How Much Does the Cardinal Treatment of Ordinal Variables Matter? An Empirical Investigation

Jeffrey R. Bloem[†]

[†] *Department of Applied Economics, University of Minnesota.*
E-mail: bloem023@umn.edu

Summary

This is the Online Supplement for "How Much Does the Cardinal Treatment of Ordinal Variables Matter? An Empirical Investigation" (Bloem 2020). This supplemental material provides additional results and empirical illustrations supporting the implementation of this method. Questions or comments should be directed to the corresponding author at jeffrey.bloem@usda.gov.

A1. SUPPLEMENTAL EMPIRICAL ILLUSTRATIONS

The primary empirical illustration, included in the main text, re-examines the fragility of the black-white test score gap in kindergarten through third grade (Bond and Lang 2013). This section presents two supplemental empirical illustrations. The first examines Aghion et al. (2016) and the effect of creative destruction on subjective well-being in U.S. metropolitan areas. The authors examine how the determinants of economic growth, namely "Schumpeterian creative destruction,"[1] affects subjective well-being, measured by Gallup's "ladder of life" zero through ten ordinal scale. To motivate their empirical work, Aghion et al. (2016) develop an economic model that yields empirically testable predictions. In this subsection, I revisit the empirical tests of the first prediction. The key findings from tests of the first prediction is that creative destruction has a positive effect on subjective well-being when controlling for MSA-level unemployment. The following methodology will examine the robustness of this empirical finding.

The second supplemental empirical illustration examines at the work of Nunn and Wantchekon (2011) on the effect of the slave trade on trust in sub-Saharan Africa. The core finding is that present-day differences in levels of trust within communities in sub-Saharan Africa have origins in the trading of slaves across the Atlantic and Indian Oceans. In particular, individuals whose ancestors were heavily impacted by the slave trade are less trusting today. This effect persists across five measures of trust: trust of relatives, neighbors, the local council, intra-group trust, and inter-group trust. Nunn and Wantchekon (2011) use data from the Afrobarometer survey, which measures trust in the following categories: "not at all," "just a little," "somewhat," and "a lot." In the primary analysis the authors code these categories from zero through three, with zero representing "not at all" and three representing "a lot." In the following analysis I ex-

[1] Aghion et al. use "creative destruction" to refer to the sum of the job creation rate and the job destruction rate. This is analogous to the concept that Davis, Haltwanger, and Schuh (1996) call "gross job reallocation".

amine robustness of these empirical findings to monotonic increasing transformations of the ordinal scale.

The data for these empirical illustrations come from the replication files for each study.[2] In the next subsection, I will briefly outline the estimation methodologies used in each of the studies under investigation in the present analysis.

A1.1. Creative Destruction and Subjective Well-Being (Aghion et al. 2016)

In their empirical specifications Aghion et al. (2016) use a measure of creative destruction that varies at the MSA level. Since the subjective well-being measures vary at the individual level, the empirical analysis can in principle be run with either MSA-level or individual-level regressions. However, since aggregating the subjective well-being measures up to the MSA level requires an additional assumption that this procedure passes the first condition derived by Schröder and Yitzhaki (2017), for ease of exposition, I will focus on the individual level analysis of Aghion et al. (2016). The individual-level analysis also has the added benefit of being able to include more meaningful variation in individual level controls that may importantly influence subjective well-being – such as income, education, gender, marital status, ethnicity, and age. The primary empirical specification uses OLS to estimate the following equation.

$$SWB_{imt} = \alpha X_{mt} + \beta Y_{mt} + \delta Z_{it} + T_t + \epsilon_{it} \tag{A1}$$

In equation (A1) SWB_{imt} is the Gallup measure of subjective well-being for individual i who lives in MSA m in year t. In the tests of prediction one, X_{mt} is either the job turnover rate and, depending on the specification, the unemployment rate in MSA m in year t. The variables Y_{mt} and Z_{it} are MSA-level and individual level controls, respectively. The variable T_t represents year and month fixed effects. Finally, ϵ_{mt} is the error term.

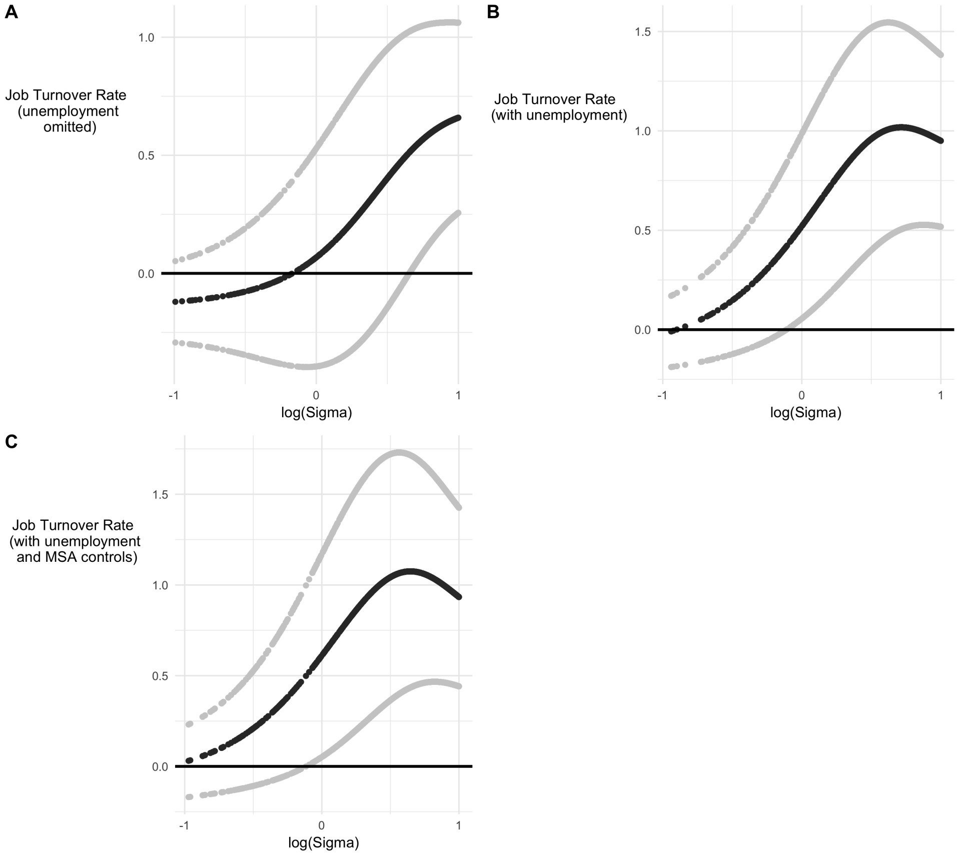
Figure A5, in section A3 of this Online Supplement, shows the LMA curves for each of the parameters of interest in predictions one through three from Aghion et al. (2016). This figure graphically illustrates that, broadly speaking, the results of Aghion et al. (2016) do not pass the theoretical sufficient conditions of Schröder and Yizhaki (2017). That is, most of the LMA curves cross the horizontal axis. It is interesting to note that each LMA curve that crosses the horizontal axis does so at a relatively high point on the subjective well-being scale. This suggests that a concave transformation of the ordinal scale can potentially change the sign of the OLS regression coefficient.

Figure A1 shows estimated effect sets corresponding with each of the three regressions testing prediction one, that job turnover increases subjective well-being more when aggregate unemployment is included as a control variable. Panel A shows the coefficient on the job turnover rate corresponding to column 1 of Table 2 in Aghion et al. (2016), when aggregated unemployment is intentionally omitted from the regression. Panels B and C show the coefficient on the job turnover rate corresponding to columns 2 and 3 of Table 2 in Aghion et al. (2016), respectively. Both of these latter specifications control for aggregated unemployment and Panel C includes additional MSA level controls.

Consistent with the theoretical predictions of Schröder and Yitzhaki (2017), Panel A shows that transformations that change the sign occur when values of $\log(\sigma)$ are between

[2] The replication files for Aghion et al. (2016) were generously shared by Gallup Inc. The replication files for Nunn and Wantchekon (2011) were available online. I thank all authors for writing clear code and organizing detailed data files.

Figure A1. Estimated Effect Sets for Prediction 1 in Aghion et al. (2016)
Globally Concave and Convex Transformations



Notes: The dark lines represent the point estimates for a given specification with the corresponding value of log(σ). Logging the value of σ allows for equal share of the graph to represent concave and convex transformations. Lighter lines represent 95% confidence interval calculated with standard errors clustered by MSA-level. Each panel refers to a different specification used to test prediction 1. Panel A refers to column (1) of prediction 1, which intentionally omits the unemployment rate and additional MSA-level controls. Panel B refers to column (2) of prediction 1, which includes the unemployment rate but intentionally omits additional MSA-level controls. Finally, panel C refers to column (3) of prediction 1, which includes the unemployment rate and additional MSA-level controls.

zero and negative one. That is, when the reporting function becomes concave, rather than linear. In panel A the coefficient changes sign for almost half of all alternative values of σ. Once the unemployment rate is included into the regression, in Panel B, the sign on the coefficient for the job turnover rate changes much less often. Finally, when additional MSA level control variables are included, in Panel C, the sign never changes. This shows that even though the sufficient conditions of Schröder and Yitzhaki (2017) suggest that the empirical results of Aghion et al. (2016) fail the second theoretical sufficient condition, once all control variables are included in the specification, the job turnover rate has a positive effect on subjective well-being for all alternative monotonic increasing transformations.

Concern persists about how *plausible* monotonic increasing transformations influence the size and statistical significance of the effect. Here the work of Kaiser and Vendrik (2019), who focus on assessing the plausibility of transformations to subjective well-being and happiness scales, is instructive. Kaiser and Vendrik (2019) cite three studies that experimentally aim to identify the functional form of the reporting function defining the relationship between subjective feelings and objective reality. Empirical analysis reported by Oswald (2008) cannot reject that the reporting function is linear—if it is curved at all, it is slightly concave. Additionally, results found by van Praag (1991) and Banks and Coleman (1981) cannot rule out the finding that individuals in their study use a linear reporting function on average. Although these cited studies provide some suggestive evidence that assuming a linear reporting function may indeed be valid, this likely will be considered a relatively strong assumption. Assuming linearity of the reporting function leads to more precise effect estimates, but ultimately with less credibility.

In the empirical setting of Aghion et al. (2016), the subjective well-being variable is measured on a zero through ten ordinal scale. The range of transformations allowed when estimating the effect sets in Figure A1, likely include implausible transformations. For example, a transformation with a σ value of 0.1 implies that the change in latent well-being associated with moving from a response of zero to a response of one is nearly 80 times larger than the change in latent well-being associated with moving from a response of nine to a response of ten. Moreover, the variation between response categories greater than five are essentially uninformative about latent well-being. The symmetric case holds for a transformation with a σ value of ten.

Based on this graphical assessment of the reporting scale, Table A1 reports plausible bounds on the effect estimates based on transformations associated with $\sigma \in [0.4, 2.5]$. This range of plausible transformations allows for both concave and convex transformations, but only transformations such that response categories are only ten times larger on opposite ends of the scale. A conceptual way to characterize these transformations is that individuals who hold a reporting function defined by $\sigma = 0.4$ are pessimistic in the sense that only relatively high levels of latent well-being are sufficient to move them off relatively low levels of the observed scale. Conversely, individuals who hold a reporting function defined by $\sigma = 2.5$ are optimistic. Finally, although this range of transformations may be plausible, each transformation within this range is likely not equally plausible. Indeed, existing experimental results (Oswald 2008; van Praag 1991; and Banks and Coleman 1981) suggest that linear transformations are likely more plausible than the concave and convex transformation allowed by this plausible range of transformations. Nevertheless, testing the robustness to the range of plausible transformations is instructive in assessing the credibility of existing empirical results.

Table A1 reports these plausible bounds on the estimated effect of creative destruction on subjective well-being. Once the subjective well-being scale is no longer assumed to be linear, several insights require brief comment. First, in terms of the qualitative result, for plausible transformations, the finding that job turnover increases subjective well-being more when controlling for aggregate unemployment persists. Estimated effects are smaller in column 1 than in columns 2 and 3 for both the upper and lower bounds. Second, statistical significance is not robust to plausible transformations. The estimates of the lower bound of the effect, reported in Panel B of Table A1, show that statistical significance is not preserved. In fact, even in specifications with all control variables included, transformations with a σ as high as 0.76 become statistically indistinguishable from zero at conventional levels. Third, the magnitudes of effects change quite

Table A1. Plausible Bounds on OLS Estimates of Prediction 1 in Aghion et al. (2016)

	(1)	(2)	(3)
A: Original 0 - 10 scale			
Job turnover rate	0.068	0.521**	0.611**
	(0.236)	(0.237)	(0.285)
σ parameter	1	1	1
R-squared	0.10	0.10	0.10
B: Lower Bound			
Job turnover rate	-0.060	0.204	0.272
	(0.154)	(0.155)	(0.183)
σ parameter	0.4	0.4	0.4
R-squared	0.099	0.099	0.099
C: Upper Bound			
Job turnover rate	0.324	0.893***	0.984***
	(0.283)	(0.287)	(0.349)
σ parameter	2.5	2.5	2.5
R-squared	0.078	0.078	0.078
Unemployment rate	No	Yes	Yes
MSA-level log of income	Yes	Yes	Yes
Additional MSA controls	No	No	Yes
Individual controls	Yes	Yes	Yes
Year and month fixed effects	Yes	Yes	Yes
Observations	556,300	556,300	461,054

Notes: This table shows bounds on the individual level results presented in Table 2 of Aghion et al. (2016). The lower and upper bounds are the smallest and largest, in absolute value, point estimates within the set of coefficient estimates. Note: this range in magnitudes persists when the marginal effects are calculated manually and expressed in terms of the original linear zero through ten ordinal scale. See the Online Supplement for additional details. Standard errors are clustered at the MSA level. *** p<0.01, ** p<0.05, * p<0.1.

dramatically for plausible transformations. Column 3 of Table A1 reports the preferred specification from Aghion et al. (2016) and shows the set of effects extend from a small and statistically insignificant effect to an effect that is statistically significant and almost 50 percent larger than the size as originally reported.

Given these results, the core prediction tested in these specifications is robust to plausible transformations of the ordinal scale measuring subjective well-being. That is, even for relatively extreme transformations (shown in Figure A1), it remains true that job turnover increases subjective well-being more when controlling for aggregated unemployment. With that said, the specific quantitative relationship between creative destruction and subjective well-being varies quite a bit for a smaller set of plausible transformations. This suggests that quantitative cost-benefit or welfare analysis with these results should consider the sensitivity of specific point estimates to plausible transformations.

A1.2. The Black-White Test Score Gap (Bond and Lang 2013)

The primary empirical illustration evaluates the black-white test score gap in kindergarten through third grade. Bond and Lang (2013) show that “plausible transformations" of test scores meaningfully change these results. This illustration, therefore, is useful to provide both a practical use and test case. Since Bond and Lang (2013) already establish the sensitivity of empirical findings to monotonic transformations of the test score, finding similar results will support the credibility of the approach developed in this paper. Although, to be clear, the method developed in this paper extends beyond the contribution of Bond and Lang (2013) by developing a partial identification method for analyzing ordinal dependent variables. Additional empirical applications, reported in the Online Supplement, include an investigation of Aghion et al. (2016) on creative destruction and subjective well-being and Nunn and Wantchekon (2011) on the slave trade and trust in sub-Saharan Africa.

In controversial and influential studies (Fryer and Levitt 2004, 2006) find that the black-white gap in test scores is relatively small and mostly explained by controlling for socioeconomic characteristics, such as child's age and birth weight, mother's age at first birth, participation in welfare programs, the number of children's books at home, and a general measure of socioeconomic status. This illustration focuses on the following specification:

$$Test\ Score_i = \gamma Black_i + X_i'\rho + \upsilon_i \tag{A2}$$

In equation (4.9) i indexes students. The variable $Black$ indicates students who identify as such and the vector X represents individual level control variables included by Fryer and Levitt (2004, 2006). Finally, υ_i is the error term. In this core illustration I will show results generated by test scores in the Early Childhood Longitudinal Study (ECLS), which includes reading test scores from the fall and spring in Kindergarten, the spring in first grade, and the spring in third grade. The ECLS also includes socioeconomic variables, which allows for the added benefit of closely mimicking the results from Fryer and Levitt (2006). Results with the inclusion of these control variables are shown in the Online Supplement. Bond and Lang (2013) also examine test scores included in the Children of the National Longitudinal Survey of Youth Kindergarten Class of 1998-1999 (CNLSY-K). Results using these test scores, the Peabody Individual Achievement Test (PIAT), are also shown in the Online Supplement.

In this section I present three elements involved in performing this method to test for robustness to the cardinal treatment of ordinal variables. First, I comment on the results of the sufficient conditions derived by Schröder and Yitzhaki (2017). These results are illustrated as graphs of LMA curves and shown in the Online Supplement. Second, I graphically report the set of effect estimates. These results are shown by plotting the point estimate and the associated confidence interval for a (relatively extreme) range of monotonic increasing transformations. Finally, I show plausible bounds on the originally-reported point estimates in tabular form.

Figure A3, in the Online Supplement, shows the LMA curves of the black-white test score gap using Early Childhood Longitudinal Study (ECLS) data. Each Panel in this figure shows the relationship between a racial status variable and the test score measured at various times between kindergarten and third grade. These graphical results show that all of the LMA curves do not cross the horizontal axis and suggest that there is no monotonic increasing transformation that can change the sign on the black-white test score gap between kindergarten and third grade. Changing of the sign, however, is not the only concern. Although the LMA curves are instructive, concern persists about the robustness of estimated effect sizes to a range of plausible monotonic increasing transformations.

A1.2.1. Graphical Results Figures 3 and 4 show effect estimates for each of the two classes of transformations, for each of the four time periods where test scores are collected in the ECLS between kindergarten and third grade. These results relate to Table 4 in Bond and Lang (2013) where the authors show several transformations that display the "fragility" of the black-white test score gap. In particular, they show several transformations: one that maximizes and another that minimizes the growth in the test score gap between kindergarten and third grade. The transformation that minimizes the gap shows the test score gap only grows 0.05 standard deviations between kindergarten and third grade. Meanwhile, the transformation that maximizes the gap shows the test score gap growing 0.64 standard deviations between kindergarten and third grade. Therefore these results are found to vary between almost no growth in the test score gap to growth that almost doubles the test score gap in just three years of early elementary education.

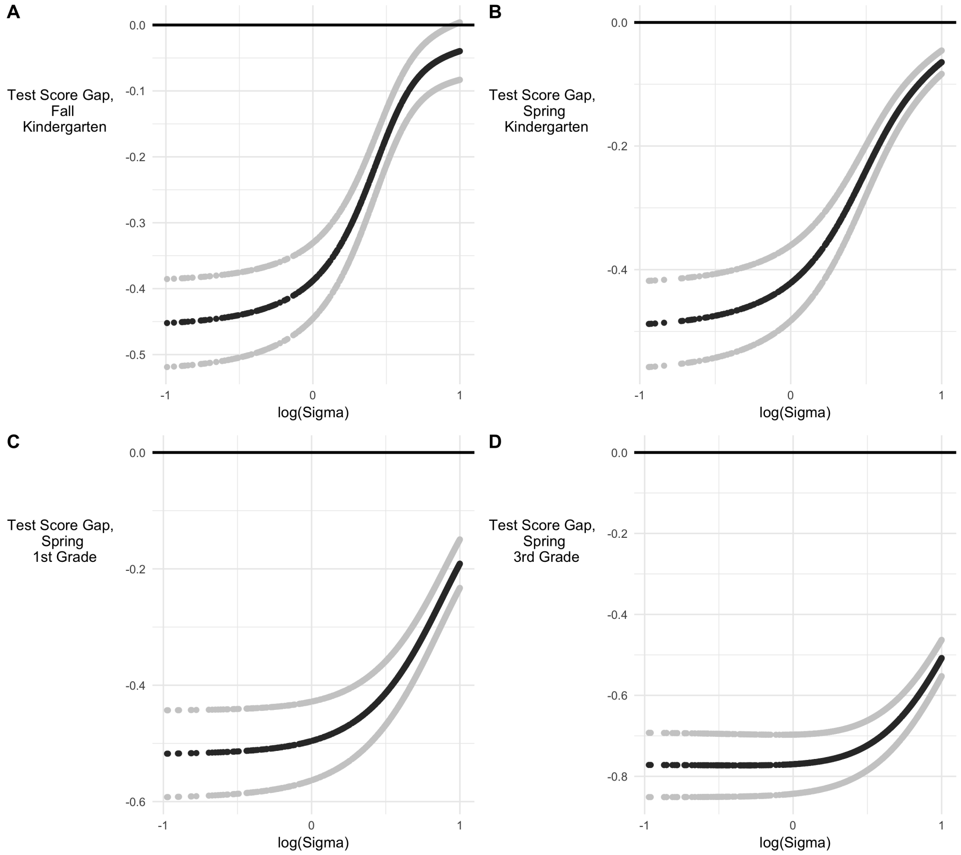
Globally Concave and Convex Transformations—In the context of test scores, globally concave and convex transformations carry implications for how the test score relates to latent student learning. If a globally concave (convex) transformation of the observed test score represents the "true" reporting function, then the marginal gain of student learning is large (small) for the first points earned on the test and rapidly diminishes (increases) thereafter. An alternative motivation for this class of transformations is if outcomes of interest are convex or concave in learning. For example, if we care most about earnings or scientific discoveries then it may be sensible to place more weight on high test scores. On the other hand, if we care most about basic competency then it may be sensible to place more weight on low test scores.

Figure 3 reports the black-white test score gaps with each panel showing how the test score gap at each time of measurement relates to a relatively extreme range of globally concave and convex transformations. The test score gap in the fall of kindergarten, shown in Panel A, is the largest with concave transformations (i.e., when the marginal gain of student learning is large for the first points earned on a test and rapidly diminishes with subsequent points). The most extreme concave transformation implies the gap could be as high as 0.46 standard deviations in the fall of kindergarten. Meanwhile, the test score gap in the spring of third grade, shown in Panel D, is the smallest with convex transformations (i.e., when the marginal gain of student learning is zero or very small for the first points earned on a test and rapidly increases with subsequent points). The most extreme convex transformation implies the gap could be as low as 0.45 standard deviations in the spring of third grade. Taken together the growth in the black-white test score gap could be a statistically insignificant 0.01 standard deviations.

At the same time, the test score gap in the fall of kindergarten is the smallest with convex transformations. The most extreme convex transformations show a test score gap of about 0.05 standard deviations in the fall of kindergarten. Meanwhile, the test score gap in the spring of third grade is the largest with concave transformations. These transformations show a test score gap of about 0.77 standard deviations in the spring of third grade. Taken together the growth in the black-white test score gap is a statistically significant 0.72 standard deviations.

Transformations with an Inflection Point—As discussed above, an alternative class of transformations are those with an inflection point. Rather than being globally convex or convex, these transformations are characterized as having local concave and convex regions. As defined in equation 3.8, this class of transformations can be roughly linear (when σ approaches Y_{Mid}) or can look like a step-function (when σ approaches zero). If a roughly linear reporting function is "true," then this implies that the marginal gain

Figure A2. Estimated Effect Sets for Bond and Lang (2013) Globally Concave and Convex Transformations

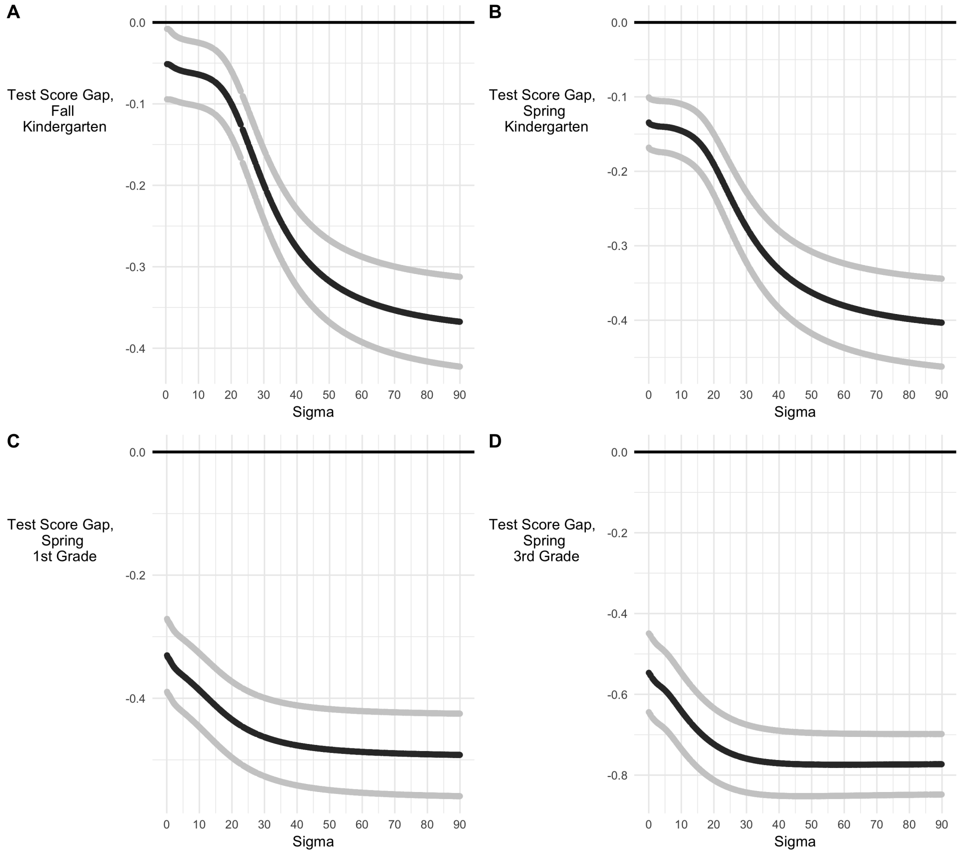


Notes: The dark lines represent the point estimates for a given specification with the corresponding value of log(σ). Logging the value of σ allows for equal share of the graph to represent concave and convex transformations. Lighter lines represent 95% confidence interval calculated with robust standard errors. Each panel refers to a test scores from different grades as shown in Table 4 of Bond and Lang (2013). Panel A refers to the test gap in the fall of kindergarten, panel B the spring of kindergarten, panel C the spring of first grade, and panel D the spring of third grade.

of student learning is roughly constant for each point earned on the test. If the "true" reporting function is best represented as a step function with the step at the mid-point in the scale, then this implies that the marginal gain of student learning is very small for the first half of points earned on a test, then immediately jumps with more than half of the question answered correctly, and is very small for the remainder of the questions. Again, an alternative motivation for this class of transformations is if outcomes of interest depend on students exceeding a threshold score. Such as, for example, a GRE math score.

The test score gap in the fall of kindergarten, shown in Panel A, is the largest with roughly linear transformations. These transformations suggest the test score gap in kindergarten could be as high as 0.37 standard deviations. Meanwhile, the test score gap in the spring of third grade, shown in Panel D, is the smallest with transformations

Figure A3. Estimated Effect Sets for Bond and Lang (2013) Transformations with an Inflection Point



Notes: The dark lines represent the point estimates for a given specification with the corresponding sigma value. Lighter lines represent 95% confidence interval calculated with robust standard errors. Each panel refers to a test scores from different grades as shown in Table 4 of Bond and Lang (2013). Panel A refers to the test gap in the fall of kindergarten, panel B the spring of kindergarten, panel C the spring of first grade, and panel D the spring of third grade.

approaching a step function. Such transformations suggest the test score gap could be as low as 0.55 in the spring of third grade. Taken together the growth in the black-white test score gap could be as small as a marginally significant 0.18 standard deviations.

At the same time, the test score gap in the fall of kindergarten is the smallest with transformations approaching a step function. Such transformations suggest the test score gap could be as low as 0.05 in the fall of kindergarten. Meanwhile, the test score gap in the spring of third grade is the largest with roughly linear transformations. These transformations show that the test score gap could be as large as 0.78. Taken together, the growth in the black-white test score gap could be as large as a statistically significant 0.73 standard deviations. Therefore, the broad fragility in the evolution of the black-white test score gap in kindergarten though third grade, discussed by Bond and Lang

(2013), is replicated by allowing for a (relatively extreme) range of monotonic increasing transformations.

A1.2.2. Plausible Effect Bounds So far the range of alternative transformations could be characterized as relatively extreme. In the case of the globally concave and convex transformations, the range is defined by any $\sigma \in [0.1, 10]$. A transformation associated with $\sigma = 0.1$ implies that a student moving from a test score of zero to one out of 180 points measures a massive change in learning, while a transformation associated with $\sigma = 10$ implies that a student would need to earn a test score of over 90 out of 180 before any noticeable change in learning occurs. In the case of transformations with an inflection point, the range extends from transformations suggesting a step function to transformations that are roughly linear. Step function transformations imply that the only useful information embedded within test scores relating to student learning is whether or not a student correctly answers more than half of the questions. A roughly linear transformation, on the other hand, implies that each test question represents an equal marginal gain in student learning. In both of these cases the range of alternative monotonic increasing transformations likely represent transformations that are implausible in this specific empirical setting. The task now is to determine a plausible range of transformations and therefore plausible bounds on the estimated effects.

What is a plausible transformation of the test score scale in this empirical context? The work of Reardon (2008), and specifically insights from applying item response theory (IRT) results to the ECLS-K data, provide structure for an assessment of plausible transformations. In particular, Reardon (2008) estimates IRT parameters indicating how each question on the ECLS-K test predicts student learning.[3] The author finds that roughly 40 percent of the questions in the ECLS-K have a relatively high likelihood of predicting no information about student learning. Most fundamentally, this suggests that the relationship between the observed test score and student learning is unlikely to be linear. It also suggests that at low (high) test score levels the marginal gain in student learning is relatively small (high). Taken together, this suggests that transformations with an inflection point are less plausible in the context of test scores, and implies that a plausible reporting function is convex to some degree. To what degree? Assuming that 40 percent of the questions could be answered correctly by guessing suggests that a σ value of 5 is relatively plausible.[4] Therefore, a plausible range of transformations extends from $\sigma \in [1, 5]$.

Based on these assumptions, Table 1 reports *plausible* bounds on the black-white test score gap. Although the range of transformations represents only a subset of those shown in Figure 3, the growth in the black-white test score gap remains relatively fragile for plausible transformations of the test score. The test score gap could be as small as 0.08 standard deviations in the fall of kindergarten (see column 1 of Panel C), and could be as large as 0.75 in the spring of third grade (see column 4 of Panel B). In this case, the black-white test score gap grows considerably between kindergarten and third grade. At the same time, the test score gap could be as large as 0.40 standard deviations in the fall of kindergarten (see column 1 of Panel B), and could be as small as 0.67 standard

[3] For transparency, the effectiveness of IRT is a historic and still active topic of debate in the psychometric literature. See, e.g., Lord (1975).

[4] That is, only after a student answers over 40 percent of the questions correctly does the test score begin to take on a positive relationship with student learning. See Figure 1.

Table A2. Plausible Bounds on OLS Estimates of the Black-White Test Score Gap

	(1) Fall Kindergarten	(2) Spring Kindergarten	(3) Spring 1st Grade	(4) Spring 3rd Grade
		A: Original 0-180 scale		
Black	-0.388***	-0.422***	-0.496 ***	-0.770 ***
	(0.0293)	(0.0311)	(0.0345)	(0.0373)
σ parameter	1	1	1	1
R-squared	0.04	0.04	0.05	0.09
		B: Lower Bound		
Black	-0.388***	-0.422***	-0.496 ***	-0.770 ***
	(0.0293)	(0.0311)	(0.0345)	(0.0373)
σ parameter	1	1	1	1
R-squared	0.04	0.04	0.05	0.09
		C: Upper Bound		
Black	-0.084***	-0.148***	-0.340***	-0.665***
	(0.021)	(0.015)	(0.024)	(0.028)
σ parameter	5	5	5	5
R-squared	0.002	0.007	0.024	0.074
Hispanic control	Yes	Yes	Yes	Yes
Asian control	Yes	Yes	Yes	Yes
Other race control	Yes	Yes	Yes	Yes
Observations	11,414	11,414	11,414	11,414

Notes: This table shows bounds on the results presented in Table 4 of Bond and Lang (2013). The lower and upper bounds are the smallest and largest, in absolute value, effect estimates within range of plausible transformations. Robust standard errors are presented in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

deviations in the spring of third grade (see column 4 of Panel C). In this case, the black-white test score gap still increases between kindergarten through third grade but at a much less dramatic rate.

This empirical illustration demonstrates how this method can be used to test for robustness of effect estimates to the cardinal treatment of ordinal variables. This illustration uses the case of test scores because this method may be most appropriate in cases when ordinal scales have many response categories. In a conceptual sense, however, this method does not only apply to test score but also to any variable that cannot be directly observed and must be quantitatively measured on an ordinal scale. Illustrations of subjective well-being and trust, using the applications of Aghion et al. (2016) and Nunn and Wantchekon (2011) respectively, are discussed in the Online Supplement.

A2. A SAMPLING OF THE CARDINAL USE OF ORDINAL VARIABLES

Table A3 shows a sampling of papers that are either highly influential or are published in top academic journals. This table is not an exhaustive list. Indeed it omits entire literatures, such as the education literature using test scores as a dependent variable. Nevertheless, this table does highlight the reality that the cardinal use of ordinal dependent variables does exist. As such, understanding the robustness of these results to monotonic increasing transformations is important.

Table A3. Examples of the Cardinal Treatment of Ordinal Variables

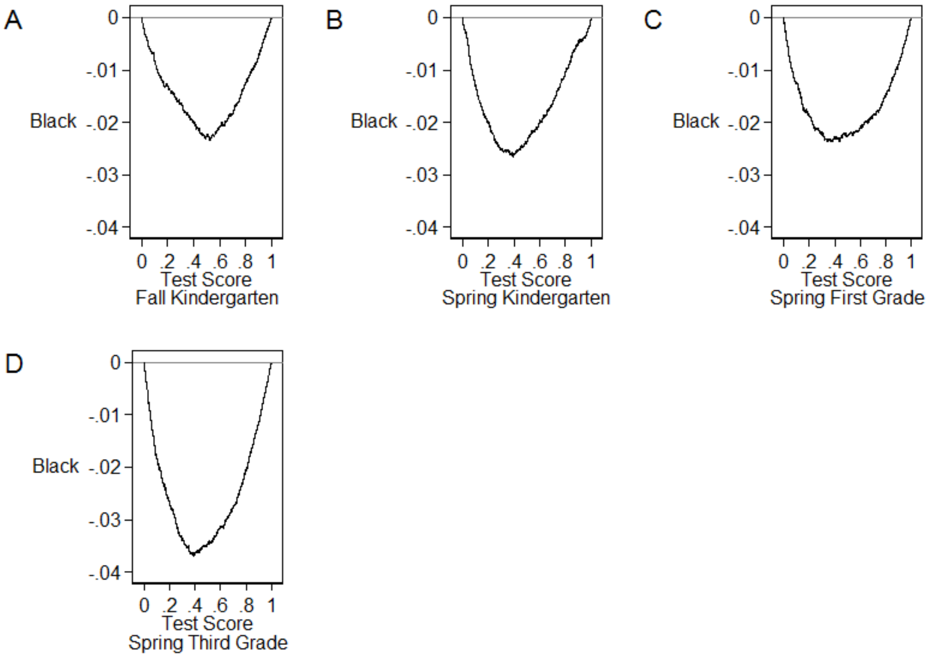
Citation	**Journal**	**Dependent Variable**	**Method**
Aghion et al. (2016)	*American Economic Review*	Subjective well-being	OLS with fixed effects
Alatas et al. (2012)	*American Economic Review*	Satisfaction	OLS
Ashraf et al. (2014)	*American Economic Review*	Subjective well-being	OLS
Bandiera et al. (2017)	*Quarterly Journal of Economics*	Mental health	OLS
Banerjee et al. (2015)	*Science*	Mental health	OLS
Bertrand (2013)	*American Economic Review: P&P*	Emotional well-being	OLS
Bianchi (2012)	*Review of Economics and Statistics*	Satisfaction	OLS
Bloom et al. (2015)	*Quarterly Journal of Economics*	Satisfaction	OLS
Bloom et al. (2015)	*Review of Economic Studies*	Management quality	OLS
Bryson and MacKerron (2017)	*The Economic Journal*	Happiness	OLS with fixed effects
Card et al. (2012)	*American Economic Review*	Satisfaction	OLS
Clark et al. (2008)	*Journal of Economic Literature*	Happiness	OLS and comparison of means
Clark et al. (2016)	*Review of Economics and Statistics*	Satisfaction	OLS with fixed effects
De Neve et al. (2018)	*Review of Economics and Statistics*	Subjective well-being	OLS with fixed effects
Deaton (2018)	*Journal of Public Economics*	Subjective well-being	OLS and comparison of means
Di Tilla et al. (2001)	*American Economic Review*	Happiness	OLS
Dohmen et al. (2012)	*Review of Economic Studies*	Trust	OLS
Dustmann and Fasani (2016)	*The Economic Journal*	Mental health	OLS with fixed effects
Frijters et al. (2014)	*The Economic Journal*	Satisfaction	OLS
Glewwe et al. (2018)	*Journal of Human Resources*	Hope	OLS
Haushofer and Shapiro (2016)	*Quarterly Journal of Economics*	Psychological well-being	OLS
Krueger and Mueller (2012)	*American Economic Review: P&P*	Emotional well-being	Comparison of means
Lachowska (2017)	*Journal of Human Resources*	Subjective well-being	OLS
Layard et al. (2014)	*The Economic Journal*	Satisfaction	OLS
Milligan and Stabile (2011)	*American Economic Journal: Economic Policy*	Emotional well-being	OLS
Moscona et al. (2017)	*American Economic Review: P&P*	Trust	OLS with fixed effects
Nunn and Wantchekon (2011)	*American Economic Review*	Trust	OLS and 2SLS
Oswald and Powdthavee (2008)	*Journal of Public Economics*	Satisfaction	OLS with fixed effects
Oswald and Wu (2011)	*Review of Economics and Statistics*	Subjective well-being	OLS
Schechter (2007)	*American Economic Review*	Trust	GMM
Steptoe et al. (2015)	*The Lancet*	Subjective well-being	OLS and comparison of means
Wunder et al. (2013)	*Review of Economics and Statistics*	Subjective well-being	OLS

Notes: This list is a sampling of papers that treat an ordinal dependent variable as if it was cardinal. This is not an exhaustive list.

A3. LMA CURVES

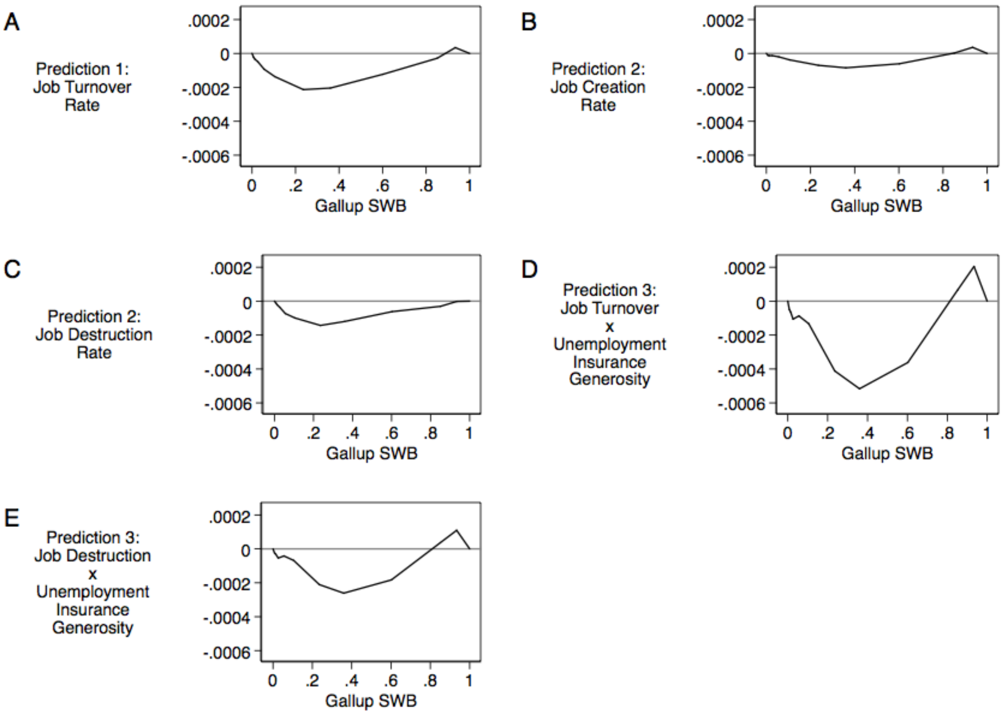
The LMA curves for the three empirical investigations are shown in the following figures. Figure A4 shows the LMA curves for the results from Bond and Lan (2013) on the black-white test score gap in kindergarten through third grade. Figure A5 shows the LMA curves for the results from Aghion et al. (2016) on the relationship between creative destruction and subjective well-being. Finally, Figure A6 shows the LMA curves for the results from Nunn and Wantchekon (2011) examining the effect of the slave trade on trust in sub-Saharan Africa. Specific details about how these LMA curves are constructed can be found in Section 2.2 of the main manuscript.

Figure A4. LMA Curves with ECLS Test Scores and Race



Notes: This figure shows LMA curves between a racial status variable and test scores. Each graph shows test scores measured in different time periods between kindergarten and third grade, as in Bond and Lang (2013). The y-axis is fixed between all graphs.

Figure A5. LMA Curves with Gallup Current Ladder SWB and Creative Destruction



Notes: This figure shows LMA curves between the Gallup "ladder of life" SWB variable and the various variable of interest for each of the first three predictions tested in Aghion et al. (2016). The y-axis is fixed between all graphs.

A4. ADDITIONAL TEST SCORE ANALYSIS FROM BOND AND LANG (2013)

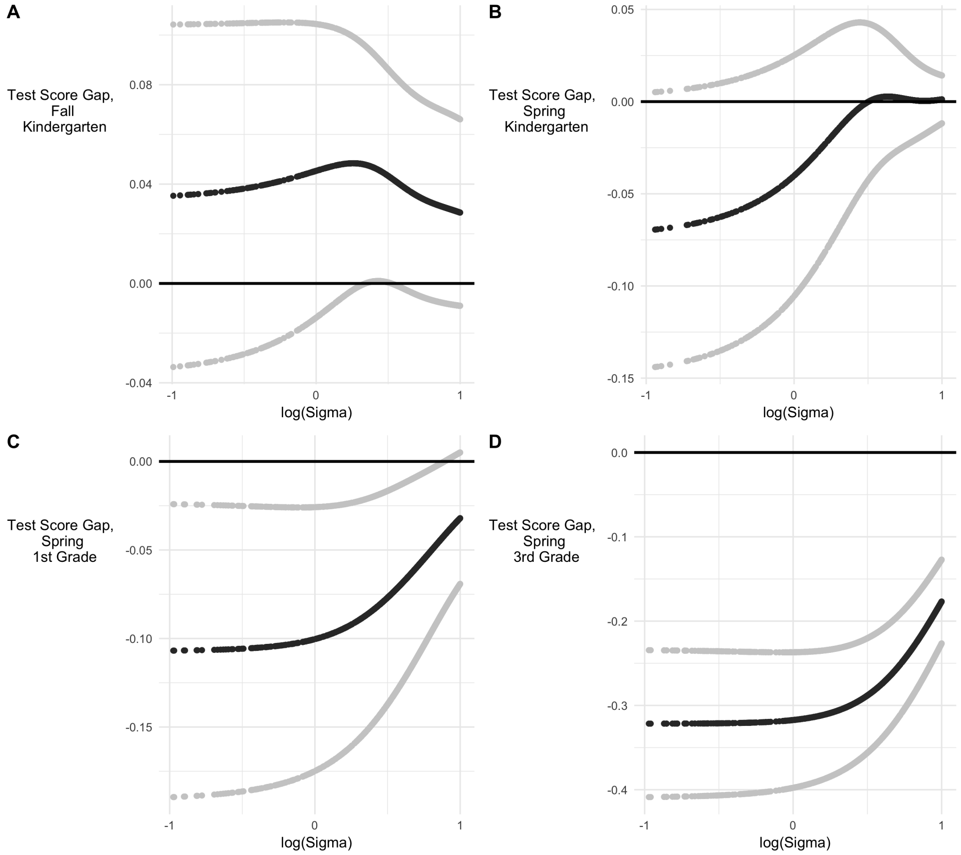
The analysis by Bond and Lang (2013) provides two additional opportunities to test the validity of the method developed in this paper. Both of these replicate the core findings of Bond and Lang (2013) and therefore add to the credibility of the methodology developed in this paper.

The first illustration is to perform the same analysis as shown in main text, but control for socioeconomic factors that may explain some of the early elementary racial test score gap. This more closely examines the result from Fryer and Levitt (2004) suggesting that the black-white test score gap in kindergarten through third grade can be explained by a relatively small number of socioeconomic factors. Bond and Lang (2013) examine the robustness of this finding in Table 5 of their paper. They find that, when controlling for the same socioeconomic factors as Fryer and Levitt (2004), the test score gap in the fall of kindergarten is robust to reasonable transformations. This result is largely replicated in Panel A of Figure A7. Although the racial test score gap is not statistically significant, in the fall of kindergarten, the coefficient estimate is largely robust to alternative monotonic increasing transformations. In contrast, the racial test score gap in third grade depends on the transformation of the test score scale. Bond and Lang (2013) report a range of between a 0.17 and a 0.31 standard deviation test score gap in the spring of third grade. This finding is again replicated in Panel D of Figure A7 where the test score gap ranges from between 0.17 and 0.32 standard deviations for alternative transformations.

The second illustration uses an additional data source for early education test scores: the Peabody Individual Achievement Test (PIAT). These test score gaps are calculated without the inclusion of additional socioeconomic control variables and are presented in Table 3 of Bond and Lang (2013). The authors report that the gap in kindergarten varies between a statistically insignificant 0.05 and a statistically significant 0.24 standard deviations. Panel A of Figure A8 largely replicates this result, with a range of the gap between a statistically insignificant 0.06 and a statistically significant 0.25 standard deviations in kindergarten. In third grade, Bond and Lang (2013) report the racial test score gap ranging between a statistically insignificant 0.06 and a statistically significant 0.63 standard deviations. Panel D of Figure A8, for the most part, replicates this finding with a black-white test score gap ranging between 0.15 and 0.61 standard deviations.

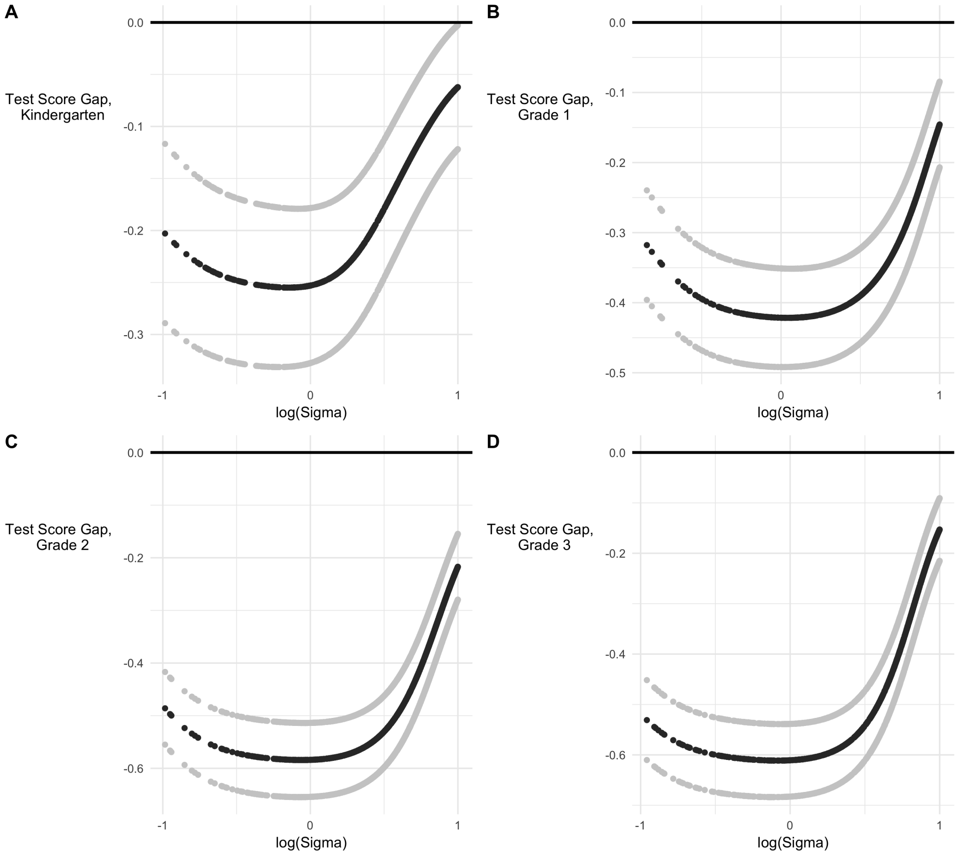
In addition to similar results presented in the main text, these results lend credence to the credibility of the methodology developed in this paper. This is highlighted by the fact that the results from column 2 and 3 in Table 5 of Bond and Lang (2013) can be found within the range of results shown in Figure A7. Additionally, the results from Table 3 in Bond and Lang (2013) are for the most part replicated in Figure A8.

Figure A6. Estimated Effect Sets for Bond and Lang (2013) ECLS Test Score Gap with Controls



Notes: The dark lines represent the point estimates for a given specification with the corresponding sigma value. Lighter lines represent 95% confidence interval calculated with robust standard errors. Each panel refers to a test scores from different grades as shown in Table 5 of Bond and Lang (2013). Panel A refers to the test gap in the fall of kindergarten, panel B the spring of kindergarten, panel C the spring of first grade, and panel D the spring of third grade.

Figure A7. Estimated Effect Sets for Bond and Lang (2013)
PIAT Test Score Gap

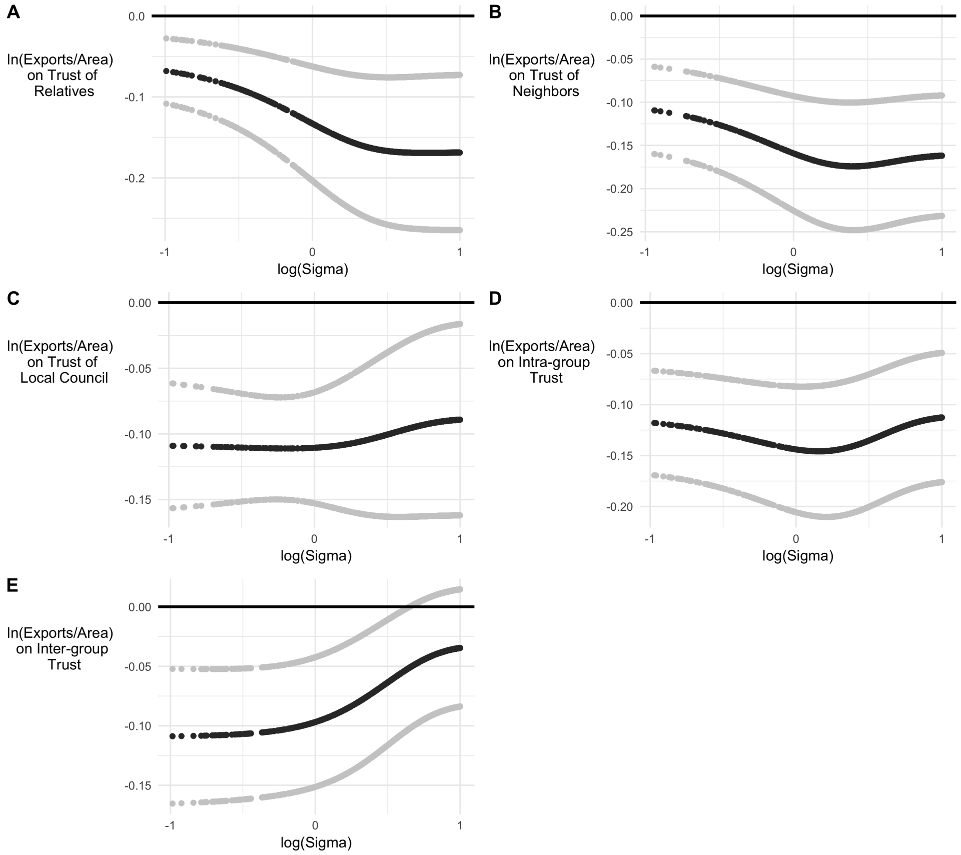


Notes: The dark lines represent the point estimates for a given specification with the corresponding sigma value. Lighter lines represent 95% confidence interval calculated with robust standard errors. Each panel refers to a test scores from different grades as shown in Table 3 of Bond and Lang (2013). Panel A refers to the test gap in kindergarten, panel B refers to grade 1, panel C to grade2, and panel D to grade 3.

A5. OLS RESULTS FROM NUNN AND WANTCHEKON (2011)

Before showing instrumental variable results, Nunn and Wantchekon (2011) perform an OLS regression testing the relationship between the slave trade and present day trust in sub-Saharan Africa. These results are shown in Table 2 of Nunn and Wantchekon (2011). Although the OLS results could be biased by omitted variables, it may be informative to examine the robustness of these results to alternative monotonic increasing transformations. These results are shown in Figure A9. In general the core finding from the instrumental variable results holds with the OLS results as well. Namely, that the empirical findings are largely robust, in terms of effect size and statistical significance, to all reasonable transformations.

Figure A8. Estimated Effect Sets for OLS Estimates from Nunn and Wantchekon (2011)



Notes: The dark lines represent the point estimates for a given specification with the corresponding sigma value. Lighter lines represent 95% confidence interval calculated with standard errors clustered by ethnicity. Each panel refers to a different specifications used in Table 2 of Nunn and Wantchekon (2011). Panel A refers to column (1) with the dependent variable trust of relatives. Panel B refers to column (2) with the dependent variable trust of neighbors. Panel C refers to column (3) with the dependent variable trust of local council. Panel D refers to column (4) with the dependent variable intra-group trust. Finally, panel E refers to column (5) with the dependent variable inter-group trust.

A6. COMPARING MARGINAL EFFECTS ACROSS TRANSFORMATIONS

Comparing interpretations of the marginal effects calculated from transformed ordinal scales to original (or linear) ordinal scales may be challenging. The transformation of the dependent variable sometimes changes the interpretation of regression coefficients. For example, in some specifications taking the natural log of the dependent variable allows regression coefficients to be interpreted as percentage changes. Therefore, this may complicate the comparison of regression coefficients across monotonic increasing transformations. One way to overcome this challenge is to manually calculate the marginal effect (see, e.g., Cameron and Trivedi 2010) and express the marginal effect in terms of the original linear ordinal scale. Table A4 shows both the raw marginal effects (in row i of each panel) and the marginal effects expressed in terms of the original linear scale (in row ii of each panel) for each of the coefficients of interest in Aghion et al. (2016). Column (1) shows marginal effects given $\sigma = 1$, that is the transformation is linear. Columns (2) and (3) show marginal effects at the extremes of the domain of σ, 0.1 and 10, respectively.

Panel A shows that when expressing the marginal effects in terms of the original linear scale, the effect size still ranges from close to zero to an effect size that is considerably larger than reported by Aghion et al. (2016). Therefore, the lack of robustness of the effect size persists even when converting marginal effects, calculated with different σ values, back into terms of the linear zero through ten scale. The other panels also show considerable variation in the marginal effects, for discrete values of σ, even when expressed in terms of the original linear ordinal scale.

Table A4. Marginal Effects in Terms of Transformed and Linear SWB Scales

	(1)	(2)	(3)
Dep. Variable: Gallup SWB	log (σ) = 0	log (σ) = -1	log (σ) = 1
A: Prediction1, Job Turnover			
(i) Raw Marginal Effect	0.521	-0.021	0.950***
	(0.237)	(0.088)	(0.221)
(ii) Marginal Effect on Linear Scale	0.521	-0.139	0.701***
	(0.237)	(0.548)	(0.158)
Additional MSA controls	No	No	No
Individual controls	Yes	Yes	Yes
Year and month fixed effects	Yes	Yes	Yes
Obs.	556,300	556,300	556,300
B: Prediction 2, Job Creation			
(i) Raw Marginal Effect	1.274***	0.131	1.549***
	(0.445)	(0.168)	(0.404)
(ii) Marginal Effect on Original Scale	1.274***	0.847	1.137***
	(0.436)	(1.135)	(0.289)
Additional MSA controls	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes
Year and month fixed effects	Yes	Yes	Yes
Obs.	461,054	461,054	461,054
C: Prediction 2, Job Destruction			
(i) Raw Marginal Effect	-0.702**	-0.245*	-0.043
	(0.306)	(0.142)	(0.306)
(ii) Marginal Effect on Original Scale	-0.702**	-1.584*	-0.031
	(0.326)	(0.926)	(0.237)
Additional MSA controls	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes
Year and month fixed effects	Yes	Yes	Yes
Obs.	461,054	461,054	461,054
D: Prediction 3, Job Turnover × UI Generosity			
(i) Raw Marginal Effect	0.675**	0.322**	0.284
	(0.310)	(0.129)	(0.297)
(ii) Marginal Effect on Original Scale	0.675**	2.086**	0.209
	(0.315)	(0.829)	(0.222)
Additional MSA controls	No	No	No
Individual controls	Yes	Yes	Yes
Year and month fixed effects	Yes	Yes	Yes
Obs.	556,300	556,300	556,300
E: Prediction 3, Job Destruction × UI Generosity			
(i) Raw Marginal Effect	0.620*	0.388***	0.248
	(0.329)	(0.148)	(0.322)
(ii) Marginal Effect on Original Scale	0.620*	2.511***	0.183
	(0.317)	(0.969)	(0.249)
Additional MSA controls	No	No	No
Individual controls	Yes	Yes	Yes
Year and month fixed effects	Yes	Yes	Yes
Obs.	556,300	556,300	556,300

Notes: Within each panel, row (i) shows the raw marginal effect given the discrete σ value and row (ii) shows the marginal effect given the discrete σ value that is transformed back into terms of the original zero through ten linear ordinal SWB scale. Standard errors are shown in parentheses. In rows (i) standard errors are calculated by clustering at the MSA level. In rows (ii) standard errors are bootstrapped with 1,000 replications. *** p<0.01, ** p<0.05, * p<0.1.

ONLINE SUPPLMENT REFERENCES

Aghion, P., Akcigit, U., Deaton, A., and Roulet, A. **(2016)** “Creative Destruction and Subjective Well-Being” *American Economic Review*, 106 (12) pp. 3869-3897.

Alatas, V., Banerjee, A., Hanna, R., Olken, B., and Tobias, J. **(2012)** “Targeting the Poor: Evidence from a Field Experiment in Indonesia” *American Economic Review*, vol. 102 (4), pp. 1206-1240.

Ashraf, N., Field, E., and Lee, J. **(2014)** “Household Bargaining and Excell Fertility: An Experimental Study in Zambia” *American Economic Review*, vol. 104 (7), pp. 2210-2237.

Bandiera, O., Burgess, R., Das, N., Gulesci, S., Rasul, I., Sulaiman, M. **(2017)** “Labor Markets and Poverty in Village Economies” *Quarterly Journal of Economics*, vol. 132 (2), pp. 811-870.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Pariente, W., Shapiro, J., Thuysbaert, B., and Udry, C. **(2015)** “A multifaceted program causes lasting progress for the very poor: Evidence from six countries” *Science*, vol. 348, issue 6236.

Bertrand **(2013)** “Career, Family, and the Well-Being of College-Educated Women” *American Economic Review: Papers & Proceedings*, vol. 103 (3), pp. 244-250.

Bianchi **(2012)** “Financial Development, Entrepreneurship, and Job Satisfaction” *Review of Economics and Statistics*, vol. 94 (1), pp. 273-286.

Bloom, N., Liang, J., Roberts, J. and Ying, Z.J. **(2015)** “Does Working from Home Work? Evidence from a Chinese Experiment” *Quarterly Journal of Economics*, vol. 130 (1), pp. 165-218.

Bloom, N., Propper, C., Seiler, S., and Van Reenen, J. **(2015)** “The Impact of Competition on Management Quality: Evidence from Public Hospitals” *Review of Economic Studies*, vol. 82 (2), pp. 457-489.

Bond, T. and Lang, K. **(2014)** “The Sad Truth About Happiness Scales” *NBER Working Paper*, No. 19950.

Bryson and MacKerron **(2017)** “Are You Happy While You Work?” *Economic Journal*, vol. 127 (599), pp. 106-125.

Cameron, A.C. and Trivedi, P.K. **(2010)** *Microeconomics Using Stata*, Revised Edition, Stata Press. College Station, Texas.

Card, D., Mas, A., Moretti, E., and Saez, E. **(2012)** “Inequality at Work: The Effect of Peer Salaries on Job Satisfaction” *American Economic Reveiw*, vol. 102 (6), pp. 2981-3003.

Clark, A., Frijters, P., and Shields, M. **(2008)** “Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles” *Journal of Economic Literature*, vol. 46 (1), pp. 95-144.

Clark, A., D’Ambrosio, Ghislandi, S. **(2016)** “Adaptation to Poverty in Long-Run Panel Data” *Review of Economics and Statistics*, vol. 98 (3), pp. 591-600.

De Neve, J., Ward, G., De Keulenaer, F., Van Landeghem, B., Kavetsos, G., and Norton, M.I. **(2018)** “The Asymmetric Experience of Positive and Negative Economic Growth: Global Evidence Using Subjective Well-Being Data” *Review of Economics and Statistics*, vol. 100 (2), pp. 362-375.

Deaton **(2018)** “What do self-reports of wellbeing say about life-cycle theory and policy?” *Journal of Public Economics*, vol. 162, pp. 18-25.

Di Tilla, R., MacCulloch, R.J., and Oswald, A., **(2001)** “Preferences of Inflation and

Unemployment: Evidence from Surveys of Happiness" *American Economic Review*, vol. 91 (1), pp. 335-341.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. **(2012)** "The Intergenerational Transmission of Risk and Trust Attitudes" *Review of Economic Studies*, vol. 79 (2), pp. 645-677.

Dustmann and Fasani **(2016)** "The Effect of Local Area Crime on Mental Health" *Economic Journal*, vol. 126 (593), pp. 978-1017.

Frijters, P., Johnston, D.W., and Shields, M.A. **(2014)** "Does Childhood Predict Adult Life Satisfaction? Evidence from British Cohort Surveys" *Economic Journal*, vol. 124 (580), pp. F688-F719.

Fryer, R. and Levitt, S. **(2004)** "Understanding the Black-White Test Score Gap in the First Two Years of School" *The Review of Economics and Statistics*, 86 (2) pp. 447-464.

Glewwe, P. , Ross, P.H., and Wydick, B. **(2018)** "Developing Hope among Impoverished Children: Using Child Self-Portraits to Measure Poverty Program Impacts" *Journal of Human Resources*, vol. 53 (2), pp. 330-355.

Haushofer and Shapiro **(2016)** "The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya" *Quarterly Journal of Economics*, vol. 131 (4), pp. 1973-2042.

Kaiser, C. and Vendrik, C.M. **(2019)** "How threatening are transformations of reported happiness to subjective wellbeing research?" *SocArXiv Paper.*

Krueger and Mueller **(2012)** "Time Use, Emotional Well-Being, and Unemployment: Evidence from Longitudinal Data" *American Economic Review: Papers & Proceedings*, vol. 102 (3), pp. 594-599.

Lachowska, M. **(2017)** "The Effect of Income on Subjective Well-Being: Evidence from the 2008 Economic Stimulus Tax Rebates" *Journal of Human Resources*, vol. 52(2), pp. 374-417.

Layard, R., Clark, A., Cornaglia, F. Powdthavee, N. and Vernoit, J. **(2014)** "What Predicts a Successful Life? A Life-course Model of Well-Being" *Economic Journal*, vol. 124 (580), pp. F720-F738.

Milligan and Stabile **(2011)** "Do Child Tax Benefits Affect the Well-Being of Children? Evidence from Canadian Child Benefit Expansions" *American Economic Journal: Economic Policy*, vol. 3 (3), pp. 175-205.

Moscona, J., Nunn, N., and Robinson, J. **(2017)** "Keeping It in the Family: Lineage Organization and the Scope of Trust in Sub-Saharan Africa" *American Economic Review: Papers & Proceedings*, vol. 107 (5), pp. 565-571.

Nunn, N. and Wantchekon, L. **(2011)** "The Slave Trade and the Origins of Mistrust in Africa" *American Economic Review* 101 (7) pp. 3221-3252.

Oswald and Powdthavee **(2008)** "Does Happiness Adapt? A Longitudinal Study of Disability with Implications for Economists and Judges" *Journal of Public Economics*, vol. 92 (5-6), pp. 1061-1077.

Oswald and Wu **(2011)** "Well-Being Across America" *Review of Economics and Statistics*, vol. 93 (4), pp. 1118-1134.

Schechter **(2007)** "Theft, Gift-Giving, and Trustworthiness: Honesty Is Its Own Reward in Rural Paraguay" *American Economic Review*, vol 97 (5), pp. 1560-1582.

Steptoe, A., Deaton, A., and Stone, A. **(2015)** "Subjective wellbeing, health, and aging" *The Lancet*, vol. 385, issue 9968, pp. 640-648.

Wunder, C., Wiencierz, A., Schwarze, J. and Kuchenhoff, H. **(2013)** "Well-Being over the Life Span: Semiparametric Evidence from British and German Longitudinal Data" *Review of Economics and Statistics*, vol. 95 (1), pp. 154-167.

Schröder, C. and Yitzhaki, S. **(2017)** "Revisiting the evidence for cardinal treatment of ordinal variables" *European Economic Review*, 92 pp. 337-358.