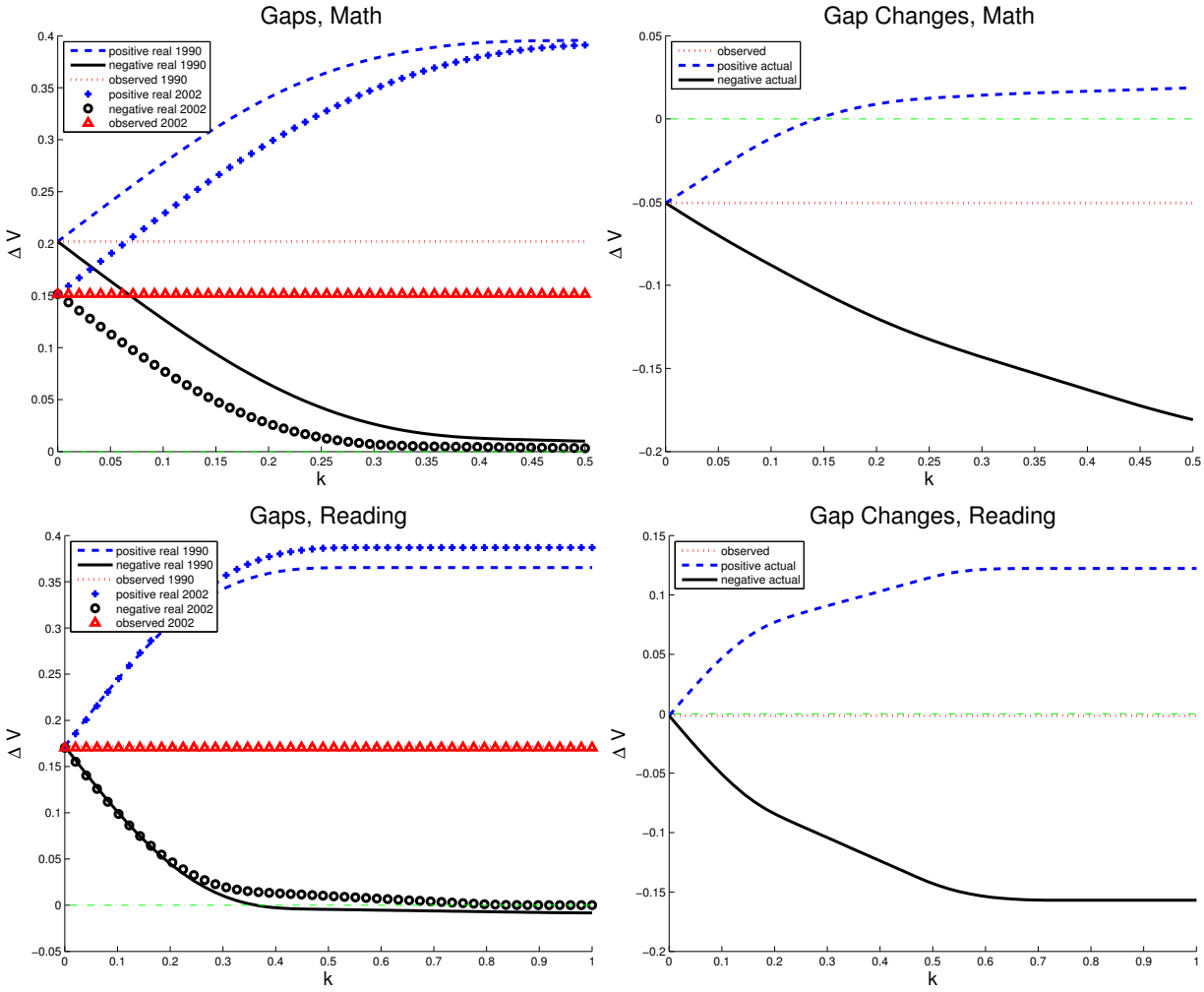
FIGURE A.13. High-/Low-Income Achievement Gap/Change Bounds, NELS/ELS



Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

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 NELS and ELS calculated such that the differences in the observed curves (perfectly horizontal) equal the observed gap changes in the right-hand panels. Data cleaned as described in section 6 and appendix D.

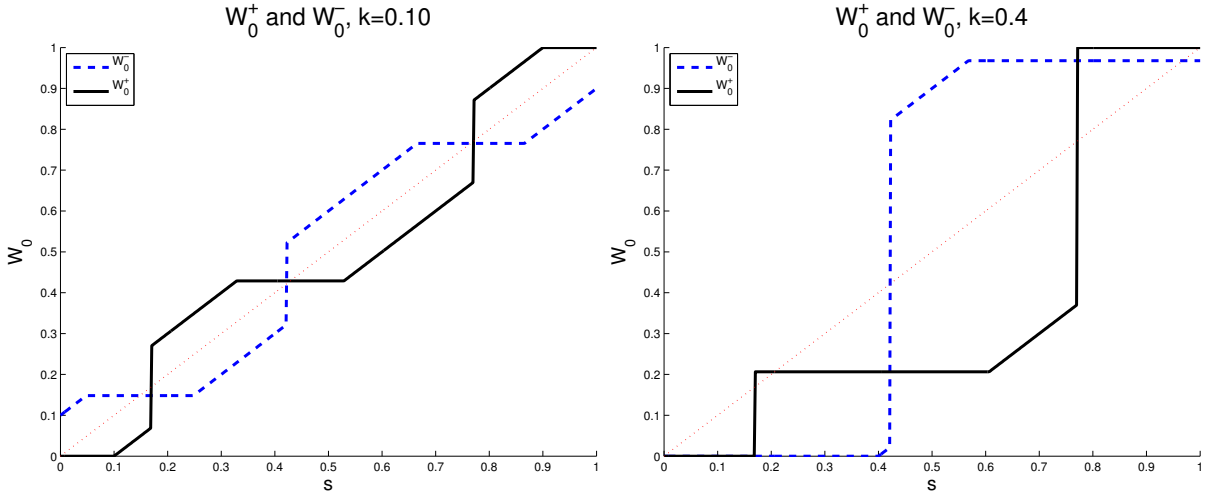
FIGURE A.14. High-/Low-Income Achievement Gap/Change Bounds Using Z-Scores, NELS/ELS



Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

Note: Curves estimated using Δf's calculated on a grid of 5,000 evenly spaced points and 500 evenly spaced values of k. Left-hand panels show the cross-sectional gaps for the NELS and ELS calculated such that the differences in the observed curves (perfectly horizontal) equal the observed gap changes in the right-hand panels. Data cleaned as described in section 6 and appendix D.

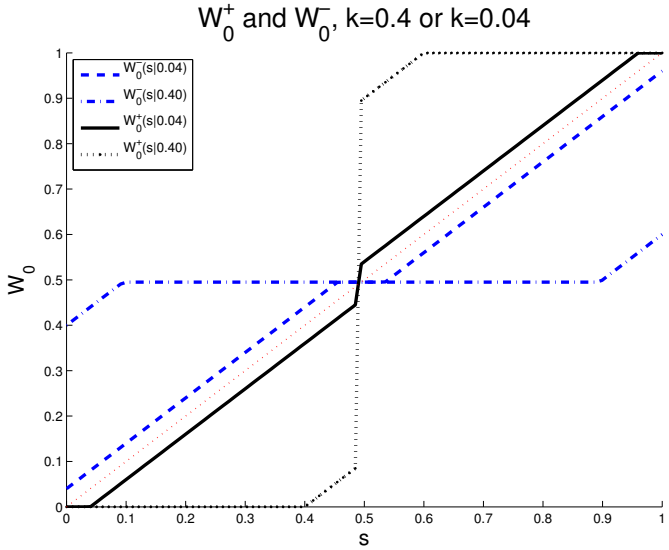
FIGURE A.15. Large- and Small-k W_0^+ and W_0^-, High-Low Income Math Δf, NELS/ELS (A3)



Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

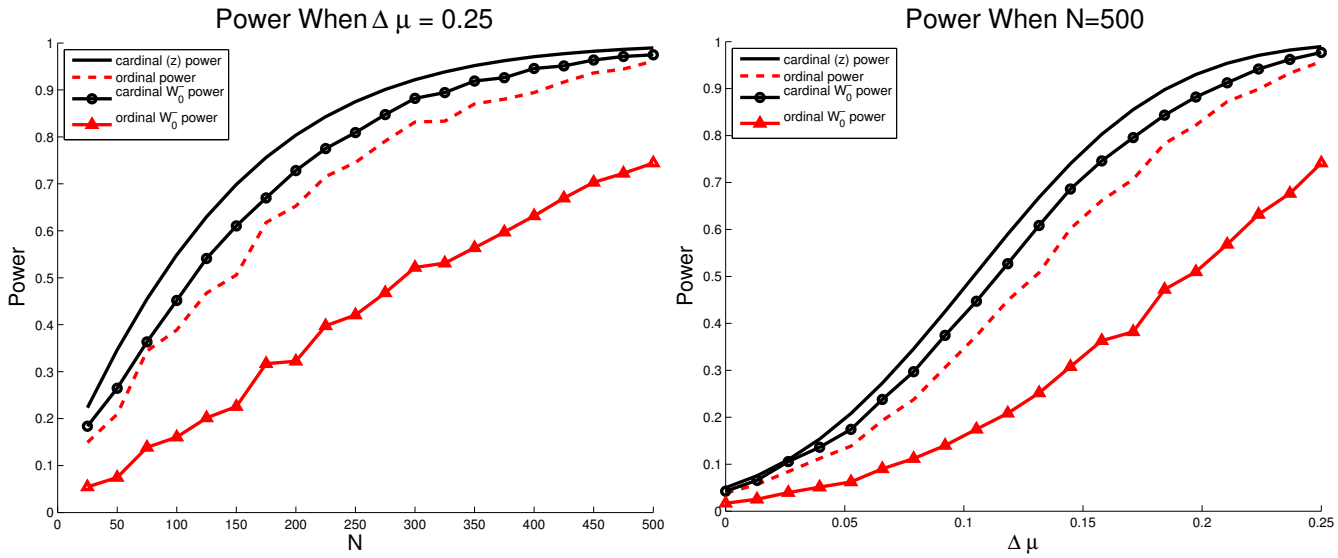
Note: Curves estimated using Δf's calculated on a grid of 5,000 evenly spaced points. Data cleaned as described in section 6 and appendix D.

FIGURE A.16. Large- and Small-k $W_0^+$and W_0^-, Symmetric Δf Satisfying (A3)



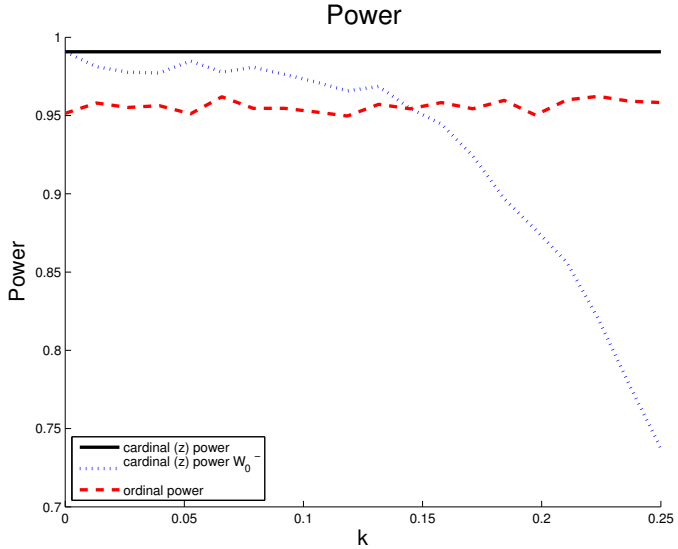
Note: Curves estimated using Δf's calculated on a grid of 5,000 evenly-spaced points.

FIGURE A.17. Ordinal vs. Cardinal Power Using $\mathbb{I}(s)$ and $W_0^-(s|k=0.1)$



Note: Plot shows cross-sectional power for z-tests and BD tests where the raw data is drawn from $F_A \sim N(0.25, 1)$ and $F_B \sim N(0, 1)$. The solid curve and dashed curve show the relationship between sample size and power for z-tests and BD tests when raw test scores are used. The circle line and triangle line show the corresponding powers when $W_0^-(s|0.2)$ is used instead. The power of the BD testing approach falls very rapidly as k increases. However, if $W_0^-(s|k = 0.2, \varepsilon = 0.0001)$ is used instead, the power of the BD test is essentially unchanged from the raw data case, while the power of the z-tests using the correctly weighted data are essentially unchanged from the $W_0^-(s|0.2)$ case.

FIGURE A.18. Ordinal vs. Cardinal Power Using $W_0^-(s|k)$ When $N = 200$



Note: Plot shows cross-sectional power for z-tests and BD tests where the raw data is draw from $F_A \sim N(0.2, 1)$ and $F_B \sim N(0, 1)$. The red dashed curve shows the estimated power of the BD tests applied to the raw test scores, while the blue dotted curve shows the power of the z-test using test scores weighted according to $W_0^-(s|k)$.

Appendix B. Tables

Table 1. NLSY79 and NLSY97 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

Sources: Bureau of Labor Statistics, National Longitudinal Surveys of Youth, NLSY79 and NLSY97, `www.bls.gov/nls/nlsy79.htm`, and `www.bls.gov/nls/nlsy97.htm`

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 6 and appendix D.

Table 2. NELS/ELS Summary Statistics

Variable	Survey	Wave	*N*	Mean	Median	S.D.	Missing	Imputed
math	NELS	1990	14,410	44.03	44.31	13.57	777	0
math	NELS	1992	12,008	49.00	49.53	14.07	2,138	0
reading	NELS	1990	14,427	30.93	31.38	9.91	760	0
reading	NELS	1992	11,999	33.33	34.68	10.01	2,147	0
age	NELS	1990	15,187	16.13	16	0.68	0	0
age	NELS	1992	14,146	18.14	18	.62	0	0
black	NELS	1990	15,187	0.12	0	0.32	0	0
black	NELS	1992	14,146	0.11	0	0.32	0	0
female	NELS	1990	15,187	0.51	1	0.50	0	0
female	NELS	1992	14,146	0.50	1	.50	0	0
math	ELS	2002	14,934	44.62	44.79	13.57	0	800
math	ELS	2004	13,444	50.22	51.38	14.13	1,148	0
reading	ELS	2002	14,934	29.29	29.65	9.44	0	933
reading	ELS	2004	NA	NA	NA	NA	NA	NA
age	ELS	2002	14,934	15.67	16	0.61	0	0
age	ELS	2004	14,592	17.70	18	0.61	0	0
black	ELS	2002	14,592	0.14	0	0.35	0	0
black	ELS	2004	14,934	0.14	0	0.35	0	0
female	ELS	2002	14,934	0.50	0	0.50	0	7
female	ELS	2004	14,592	0.50	0	0.50	0	5

Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

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 6 and appendix D.

TABLE 3. NELS/ELS Income Variables

NELS Income	Percentage Full Sample	Percentage Analysis Sample	ELS Income	Percentage Full Sample	Percentage 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			

Sources: U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

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 6 and appendix D.

TABLE 4. Cross-Sectional 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	1990	income	–	No
math	2002	income	–	No
reading	1990	income	0.38	Yes
reading	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

Sources: Bureau of Labor Statistics, National Longitudinal Surveys of Youth, NLSY79 and NLSY97, `www.bls.gov/nls/nlsy79.htm`, and `www.bls.gov/nls/nlsy97.htm`; U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

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 6 and appendix D.

TABLE 5. Gap-Change k^*'s

Survey	Subject	Comparison	k^*	Crosses?
NELS/ELS	math	black/white	0.29	Yes
NELS/ELS	reading	black/white	0.28	Yes
NELS/ELS	math	income	0.08	Yes
NELS/ELS	reading	income	0.04	Yes
NLSY79/97	math	black/white	0.11	Yes
NLSY79/97	reading	black/white	0.12	Yes
NLSY79/97	math	income	0.27	Yes
NLSY79/97	reading	income	0.05	Yes

Sources: Bureau of Labor Statistics, National Longitudinal Surveys of Youth, NLSY79 and NLSY97, `www.bls.gov/nls/nlsy79.htm`, and `www.bls.gov/nls/nlsy97.htm`; U.S. Department of Education, National Education Longitudinal Study of 1988 (NELS:88), `nces.ed.gov/surveys/nels88/` and Education Longitudinal Study of 2002 (ELS:02), `nces.ed.gov/surveys/els2002/`

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 6 and appendix D.

Appendix C. Proofs and Additional Theorems

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 4.4 and theorem 4.1) Let $\mathcal{W}_k^+$ denote the set of weighting functions satisfying (A2) and $D(\mathbb{I},W) \leq k$ that have the form given in equation (4.5). 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. □

Proof. (theorem 4.5 and theorem 4.2) Let $\mathcal{W}_k^-$ denote the set of weighting functions satisfying (A2) and $D(\mathbb{I},W) \leq k$ that can be written as in equation (4.6). 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 4.3) Consider $\frac{\partial \mathcal{B}^+}{\partial k}$. Suppose that $k < \min\{s^*, 1-s^*\}$ so that W_0^+ has the form given in equation 4.2. In this case, $\mathcal{B}^+$ may be written as $\mathcal{B}^+ = -\int_0^k s\Delta f(s)ds - \int_k^{s^*} k\Delta f(s)ds + \int_{s^*}^{1-k} k\Delta f(s)ds + \int_{1-k}^1 (1-s)\Delta f(s)ds$. Differentiating each of these integrals with respect to k yields $\frac{\partial \mathcal{B}^+}{\partial k} = \int_{s^*}^{1-k} \Delta f(s)ds - \int_k^{s^*} \Delta f(s)ds$. Now consider $\frac{\partial \mathcal{B}^-}{\partial k}$ if $s_c > k$ and $s_c + k < 1$. In this case, $\mathcal{B}^-$ may be written as $\mathcal{B}^- = -\int_0^{s_c-k} k\Delta f(s)ds + \int_{s_c-k}^{s_c+k} (s-s_c)\Delta f(s)ds + \int_{s_c+k}^1 k\Delta f(s)ds$. Taking the derivative while noting that s_c depends on k yields $\frac{\partial \mathcal{B}^-}{\partial k} = \int_{s_c+k}^1 \Delta f(s)ds - \int_0^{s_c-k} \Delta f(s)ds - \int_{s_c-k}^{s_c+k} \frac{\partial s_c}{\partial k}\Delta f(s)ds$. □

C.2. **Bounding Analysis Using Slope Restrictions.** This section derives worst-case bounds for the bias associated with using the observed test scale when W_0 is required to be strictly increasing. Very little of importance changes for the bounding analysis if the derivative of the true scale must be bounded away from 0. The functional forms of W_0^+ and W_0^- under this new restriction are very slight modifications of their unconstrained counterparts. Furthermore, as the minimum allowable rate of change in W_0 declines to 0, these worst-case functions converge smoothly to those defined in section 4. This implies that $\mathcal{B}^+$ and $\mathcal{B}^-$ also converge smoothly to the values derived in the main body of the paper. Thus, for very small ε, the unconstrained biases will be approximately correct and yet the full ordinal information of the observed test scale will be preserved in the worst-case weighting functions.

Definition C.1. W_0 satisfies (A5) for $1 > \varepsilon > 0$ iff the following hold:

(i) $\frac{dW_0}{ds}$ exists everywhere on [0,1] except at a finite number of points. Let $\mathcal{S}$ be the points in [0,1] such that $\frac{dW_0}{ds}$ is not defined.

(ii) $\frac{dW_0}{ds} \geq \varepsilon$ for all $s \in [0,1]/\mathcal{S}$.

Definition C.2. Let $\mathcal{W}_\varepsilon$ be the set of functions on [0,1] that satisfy by (A2) and (A5). Suppose that all component test-score distributions in Δf satisfy (A1). The worst-case W_0's satisfying (A2), (A5),

and $D(\mathbb{I}, W) \leq k$ for a given distance restriction k are defined by

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

Theorem C.3. *Suppose that Δf satisfies (A1) and (A3). Then there exists $s_c \in [s^* - k, s^* + k]$ such that*

$$W_0^+(s|k, \Delta f, \varepsilon) = \begin{cases} \max\{s - k, \varepsilon s\}, & s \in [0, s^*) \\ \min\{s + k, \varepsilon s + (1 - \varepsilon)\}, & s \in [s^*, 1] \end{cases}$$

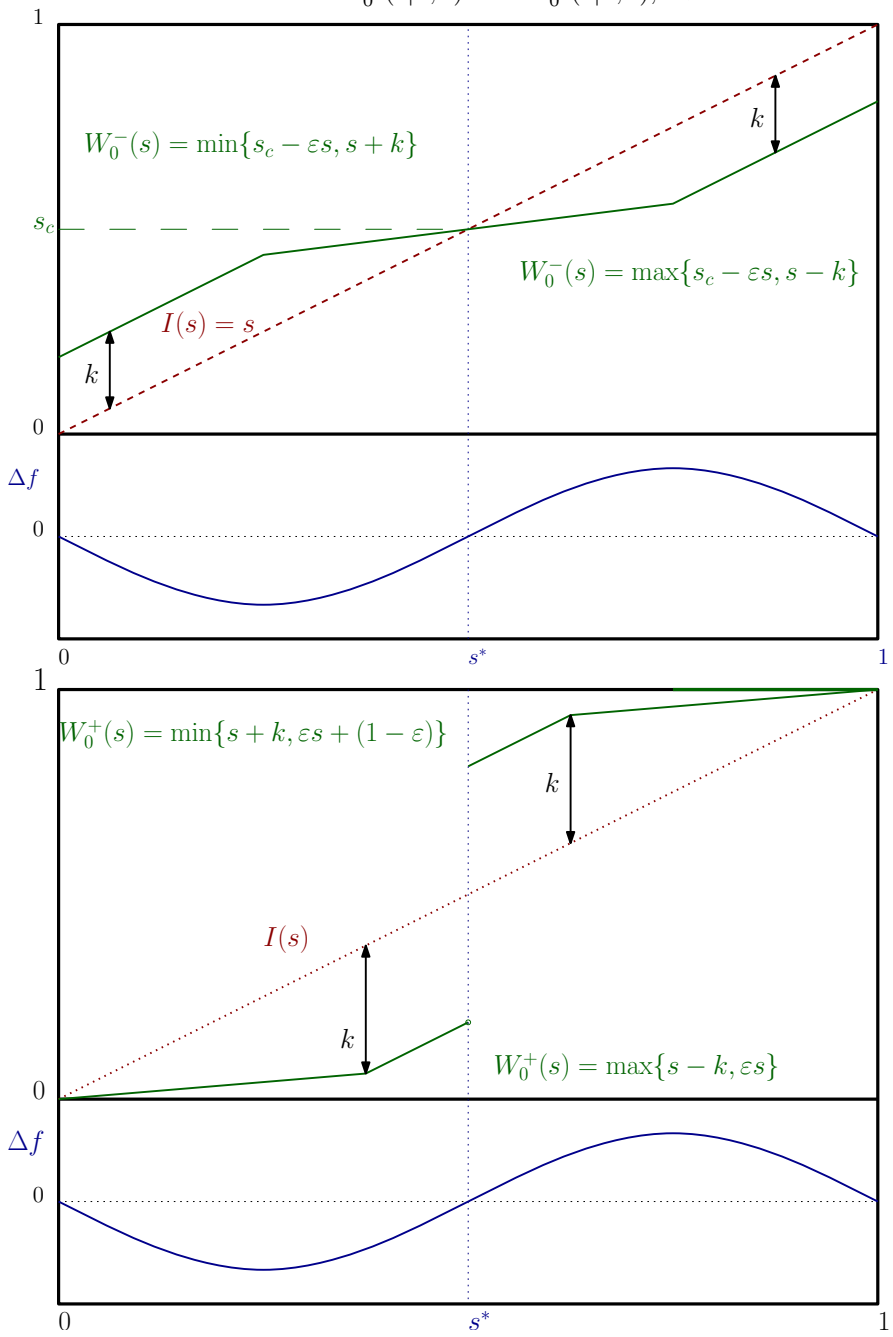
$$W_0^-(s|k, \Delta f, \varepsilon) = \begin{cases} \min\{\varepsilon s + (s_c - \varepsilon s^*), s + k\}, & s \in [0, s^*) \\ \max\{\varepsilon s + (s_c - \varepsilon s^*), s - k\}, & s \in [s^*, 1] \end{cases}$$

Proof. The proofs for $W_0^+(s|k, \Delta f, \varepsilon)$ and $W_0^-(s|k, \Delta f, \varepsilon)$ are trivial modifications of the proofs of theorems 4.1 and 4.2, respectively. □

Corollary C.4. *Suppose that Δf satisfies (A1) and (A3). Then, $\lim_{\varepsilon \downarrow 0} W_0^+(s|k, \Delta f, \varepsilon) = W_0^+(s|k, \Delta f)$ and $\lim_{\varepsilon \downarrow 0} W_0^-(s|k, \Delta f, \varepsilon) = W_0^-(s|k, \Delta f)$.*

Theorem C.3 and corollary C.4 only derive W_0^+ and W_0^- under (A3). The analysis is similar but more cumbersome for (A4) and is omitted for brevity. Figure C.2 below plots $W_0^-(s|k, \varepsilon)$ and $W_0^+(s|k, \varepsilon)$ in the case that k is small enough that the distance restriction bites both above and below s^*. These weighting functions are exactly analogous the their non-slope-constrained counterparts except that the regions on the unconstrained curves which had slope 0 now have slope ε. This modification also implies that that the kink points are slightly farther from s^* compared to the unconstrained case with the same value of s_c.

FIGURE C.1. $W_0^-(s|k,\varepsilon)$ and $W_0^+(s|k,\varepsilon)$, $\varepsilon > 0$



APPENDIX D. 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. I discuss the various ways of attacking this problem in my other working papers. 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 range of income buckets such that the mass of the bucket is as close as possible to 0.2.[44] Unlike the NELS, the ELS imputes test scores, family income, and demographic variables. I drop imputed observations from the ELS sample. My other working paper documents that the inclusion or exclusion of these observations has relatively little bearing on the sign or magnitude of 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.[45] 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[21] constructs a score crosswalk by equating percentiles on the two tests

[44]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.

[45]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.

for a sample of military recruits who were randomly assigned to one version of the test or the other.[46] I use these crosswalked scores exclusively, 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 quintiles of the base-year household income distribution, which is reported continuously. This income measure 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[16] that ordinal income-achievement estimates using these data are not sensitive to plausible alternative income definitions.[47]

[46]The crosswalk is available courtesy of Altonji, Bhadarwaj, and Lange[1] 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[22] reports that summing the crosswalked scores or crosswalking the summed scores leads to virtually identical results.

[47]For example, I estimate similar income-achievement gap changes if I use parental wage income instead of total household income to define the high- and low-income categories.