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

Jeffrey R. Bloem†

December 20, 2018

Abstract

Many researchers use an ordinal scale to quantitatively measure and empirically analyze concepts. Theoretically, valid empirical estimates are robust in sign to any monotonic increasing transformation of the ordinal scale. Extending this theoretical criterion, I develop a method for testing how much the cardinal treatment of ordinal variables matters for any empirical specification. I apply this method to results from three papers: Aghion et al. (2016) on creative destruction and subjective well-being, Nunn and Wantchekon (2011) on the slave trade and trust in sub-Saharan Africa, and Bond and Lang (2013) on the fragility of the racial test score gap. This leads to three insights. (i) Failing the theoretical criterion may not be a serious problem in practice if only implausible transformations change empirical results. Passing the theoretical criterion does not imply that empirical results are robust because (ii) the size of coefficients and (iii) statistical inference could meaningfully change with such transformations.

Keywords: ordinal variables, partial identification, linear regression

*I am grateful for constructive feedback and guidance to Joe Ritter, Marc Bellemare, Paul Glewwe, Jason Kerwin, Andrew Oswald, Jane Sumner, Soren Anderson, Khandher Wahedur Rahman, Michael Bloem, Kenn Chua, Anuar Bechara Bitar, Natalia Ordaz Reynoso, and other conference and seminar participants. I also thank Gallup Inc. and all authors of the papers under investigation in this study for writing clear and well-documented code. All errors are my own.

†PhD Student, Department of Applied Economics, University of Minnesota, bloem023@umn.edu

1 Introduction

Concerns about the cardinal treatment of ordinal dependent variables are well-known. Consider an ordinal scale measuring "satisfaction," with the following four categories: "Very Dissatisfied," "Dissatisfied," "Satisfied," and "Very Satisfied." Suppose there are two groups of people, A and B, each consisting of two individuals. Group A has one person who is "Very Dissatisfied" and another who is "Very Satisfied." Group B has one person who is "Dissatisfied" and another who is "Satisfied." Which group is more satisfied? The answer depends on the numerical values assigned to the response categories. It may seem reasonable to assign the integers zero, one, two, and three to each of the four categories. In this case, the groups are equally satisfied, with an average score of 1.5. Similar to utility functions, however, ordinal variables provide information about the rank of a specific concept, rather than representing a known or fixed interval. As such, any set of numerical values that preserve the ordering of the satisfaction scale is also potentially valid. Therefore another reasonable set of numerical values could be: zero, 1.75, 2.5, and three. In this case, group B is more satisfied. Yet, another reasonable set of values could be: zero, 0.5, 1.25, and three. In this case group A is more satisfied.

In many cases, the most appropriate methodological approach to limit such concerns is to use an ordered response model.[1] These models, however, are often omitted or the results are less preferred compared to results from a more simple and straightforward linear regression (see, for example, Nunn and Wantchekon 2011; Stevenson and Wolfers 2013; Aghion et al. 2016; Bryson and MacKerron 2017; Deaton 2018; as well as numerous other examples listed in Table A1 in the Appendix). Justification for using a linear regression with an ordinal dependent variable—despite the well-known concerns—often include the incidental parameter problem (Neyman and Scott 1948; Heckman 1981; Lancaster 2000; Riedl and Geishecker 2014) or the use of a more sophisticated identification strategy.

Theoretically, the choice of statistical model matters. In practice, however, this choice is often considered inconsequential. In an influential paper within the subjective well-being (SWB) and happiness literature, Ferrer-i-Carbonell and Frijters (2004) find that, "[...] assuming ordinality or cardinality of scores makes little difference".[2] Responding directly to this conclusion, Schröder and

[1] Statistical models designed to appropriately handle an ordered dependent variable originated in the biometrics literature (Aitchison and Silvey 1957; Snell 1964). Use of such statistical models in the social sciences followed with McKelvey and Zavoina (1975).

[2] An example of the influence of Ferrer-i-Carbonell and Frijters (2004) can be found in Wang and Zhou (2018):"[...]

Yitzhaki (2017) argue that simply showing robustness of a single scale across statistical models is insufficient for validating the cardinal treatment of ordinal variables. Bond and Lang (forthcoming), in their paper entitled, "The Sad Truth About Happiness Scales", simply state that the conclusion of Ferrer-i-Carbonell and Frijters (2004) is "false".

Most fundamentally, concerns associated with the cardinal treatment of an ordinal dependent variable can be characterized as a missing information problem. That is, the researcher does not know the functional form of the latent response function characterizing the relationship between the ordinal scale and the latent concept (Oswald 2008). Therefore, as demonstrated by Schröder and Yitzhaki (2017) and Bond and Lang (forthcoming), the valid cardinal treatment of ordinal variables must be robust to monotonic increasing transformations of the ordinal scale.

Clearly the cardinal treatment of ordinal variables matters, but how much does it change empirical findings? In this paper I develop a method for testing how much the cardinal treatment of ordinal variables matters for any empirical specification. I consider this method as similar to partial identification (Manski 2003; Tamer 2010). That is, with an ordinal dependent variable, additional assumptions about the cardinal properties of the ordinal scale are necessary to point-identify coefficient estimates. The method I describe allows for a test of robustness of coefficient estimates to a range of plausible monotonic increasing transformations of the ordinal dependent variable. Without these additional assumptions, valid empirical estimates are bounded between the most extreme results within the range of monotonic increasing transformations. I apply this method to results from three papers: Aghion et al. (2016) on creative destruction and subjective well-being, Nunn and Wantchekon (2011) on the slave trade and trust in sub-Saharan Africa, and Bond and Lang (2013) on the fragility of the black-white test score gap.[3]

The method described in this paper first limits the universe of monotonic increasing transformations to be defined by a parameterized function representing all plausible transformations. Next, I use Monte Carlo simulations to investigate the robustness of, and generate bounds around, existing

we run simple ordinary least squares (OLS) regressions [...] though the happiness scores are only integers ranging from 1 to 5; as Ferrer-i-Carbonell and Frijters (2004) argue, whether happiness scores are treated as ordinal or cardinal does not matter" (pp. 831). Another example can be found in Wunder et al. (2013): "Ferrer-i-Carbonell and Frijters (2004) show that assuming ordinality or cardinality of satisfaction scores makes little difference to the results of regression analysis. Hence, we are able to apply econometric models designed for continuous response variables" (pp. 159).

[3]Since Bond and Lang (2013) already establish the "fragility" of the black-white test score gap in kindergarten through grade three, the investigation of these results act as a validity test on the method developed in this paper.

empirical results. This method provides insight into the robustness of both the coefficient estimates and statistical significance for any multi-variate empirical specification. That is, this method tests robustness of empirical results when relaxing arbitrary assumptions about the implicit cardinalization of the ordinal scale. The goal of this method is to gain consensus about plausible restrictions on the ordinal scale, so that it may be possible to conclude that no plausible transformations will meaningfully change an empirical finding. In short, this method tests the validity of the cardinal treatment of ordinal variables.

The contribution of this paper is threefold. First, this paper clarifies three empirical points that extend existing theoretical insights (Schröder and Yitzhaki 2017 and Bond and Lang forthcoming). (i) Failing existing theoretical tests for the valid cardinal treatment of ordinal variables may not be a serious problem in practice because it may be the case that only extreme monotonic increasing transformations substantially change empirical results. Although these important contributions are theoretically insightful, they do not show if a given monotonic increasing transformation is plausible.[4] (ii) Passing existing theoretical tests does not necessarily imply that empirical results are robust because the size of estimated coefficients could change dramatically for monotonic increasing transformations. (iii) Passing existing theoretical tests does not imply that the statistical significance of results is robust to monotonic increasing transformations of the ordinal scale, even if the size of the coefficient estimates are relatively robust.

Second, the method developed in this paper generalizes the work of Schröder and Yitzhaki (2017) to cases using econometric specifications with multiple right hand side variables. As I will discuss in more detail in Section 2.2, the sufficient conditions developed by Schröder and Ytizhaki (2017) apply only to cases either comparing means between two groups or performing simple bivariate regression analysis. Obviously, since most econometric specifications include more than one covariate, this is quite limiting. Applying the method developed in this paper to the results from three existing papers show that the inclusion of additional covariates in a given econometric specification influences robustness of results to monotonic increasing transformations. That is, it is possible to fail the sufficient conditions developed by Schröder and Yitzhaki's (2017) and yet, when the full set of covariates are included in the empirical specification, the results can be robust to plausible monotonic increasing transformations.

[4]Typically, when researchers transform variables they make some statement about the plausibility of such transformations. For example, see Bond and Lang (2013). In Section 3.2 of this paper, I present a method in which researchers can assess plausibility in a given empirical context.

Finally, the most recent work on the valid statistical treatment of ordinal variables, by Schröder and Yitzhaki (2017) and Bond and Lang (forthcoming), focus on ordinal variables measuring subjective well-being or happiness. These insights also apply to any variable that measures a latent concept using an ordinal scale. Therefore, concepts such as "satisfaction" (Frijters, et al. 2004; Clark and Oswald 1996; Ritter and Anker 2002; Luechinger et al. 2010), "trust" (Nunn and Wantchekon 2011; Putnam 2001), "hope" (Bloem et al. 2018; Glewwe et al. 2018), various measures of mental well-being and personality traits (Borghans et al. 2008; Baird et al. 2013; Cornaglia et al. 2014), measures of "affect" (Krueger et al. 2009; Krueger 2017), measures of "quality"—of political institutions (Acemoglu et al. 2001), for example—and even standardized test scores (Bond and Lang 2013; Glewwe 1997; Jacob and Rothstein 2016; Lang 2010; Schröder and Yitzhaki 2016) are all measured with an ordinal scale. This paper extends existing theoretical insights and empirical investigation to an application using a dependent variable measuring subjective well-being (Aghion et al. 2016), trust (Nunn and Wantchekon 2011), and test scores (Bond and Lang 2013).

The next section briefly describes the theoretical framework motivating this research. In that section I outline the potential theoretical consequences of the cardinal treatment of ordinal variables and summarize the sufficient conditions developed by Schröder and Yizhaki (2017). Section 3 provides information on the empirical specifications used Aghion et al. (2016), Nunn and Wantchekon (2011), and Bond and Lang (2013). Additionally, this section introduces the methodology for the Monte Carlo simulations used in this paper. Section 4 presents the simulation results from each of the three empirical investigations and discusses these findings. Finally, section 5 concludes.

2 Theoretical Framework

Although ordinal variables are used to measure a variety of concepts with no natural quantitative unit of measure, I proceed here by briefly discussing the SWB literature specifically. As discussed below, the following also applies to other ordinal variables, such as happiness, satisfaction, trust, measures of quality, standardized test scores, and other concepts that require measurement via the use of an ordinal variable. Nevertheless, it is helpful to draw a connection to the familiar concept of utility theory and the relevant implications for econometric analysis (Greene 2012; Becker and Kennedy 1992).

Suppose an individual's well-being is characterized by the following underlying relationship:

$$Y^* = X'\beta + \epsilon \tag{1}$$

In this characterization, Y^* is the unobserved latent well-being of the individual. The vector X represents observable variables that define an individual's well-being and β is a vector of regression coefficients. Since, Y^* cannot be directly observed, subjective well-being, Y, is measured via an ordinal variable with the various values of μ corresponding to threshold points on the ordinal scale:

$$Y = \begin{cases} 0 & \text{if} \quad Y^* \leq 0, \\ 1 & \text{if} \quad 0 < Y^* \leq \mu_1, \\ 2 & \text{if} \quad \mu_1 < Y^* \leq \mu_2 \\ \vdots & \\ N & \text{if} \quad \mu_{N-1} < Y^* \end{cases} \tag{2}$$

The problem is the values of μ are unknown. Estimating equation (1) using OLS with the observed ordinal scale of Y as the dependent variable, implicitly assumes that the values of Y have known and fixed intervals—an arbitrary assumption. Thus, OLS assumes the ordinal variable measuring well-being is cardinal. This detail is often obscured in the SWB and happiness literature.

The "paradoxical" finding of Easterlin (1974), that higher levels of a country's GDP per capita are not correlated with higher measures of a country's average SWB, started a debate that lasted decades. Since then many find results that contradict the "Easterlin paradox" (Stevenson and Wolfers 2008; Deaton 2008). By now a generally accepted understanding is that there is a positive, albeit diminishing, return to SWB from income (Dolan et al. 2008; Clark et al. 2008). Easterlin has since clarified his finding suggesting that the relative income effect dominates the absolute income effect (Easterlin 1995). This would potentially explain why cross sectional data show that wealthier individuals within a society report higher measures of SWB, but average SWB levels remain constant as all members of a society become wealthier. Others argue that there is a satiation point, where beyond a certain income threshold more income is not related to measures of SWB (Diener and Seligman 2004; Clark et al. 2008; Di Tella and MacCulloch 2008). Studies by Deaton (2008) and Stevenson and Wolfers (2008), find however that while the relationship between income and happiness follows a linear-log relationship, there is no evidence of a satiation point.

2.1 The Reporting Function

Oswald (2008) critiques all of these conclusions by arguing that the literature on SWB has yet to establish the shape of the function relating reported SWB to actual well-being. While introducing the concept of the reporting function, as the function which defines the relationship between subjective feelings to objective reality, Oswald (2008) states:

> *As an example, imagine that there is constant marginal utility of income, but that people as they feel cheerier, mark themselves happier on a questionnaire scale in a way in which they are intrinsically reluctant to approach the upper possible level on the questionnaire form. Then the reporting function itself is curved. In this case, we will have the illusion [...] that true diminishing marginal utility of income has been established empirically.*

If this is the case, researchers hands are tied when making statements about the relationship between well-being and income, since little is known about the shape of the reporting function that respondents use to translate actual well-being to measures of SWB. An argument mirroring Oswald (2008) is easily applied to other ordinal variables, such as happiness, satisfaction, trust, and measures of quality. Standardized test scores perhaps require a brief explanation. As discussed in Bond and Lang (2013), standardized test scores may not have a known or fixed interval between values.[5] Consider a simple case where a test score simply assigns values based on the number of questions answered correctly by each student. If some questions are more difficult than others, then assuming a fixed interval or cardinal scale may not be valid. Since test scores approximate student "learning," answering difficult questions correctly may signal a larger marginal gain in learning than answering the easier questions correctly. The reporting function for test scores, therefore, would define the relationship between actual student learning to performance on a test.

Since researchers do not know the form of the reporting function, ordinal variables present only information about the relative rank of values, and the interval between values is unknown. Much of the research using ordinal variables acknowledges that interpretation of estimates are made under the assumption that the reporting function is linear. This is unsatisfactory since a linear reporting

[5] A considerable amount of work by psychometricians aim at confronting the issue that test scores are measured on an ordinal, rather than an interval or cardinal, scale (see Stevens 1946; Thorndike 1966; Schröder and Yitzhaki 2016). Methods, such as item response theory (IRT), lead some to believe that test scores can be considered to be measured on a cardinal or interval scale (see Baker 2001). However, this is still an open area of research and the vast majority of researchers consider test scores to be measured on an ordinal scale.

function is only one of infinitely many theoretically possible reporting functions and the likelihood this assumption is valid is low.[6] Additionally, many studies show robustness of results using an ordinal response model and almost always state that results are qualitatively similar (for examples, see Ferrer-i-Carbonell and Frijters 2004; Nunn and Wantchekon 2011; Stevenson and Wolfers 2013). Schröder and Yitzhaki (2017) point out, however, that although showing robustness to the use of ordinal response models is instructive in investigating validity of the use of econometric models to a given reporting scale, these tests do little in showing robustness of results to monotonic increasing transformations of the reporting scale. These critiques shed considerable doubt on the current empirical literature analyzing ordinal variables.

2.2 Sufficient Conditions

Testing for the robustness of monotonic increasing transformations is complicated by the fact that there are an infinite number of ways to transform a variable. Precisely due to this reality, Schröder and Yitzhaki (2017) derive two theoretical conditions under which the cardinal treatment of ordinal variables is permitted. The first condition refers to the permissibility of comparing means of ordinal variables between groups and the second condition refers to the valid use of OLS regression models. In this sub-section I only summarize the details of these sufficient conditions, the interested reader should specifically reference the work of Schröder and Yitzhaki (2017) and Bond and Lang (forthcoming).

These conditions draw from the literature on stochastic dominance (Hadar and Russel 1969). In particular, the first condition states that the mean of one group is larger than another mutually exclusive group if and only if the former first-order stochastically dominates the latter. This condition implies that if the cumulative distribution functions of each group intersect, then it is possible to find a monotonic transformation of the ordinal scale that will change which group has a larger mean.

The second condition introduces the concept of the line of independence minus absolute concentration (LMA) curve. As the name suggests, the LMA curve takes the difference between two curves: the line of independence and the absolution concentration curve. The line of independence

[6] It is worth noting that an additional complication is cross-sectional heterogeneity in the reporting function. Although this point is indeed important, I abstract from this issue in the present study. Survey methodology (e.g. anchoring vignettes) typically help control for issues of reporting function heterogeneity (see King et al. 2004; King and Wand 2007; Hopkins and King 2010), however not all ordinal variables are measured via a survey.

(LoI) is defined as the weighted mean of the dependent variable, Y, multiplied by the cumulative distribution, $F(X)$, of the explanatory variable, X.

$$LoI_i = \big[\frac{1}{N}\sum_{i=1}^{N} w_i Y_i\big] \times F(X_i) \tag{3}$$

The absolute concentration curve (ACC) is defined as the cumulated product of the dependent variable, Y, and the frequency weight, w, divided by the sum of the frequency weights.

$$ACC_1 = \frac{Y_1 w_1}{\sum_{i=1}^{N} w_i}, \quad ACC_2 = \frac{Y_1 w_1 + Y_2 w_2}{\sum_{i=1}^{N} w_i}, \dots, ACC_N = \frac{Y_1 w_1 + \dots + Y_N w_N}{\sum_{i=1}^{N} w_i} \tag{4}$$

In both equations (3) and (4), $Y_1 \leq Y_2 \leq \dots \leq Y_N$. Equation (4) can be interpreted as the generalized Lorenz curve. Finally, the LMA curve is the difference between these two lines.[7]

$$LMA_i = LoI_i - ACC_i \quad \forall\ N \tag{5}$$

The LMA curve is related to the concept of second-order stochastic dominance and the absolute Lorenz curve. Recall that second-order stochastic dominance states that if two Lorenz curves cross, then it is impossible to determine which of two mutually exclusive groups second-order stochastically dominate the other. Therefore the second condition, derived by Schröder and Yitzhaki (2017), states that if the LMA curve intersects the horizontal axis, then the absolute Lorenz curves intersect and there is some monotonic increasing transformation that will change the sign of the OLS regression coefficient.[8] If the absolute Lorenz curves do not cross then there does not exist a monotonic increasing transformation that can change the sign, however, the magnitudes and statistical significance of the coefficient estimates could potentially meaningfully change. Note the two Lorenz curves, in this explanation, refer to one representing the "raw" ordinal values and the second representing the "transformed" ordinal values. For the bulk of this paper, I will focus on the second condition derived by Schröder and Yizhaki (2017).[9]

[7] A user-written STATA program is available which allows researchers to easily generate LMA curves (Schaffer 2015).

[8] Formal proofs of this condition can be found in Yitzhaki and Schechtman (2012, 2013) and the Appendix of Schröder and Yitzhaki (2017).

[9] This does not imply that an investigation using the first condition is not worthwhile, as each of the theoretical predictions in Aghion et al. (2016) are tested, in part, by finding the average of SWB within each metropolitan statistical area (MSA) and running MSA-level analysis. Additionally, cross-country analysis of the effect of income on SWB (Easterlin 1974; Deaton 2008; Stevenson and Wolfers 2008) requires an average SWB measure for each country. In these cases, the valid cardinal statistical treatment of ordinal variables must pass both conditions derived

Stated more formally, and as explained in Schröder and Yitzhaki (2017), the logic of the second condition is as follows. Consider two simple OLS regression coefficients, α_1 and β_1, from two separate specifications. One uses the raw ordinal variable, Y, and the other uses the transformed ordinal variable, $T(Y)$. If the LMA curve of Y, with respect to X, intersects the horizontal axis, it is possible to find a monotonic increasing transformation of the dependent variable, Y, $T(Y)$, that can change the sign of the OLS regression coefficient. That is, if α_1 is positive (negative) then β_1 will be negative (positive). This implies:

$$\frac{\alpha_1}{\beta_1} = \frac{\frac{Cov(X,Y)}{Var(X)}}{\frac{Cov(X,T(Y))}{Var(X)}} < 0 \tag{6}$$

Important questions remain.

1. Suppose there exists a transformation $T(Y)$ that allows equation (6) to hold, how dramatic and realistic is this transformation?

2. On the other hand, suppose there does not exist a transformation $T(Y)$ that allows equation (6) to hold, does the magnitude of the coefficient meaningfully change? Does the economic significance or policy implication change? How is statistical significance affected by these transformations?

3. Moreover, equation (1) displayed an analytical example where there are multiple covariates and equation (3) through (6) only consider one X variable. This raises a final question. Since existing theoretical tests—by both Schröder and Ytizhaki (2017) and Bond and Lang (forthcoming)—only focus on simple bivariate examples, how are researchers to test robustness of more complicated specifications to monotonic increasing transformations of the ordinal scale?

These are the questions that this paper now aims to address.

3 Empirical Framework

Three empirical illustrations provide structure for application and discussion. The first illustration 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,"[10] affect subjective well-being, measured by Gallup's "ladder

by Schröder and Yitzhaki (2017).

[10] 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".

of life" zero through ten ordinal scale.[11] To motivate their empirical work, Aghion et al. (2016) develop an economic model that yields empirically testable predictions. I conduct a simulation analysis that revisits the empirical tests of the first prediction, which is that a higher job turnover rate increases well-being more when controlling for aggregated unemployment than when not controlling for aggregated unemployment. The key findings from tests of the first prediction is that creative destruction has a positive effect on SWB when controlling for MSA-level unemployment. The following methodology will examine the robustness of this empirical finding. Examinations of theoretical predictions two and three are presented in the Appendix.

The second illustration looks 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". The authors are careful to note that estimates are "qualitatively identical" when using an ordered logit model. As previously discussed in section 2.1, however, this does little to show robustness to increasing monotonic transformations of the ordinal scale. In the following analysis I will examine robustness of these empirical findings to monotonic increasing transformations of the ordinal scale when using cardinal statistical methods.

Finally, the third illustration evaluates the black-white test score gap in kindergarten through third grade. This is a controversial area of inquiry. Jencks and Phillips (1998) initially find that a substantial gap in test scores emerges in early childhood. In contrast, Fryer and Levitt (2004,

[11] This question states: "Please imagine a ladder with steps numbered from zero at the bottom and ten at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represent the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" It may seem tempting to believe these scores to be interval scores measured on a cardinal scale, but consider the following: An individual responds with a rating of 6 on the zero through ten scale. Then something happens and their latent wellbeing doubles. Using this scale, they can then only score themselves at 10. Therefore, the relationship between objective latent wellbeing and measured subjective wellbeing is non-linear to some degree. Since we do not know the functional form of this relationship, the measured subjective wellbeing scale is ordinal.

2006) find that the racial test score gap in kindergarten is "modest" and "largely explained" by socioeconomic characteristics, but that this gap widens considerably by third grade. More recently, Bond and Lang (2013) show that "plausible transformations" of test scores meaningfully change these results. This illustration, therefore, tests the claim of the "fragility" of results regarding the black-white test score in kindergarten through third grade. Since Bond and Lang (2013) already establish the sensitivity of empirical findings to reasonable monotonic transformations of the test score, this illustration in part serves as a test of my methodology. Finding similar results as Bond and Lang (2013), namely that plausible transformations can meaningfully change the racial test score gap between kindergarten and third grade, supports the credibility of the approach I develop in this paper.

3.1 Data and Estimation Methodology

The data for each of these empirical illustrations come from the replication files for each study.[12] In this subsection, I will briefly outline the estimation methodologies used in each of the studies under investigation in the present analysis.

Creative Destruction and Subjective Well-Being – In their empirical specifications Aghion et al. (2016) use a measure of creative destruction that varies at the MSA level. Since the SWB 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 SWB 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 SWB – such as income, education, gender, marital status, ethnicity, and age. The primary empirical specification uses OLS to estimate the following equation.

$$SWB_{imt} = \alpha X_{mt} + \beta Y_{mt} + \delta Z_{it} + T_t + \epsilon_{it} \tag{7}$$

In equation (7) SWB_{imt} is the Gallup measure of SWB 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

[12]The replication files for Aghion et al. (2016) were generously shared by Gallup Inc. The replication files for Nunn and Wantchekon (2011) and Bond and Lang (2013) were both available online. I thank all authors for writing clear code and organizing detailed data files.

the specification, the unemployment rate in MSA m in year t. In the tests of prediction two, X_{mt} is the job creation and the job destruction rates separately in MSA m in year t. In the tests of prediction three, X_{mt} includes either the job turnover or destruction rate, unemployment insurance generosity,[13] and the interaction of these two variables. 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.

The Slave Trade and Trust in Africa – Using OLS, Nunn and Wantchekon (2011) find a negative and statistically significant relationship between their preferred measure of slave trade activity[14] and various measures of trust with the following specification:

$$Trust_{iedc} = \psi_c + \varphi Slave\ Exports_e + X'_{iedc}\Gamma + X'_{dc}\Omega + X'_e\Phi + \eta_{iedc} \tag{8}$$

In equation (8), i represents individuals, e ethnic groups, d districts, and c countries. The dependent variable $Trust_{iedc}$ represents each of the variables measuring trust of relatives, neighbors, the local council, intra-group, and inter-group measured on a zero through three ordinal scale. ψ_c captures country fixed effects, $Slave\ Exports_e$ indicates the number of slaves sold from a particular ethnic group e. The various X vectors are individual, district, and ethnic group level controls variables. Finally, η_{iedc} is the error term. Concerned about the possibility of omitted variables biasing these results, the authors undertake several strategies to identify the causal relationship between the slave trade and trust. One of these strategies is instrumental variable analysis, where the distance of an individual's ethnic group from an ocean coast instruments for slave trade activity. The instrument approximates an ethnic group's exposure to the slave trade and is unlikely to be correlated with factors that impact present day trust. In the following illustration, I will examine the results from the instrumental variable estimation strategy. The simulation results from the simple OLS specification do not change the core findings from the simulation analysis and are shown in the Appendix.

Black-White Test Score Gap – In controversial and influential studies (Fryer and Levitt 2004, 2006) find that the racial 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. Bond and Lang (2013) show how "fragile" these results are to "plausible

[13] This is measured as the maximum weekly unemployment benefit amount within each state.

[14] The natural log of one plus slave exports normalized by land area.

transformations" of the test score. This illustration will focus on the following specification:

$$Test\ Score_i = \gamma Black_i + X_i'\rho + \upsilon_i \tag{9}$$

In equation (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 the following illustration I will show results generated by test scores in the Early Childhood Longitudinal Study (ECLS), which includes 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 Appendix. 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). Simulations using these test scores, the Peabody Individual Achievement Test (PIAT), are also shown in the Appendix.

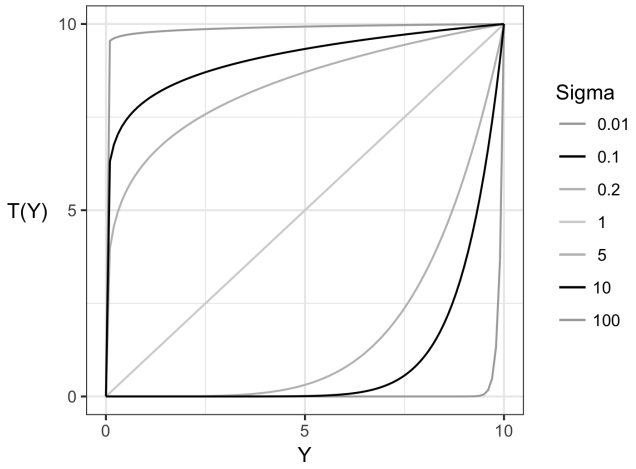
3.2 Simulation Design

It is instructive to think of the set of all possible monotonic increasing transformations as the equivalent to the set of all possible reporting functions, in the spirit of Oswald (2008). This being the case, the reporting function can be convex, concave, or linear in the raw ordinal rankings. For the purposes of running a Monte Carlo simulation, it is necessary to define a parameterized function that effectively limits the domain of potential monotonic increasing transformations. I propose the following parameterized function defining the relationship between an ordinal scale with linear numerical values and a transformed ordinal scale:

$$T(Y) = Y_{Max} \times \left(\frac{Y}{Y_{Max}}\right)^{\sigma} \qquad \forall\ \sigma > 0 \tag{10}$$

In this transformation Y is the linear ordinal scale ranging from zero through Y_{Max}. This transformation ensures that the scale maintains its endpoints at zero and Y_{Max}, respectively. The parameter σ controls the convexity or concavity of the ordinal scale. If $\sigma = 1$, then the scale remains in its linear form. If $0 < \sigma < 1$, then the SWB scale will be concave to some degree, with the distances between relatively low levels being larger than the distances between relatively high levels. Finally, if $\sigma > 1$, then the SWB scale will be convex to some degree, with the distances between relatively low levels being smaller than the distances between relatively high levels.

Figure 1: Specific Parameter Values of Transformation Function



Notes: This figure shows various transformation functions, given specific parameter values. The functions map the original variable, Y, into a transformed ordinal variable, T(Y). In this figure the ordinal scale is assumed to run from zero through ten.

In reality values of σ could exist within the positive infinite interval $(0, +\infty)$. If some restrictions to this domain are acceptable, however, equation (10) can help provide insight into the robustness of a particular empirical result to monotonic increasing transformations. Figure 1 shows equation (10), assuming a zero through ten ordinal scale, with several values of σ. Plotting these functions allows researchers to make an explicit choice about restrictions to the domain of transformations. One way to place theoretical structure on the domain of σ is to consider plausible shapes of the reporting function in a given context. For the following simulation analysis, I assume that $\sigma \in [0.1, 10]$ define the domain of plausible transformations. This is assumed because values of σ less than 0.1 and greater than ten are extreme and therefore implausible. This assumption should not be confused as an assertion. Although the assumption is used in the following simulation analysis, researchers can and should think critically about what is a plausible domain for monotonic increasing transformations of their ordinal dependent variable. After randomly picking $\sigma \in [0.1, 10]$, Y is transformed into $T(Y)$, and substituted as the dependent variable into specifications from equations (7) through (9). This process is repeated 1,000 times for each specification.

This approach for transforming the ordinal dependent variable is closely related to the Box-Cox transformation (Box and Cox 1964). Although performing the well-known Box-Cox methodology could be a valid method to transform ordinal dependent variables, the present method described above has the benefit of preserving the linear case with uniform intervals between scale values.

This feature is preferred for the present analysis so to provide clear comparisons to the existing empirical results reported in Aghion et al. (2016), Nunn and Wantchekon (2011), and Bond and Lang (2013). Of course, there are many ways to define and parameterize monotonic transformations and the present analysis remains largely agnostic to this choice. Indeed rather than transforming the scales to be either concave or convex, one could imagine a plausible transformation with an inflection point at the mid-point of the scale. Transformations defined by the normally distributed cumulative distribution function are shown in the Appendix.

4 Results and Discussion

In this section I present three elements of the results for each of the three studies under investigation. 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 Appendix. Second, I report results from the Monte Carlo simulations, as described in Section 3.2. These results are shown graphically by plotting the point estimate and the associated confidence interval for all plausible monotonic increasing transformations. Finally, I show bounds on the originally-reported point estimates. These results are reported in tabular form and allow for a direct comparison of the range of the effect bounds to the original point estimates.

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

Figure A1, in the Appendix, shows the LMA curves for each of the variables of interest in predictions one through three from Aghion et al. (2016). Prediction one examines the job turnover rate, prediction two examines both the job creation and job destruction rates, and prediction three examines both the job turnover rate and the job destruction rate interacted with unemployment insurance generosity. 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 SWB scale. This suggests that a transformation that can potentially change the sign of the OLS regression coefficient is one that will increase the distance between relatively low values, and increase the distance between relatively high values, on the SWB scale. It is expected, therefore, that if the sign on the coefficients change, they will do so for concave transformations. Nevertheless, questions persist about the

plausibility of a transformation that can change the sign of the coefficients on these variables of interest, the impact these transformations have on statistical inference, and how much the overall magnitude of the effect changes.

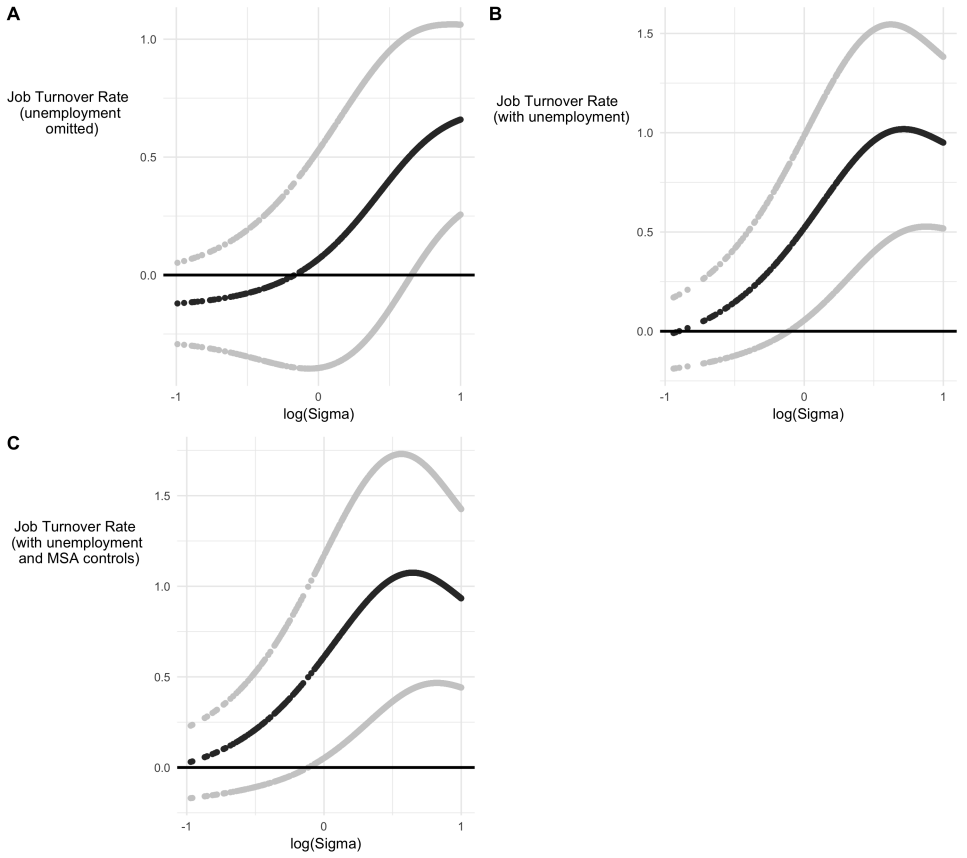
Figure 2 shows the simulation results corresponding with each of the three regressions testing prediction one, that job turnover increases SWB more when aggregate unemployment is included as a control variable. Panel A shows the coefficient on the job turnover rate corresponding to column 1 of Table 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 zero and negative one.[15] That is, when the reporting function becomes concave, rather than linear. In panel A the coefficient changes sign for almost half of all plausible 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 SWB for all plausible monotonic increasing transformations.

Concern persists about how monotonic increasing transformations impact statistical inference and the overall magnitude of the effect. It is helpful to review the core empirical findings of Aghion et al.'s tests of prediction one. These results are reported in Panel A of Table 1. Assuming a linear SWB scale with fixed and uniform intervals, the authors find the effect of creative destruction on SWB is statistically insignificant when intentionally omitting aggregated unemployment from the specification. Once aggregated unemployment is included in the regression, the effect of creative destruction on SWB both increases in magnitude and becomes statistically significant. Commenting on the magnitude of these effects, the authors report that a one standard deviation increase in job turnover has an effect on the Gallup SWB "ladder of life" which is equivalent to a 0.3 standard

[15]Simulation results are shown in terms of $\log(\sigma)$. This allows for an equal representation of concave and convex transformations in the figures. This also implies that the original results reported by authors are replicated when $\log(\sigma)$ equals zero.

Figure 2: Simulation Results for Prediction 1 in Aghion et al. (2016)



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 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.

deviation increase in the MSA-level unemployment rate.

Once the SWB scale is no longer assumed to be linear, several insights require brief comment. First, in terms of the qualitative result. For every transformation, the finding that job turnover increases SWB more when controlling for aggregate unemployment persists. That is, the effect sizes in Panel A of Figure 2 are always smaller than the effect sizes in Panels B and C for every value of $\log(\sigma)$. Therefore this result does not qualitatively change. Second, statistical inference is not robust to monotonic increasing transformations. In Panel A of Figure 2, statistical inference as reported in Aghion et al. (2016) is largely preserved. The effect only becomes statistically significant for relatively extreme convex transformations. In Panels B and C of Figure 2, however, statistical inference is quite fragile. The effect becomes statistically insignificant for most concave transformations. Therefore, statistical inference is only preserved for slightly over half of all plausible transformations. Third, the magnitudes of effects change quite dramatically depending on the transformation. Bounds on the results of prediction 1 from Aghion et al. (2016) are summarized in Table 1, with Panel B reporting estimates of the lower bound and Panel C reporting estimates of the upper bound. Column 3 of Table 1 reports the preferred specification from Aghion et al. (2016) and shows the range of plausible effects, without making arbitrary assumptions about the cardinalization of the SWB scale, extend from a small and statistically insignificant effect to an effect that is statistically significant and almost twice the size as originally reported.

For purposes of direct comparison with the author's statement about magnitude, a one standard deviation increase in job turnover has an effect on SWB which is equivalent to between a statistically insignificant -0.02 and a statistically significant 0.6 standard deviation change in the MSA-level unemployment rate. The range of magnitudes, therefore, extends from zero to twice the size of the magnitude reported by Aghion et al. (2016).[16]

Simulation results for predictions two and three from Aghion et al. (2016) are presented in the Appendix. Similar findings persist for these empirical findings. Namely, although it could be argued that only relatively extreme transformations will change the sign of coefficients of interest, the magnitude of coefficient estimates changes for reasonable transformations and therefore the economic significance of these result change dramatically. An additional finding is that the job destruction coefficient estimate in prediction two is only statistically significant for transformations around $\sigma = log(1) = 0$. Therefore, the robustness of statistical inference is quite fragile in this case.

[16]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 Appendix for additional details.

Table 1: 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)
log(σ)	1	1	1
R-squared	0.10	0.10	0.10
B: Lower Bound			
Job turnover rate	-0.120	-0.009	0.031
	(0.088)	(0.091)	(0.102)
log(σ)	-0.99	-0.94	-0.97
R-squared	0.05	0.05	0.05
C: Upper Bound			
Job turnover rate	0.659***	1.018***	1.075***
	(0.205)	(0.262)	(0.327)
log(σ)	0.99	0.72	0.65
R-squared	0.02	0.04	0.05
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 results presented in Table 2 of Aghion et al. (2016). Standard errors are clustered at the MSA level. *** p<0.01, ** p<0.05, * p<0.1.

These simulation results lead to several insights regarding the application of the methods in Schröder and Yizhaki (2017) to the empirical tests in Aghion et al. (2016). Although the theoretical "existence" results (Schröder and Yitzhaki, 2017) call into question the validity of the empirical findings in Aghion et al. (2016), in practice the qualitative result does not change. Within the universe of reasonable transformations the finding that job turnover increases SWB more when aggregate unemployment is controlled for persists. That being said, the quantitative results do change quite dramatically in terms of statistical inference and estimated effect size when relaxing assumptions about the cardinal properties of the SWB scale.

4.2 The Slave Trade and Trust in Africa (Nunn and Wantchekon 2011)

Figure A2 in the Appendix shows the LMA curves for each of the five measures of trust and the variable of interest, the natural log of slave exports normalized by land area. In this figure all but one of the LMA curves do not cross the horizontal axis. The LMA curve that does cross the horizontal axis, inter-group trust, does so for relatively high values of trust and with most of the scale below the horizontal axis only a relatively small area is left above. For all other measures of trust, the LMA curves suggest a negative covariance with the natural log of slave exports over

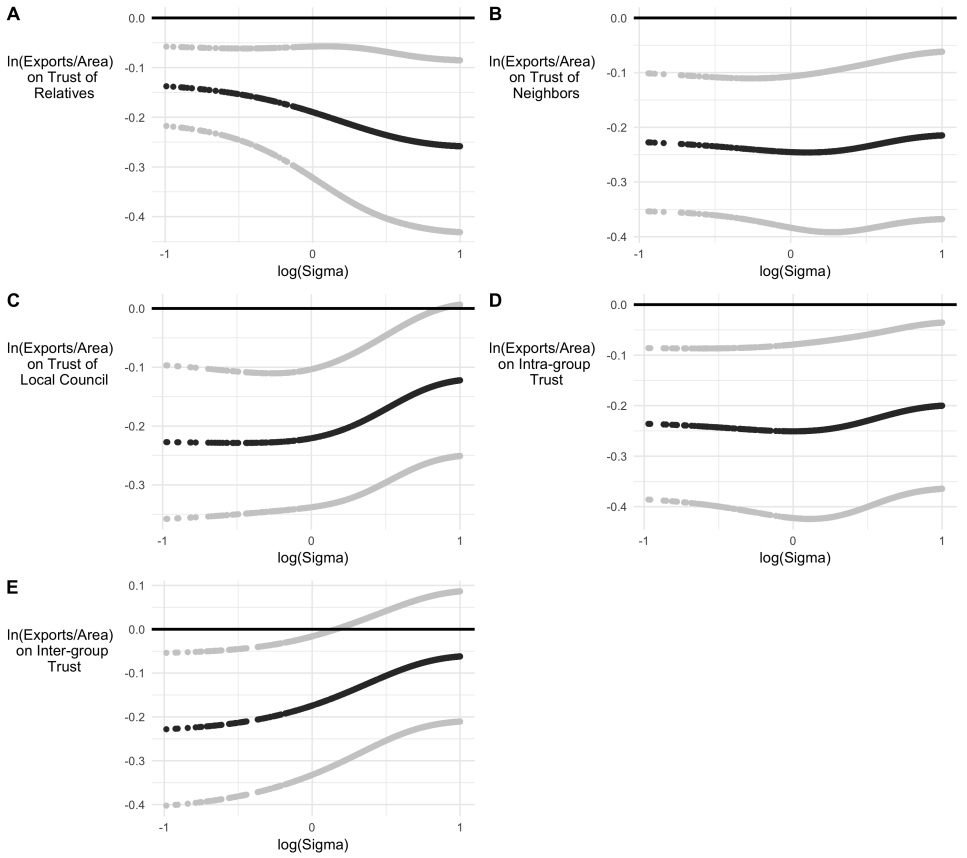
land area. Taken together, these graphical results suggest that, except for perhaps the effect on inter-group trust, the empirical findings of Nunn and Wantchekon (2011) largely pass the second theoretical sufficient condition of Schröder and Yitzhaki (2017). Even so, the statistical significance of the findings or the overall magnitude of the results may be meaningfully affected by monotonic increasing transformations.

Figure 3 shows simulation results from specifications using each of the five measures of trust as dependent variables. These specifications refer to the instrumental variable results from Table 5 in Nunn and Wantchekon (2011). The central finding of Nunn and Wantchekon (2011) is that individuals whose ancestors were heavily impacted by the slave trade are less trusting today. A key aspect of the author's findings is that the slave trade negatively impacted trust in many dimensions of people's lives. This empirical finding is consistent with historical and anthropological accounts suggesting that the slave trade had impacts deep inside the social relationships of societies and often harmed relations between friends, families, neighbors, and local leaders (Hawthorne 2003; Koelle 1854; Piot 1996). This qualitative result largely persists throughout the simulation analysis. There is no reasonable transformation that changes the sign on the estimated coefficient. Even the coefficient in the inter-group trust specification does not change sign for plausible transformations.

Moreover, for three out of five specifications the effects remain statistically significant for all plausible values of $\log(\sigma)$. The two exceptions are the effects of the slave trade on trust of the local council and inter-group trust, which both become statistically insignificant for convex transformations of the ordinal scale measuring trust. While assuming a linear scale measuring trust with fixed intervals Nunn and Wantchekon (2011) find evidence of a negative and statistically significant effect on all five measures of trust. These results are reported in Panel A of Table 2. This simulation analysis suggests that some of these findings may only persist under certain transformations of the ordinal scale. Nevertheless, the qualitative result persists for most reasonable transformations within three out of five of the measures of trust. Bounds on these effects are reported in Panels B and C of Table 2. Importantly, with the exception of the results for inter-group trust, the core finding of Nunn and Wantchekon (2011) persists for all values within these bounds. That is, even when relaxing arbitrary assumptions about the cardinalization of the trust scale, the finding that the slave trade negatively affected present day trust persists.

The robustness of the standard errors around these effect estimates is, in part, driven the relatively robust coefficient estimates themselves. The estimates of the effect on trust of neighbors and intra-group trust are highly robust as the magnitude hardly moves at all for all monotonic

Figure 3: Simulation Results for Table 5 in Nunn and Wantchekon (2011) - IV Estimates of the Effect of the Slave Trade on Trust



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 5 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.

Table 2: Bounds on IV Estimates of the Effect of the Slave Trade on Trust

	(1) Trust of relatives	(2) Trust of neighbors	(3) Trust of local council	(4) Intra-group trust	(5) Inter-group trust
A: Original 0 - 3 scale					
ln (1+exports/area)	-0.190***	-0.245***	-0.221***	-0.251***	-0.174**
	(0.067)	(0.071)	(0.060)	(0.088)	(0.081)
log(σ)	1	1	1	1	1
R-squared	0.13	0.15	0.20	0.15	0.12
B: Lower Bound					
ln (1+exports/area)	-0.258***	-0.246***	-0.229***	-0.251***	-0.228**
	(0.088)	(0.073)	(0.062)	(0.088)	(0.089)
log(σ)	0.99	0.10	-0.49	0.00	-0.98
R-squared	0.12	0.16	0.15	0.15	0.07
C: Upper Bound					
ln (1+exports/area)	-0.138***	-0.215***	-0.122*	-0.200**	-0.062
	(0.041)	(0.078)	(0.066)	(0.084)	(0.076)
log(σ)	-0.99	0.99	0.99	0.99	0.99
R-squared	0.07	0.15	0.18	0.14	0.12
Individual controls	Yes	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Obs.	16,709	16,679	15,905	16,636	16,473

Notes: This table shows bounds on the results presented in Table 5 of Nunn and Wantchekon (2011). Standard errors are adjusted for two-way clustering at the ethnicity and district levels. *** p<0.01, ** p<0.05, * p<0.1.

increasing transformations. The estimates of the effects on trust of relatives, trust of the local council, and inter-group trust are slightly less robust, but only vary by about 0.1 points on the zero through three scale used to measure trust. While discussing magnitude, the authors perform a variance decomposition and find that, along with the other covariates, slave exports explain 5.4% of the total variation of trust in neighbors. Additionally, of this 5.4%, about 16% is explained by slave exports. Results from the simulation analysis show that over all values of log(σ), along with the other covariates, slave exports explain between 4.2% and 5.4% of the total variation of trust in neighbors. Furthermore, of this 4.2 - 5.4%, roughly 16% is consistently explained by slave exports.

Comparing these results with the simulation results from the investigation of Aghion et al. (2016) provide additional insights. One possible reason why the empirical results of Nunn and Wantchekon (2011) are more robust to monotonic increasing transformations of the ordinal scale may lie in the number of categories on the scale. The Gallup SWB variable used in Aghion et al. (2016) consists of 11 categories ranked from zero through ten. The Afrobarometer trust variables on the other hand only consist of four categories ranked from zero through three. Since the transformation, as defined in equation (10), fixes the endpoints at the minimum and maximum values, respectively,

the transformation is able to change the interval between values by adjusting nine values in Aghion et al. (2016) and only two values in Nunn and Wantchekon (2011). To test this idea, I redefine the Gallup SWB measure as being on a zero through five scale, rather than a zero through ten scale, by combining adjacent categories. Re-running the simulations for first prediction in Aghion et al. (2016) shows that although limiting the number of categories does limit the variation of coefficient estimates to monotonic increasing transformations, this is not a panacea.[17] There may be an implicit trade-off between the number of categories included in an ordinal scale. On the one hand, more categories provides more variation and specific information in the data. On the other hand, more categories may lead to less robustness to monotonic increasing transformations and potentially biased results when using cardinal statistical methods. At least in terms of the results from Aghion et al. (2016), however, this trade-off does not seem to have a meaningful overall impact on the robustness of empirical results to monotonic increasing transformations of the ordinal scale.

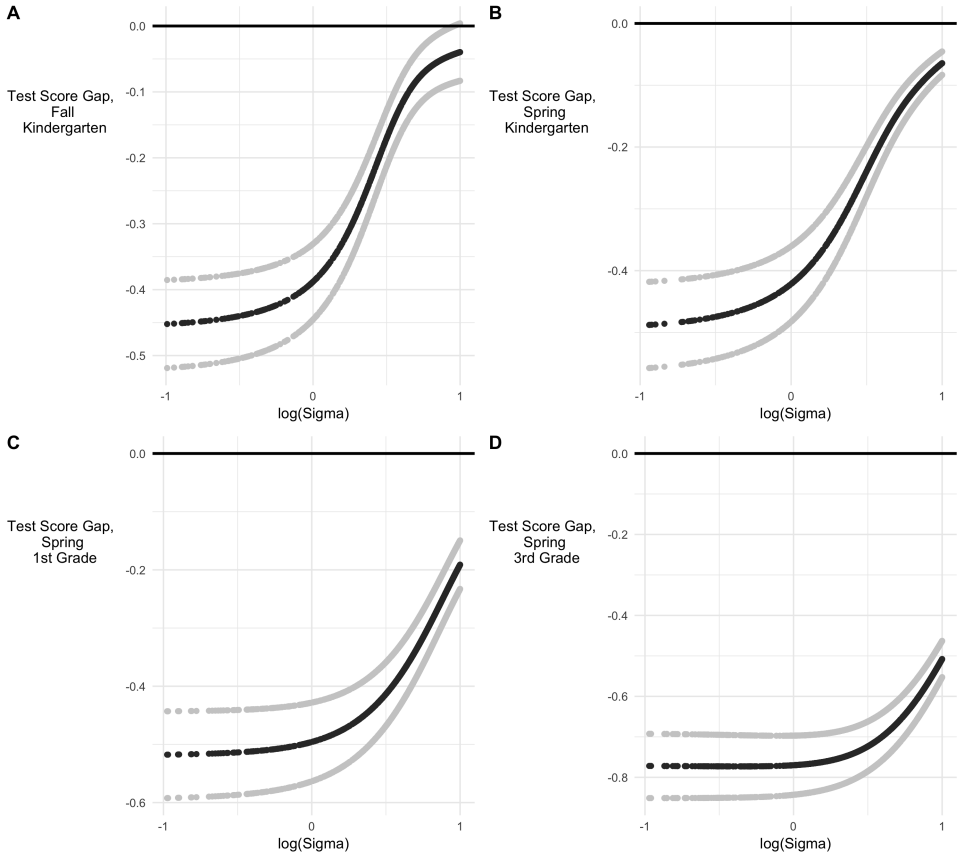
4.3 The Black-White Test Score Gap in Grades K-3 (Bond and Lang 2013)

Figure A3, in the Appendix, shows the LMA curves of the racial 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 that for all test score values there is a negative covariance between the test score and the racial status variable. This suggests that there is no monotonic increasing transformation that can change the sign on the black-white test score gap between kindergarten and third grade. That is, there is no way to change the intervals between test scores so that it appears that black students actually test higher than white students in these data. However, changing of the sign is not the primary concern of Bond and Lang (2013) when they demonstrate the "fragility" of these results. The author's key finding is that "plausible transformations... greatly reduce [the test score gap] in the ECLS during the early school years". Therefore, although the LMA curves are indeed instructive, concern persists about the robustness of estimated effect sizes to a class of monotonic increasing transformations.

Figure 3 shows simulation results for each of the four time periods where test scores are collected in the ECLS between kindergarten and third grade. These results relate to the results in Table 4 of Bond and Lang (2013). The authors show several transformations that display the "fragility"

[17] These results are shown in the Appendix.

Figure 4: Simulation Results for Table 4 in Bond and Lang (2013) - Evolution of the Black-White 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 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 these results. 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.

This finding is largely replicated in the simulation analysis. Noting that the transformations could vary between grades, such that the "true" test score reporting function may be defined

with different $\log(\sigma)$ values during each testing period, both the maximum and minimum growth transformations from Bond and Lang (2013) can be found in Figure 3. The test score gap in the fall of kindergarten, shown in Panel A, is the largest with concave transformations. With this transformation, 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. With this transformation, the gap could be as low as 0.45 standard deviations in the spring of third grade. Therefore the growth in the test score gap between blacks and whites is a statistically insignificant 0.01 standard deviations. This is relatively close to the result Bond and Lang (2013) report in column 2 of Table 4 in their paper.

Additionally, the test score gap in the fall of kindergarten is the smallest with relatively high $\log(\sigma)$ values. These 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. If these transformations represent the "true" test score reporting function, then the growth in the test score gap between blacks and white is a statistically significant 0.72 standard deviations. This is relatively close to the result reported by Bond and Lang (2013) in column 3 of Table 4 in their paper.

These results are reported in tabular form in Table 3. Panel A reports the baseline results which consider a linear test score scale. These estimates indicate the existence of a substantial black-white gap in test scores that starts as early as kindergarten. Panels B and C report bounds on these estimates. These results suggest that the size of this gap at each testing period differs quite dramatically depending on the assumed cardinalization of the test score scale. In particular, Column 1 of Table 3 reports an upper bound of the test score gap to be relatively small and only marginally statistically significant in the fall of kindergarten. Again considering the fact that the transformations could vary between grades, the growth in the black-white test score gap between kindergarten and third grade could be either relatively small or quite dramatic. Again, this finding is consistent with the results from Bond and Lang (2013).

There are three main conclusions to discuss from this final empirical illustration. The first highlights that even though the relationship between an ordinal variable and a covariate of interest may pass the theoretical results of Schröder and Yitzhaki (2017), in practice there may be important concerns. In particular, consider the change in magnitude of empirical results examining the black-white test score gap in kindergarten through third grade. Although the LMA curves suggest that

Table 3: 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	-0.404***	-0.435***	-0.493***	-0.746***
	(0.030)	(0.032)	(0.034)	(0.036)
log(σ)	1	1	1	1
R-squared	0.04	0.04	0.05	0.09
	B: Lower Bound			
Black	-0.452***	-0.488***	-0.518***	-0.773***
	(0.034)	(0.036)	(0.038)	(0.039)
log(σ)	-0.99	-0.94	-0.98	-0.36
R-squared	0.06	0.05	0.05	0.09
	C: Upper Bound			
Black	-0.039*	-0.064***	-0.191***	-0.508***
	(0.022)	(0.010)	(0.021)	(0.023)
log(σ)	0.99	0.99	0.99	0.99
R-squared	0.00	0.00	0.01	0.05
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). Robust standard errors are presented in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

there does not exist a monotonic increasing transformation that can change the sign of the test score gap, serious concern persists about the robustness of effect sizes for reasonable transformations. Even though, in Figure 3, it is never the case that black students score higher than white students, the size and growth of the test score gap ranges from being relatively small and economically insignificant to quite large and economically meaningful.

The second is about assessing "fragility" or robustness of results using ordinal dependent variables. Other than discussing the change in the magnitudes of results applied to the specific context of a study, as done in the previous two illustrations, another way to test the robustness of empirical results to reasonable monotonic increasing transformations is to compare the size of the confidence interval around specific coefficient estimates to the overall change in the size of the coefficient estimate itself. One way to do this is to simply find the ratio of the overall range of an estimate (e.g. $\hat{\beta}_{Max} - \hat{\beta}_{Min}$) and the maximum confidence interval around each coefficient estimate for all plausible transformations. If this ratio is greater than one then the overall change in an actual coefficient estimate is greater than the largest confidence interval around each particular estimate, suggesting empirical results are not robust to monotonic increasing transformations. If this statistic is less than one then the range of coefficient estimates is completely contained between the largest

confidence interval of any particular estimate, suggesting empirical results are robust to monotonic increasing transformations.

For example, consider the results from Panel A in Figure 3. The overall change in the test score gap is 0.41 standard deviations for all transformations with σ ranging from 0.1 to 10. This is the difference between the highest and lowest estimated test score gap in fall of kindergarten, shown in Panel A. The confidence interval around these coefficient estimates has a maximum difference of 0.13. Taking the ratio of these two numbers provides a statistic measuring robustness of a specific empirical result of 3.17 (= 0.41/0.13). For use of comparison, consider the results from Nunn and Wantchekon (2011) from Panel B in Figure 2. The overall change in the effect size of the slave trade on trust of neighbors is 0.03 points on the zero through three ordinal scale, while the confidence interval around these estimates has a maximum difference of 0.31 points. Again, taking the ratio of these two numbers provides a robustness statistic of 0.1 (= 0.03/0.31). This suggests that the results of the effect of the slave trade on trust in Africa (Nunn and Wantchekon, 2011) are much more robust to monotonic increasing transformations than the black-white test score gap in kindergarten through third grade (Bond and Lang, 2013). These robustness statistics provide a method and quantitative structure for comparing the robustness of empirical results to monotonic increasing transformations.

Finally, this illustration of the results from Bond and Lang (2013) provides a test of the validity of the methodology presented in this paper for understanding how much the cardinal treatment of ordinal variables matters. Since the contribution of Bond and Lang (2013) already establishes the "fragile" results of previous studies examining the evolution of the racial test score gap (Jencks and Phillips 1998; Fryer and Levitt 2004, 2006), the similar findings generated from the simulations lend credence to the methodology of this paper. In particular the plausible transformations, discussed by Bond and Lang (2013), are not only replicated by the simulation analysis but also exists within the domain of plausible transformations used in the present analysis. Moreover, additional results from Bond and Lang (2013), such as examining the ECLS test score gap while controlling for socioeconomic factors and using an alternative test score (the PIAT), are also largely replicated while using the methodology developed in this paper. These additional results are shown in the Appendix.

4.4 Limitations

There are several limitations to the methodology developed in this paper for validating the cardinal statistical treatment of ordinal variables. One limitation is the function defining monotonic increasing transformations is only one of many ways to specify such transformations. In this sense, finding empirical results that are not robust to a given class of monotonic increasing transformations is conceptually equivalent to hypothesis testing and rejecting the null hypothesis. On the contrary, finding empirical results that are robust to a given class of monotonic increasing transformations is conceptually equivalent to failing to reject the null hypothesis, since there are likely other theoretically valid classes of such transformations. In the Appendix I re-run the core results of this paper using a transformation with a functional form that includes an inflection point at the mid-point of the ordinal scale. This transformation parameterizes different cumulative distribution functions (CDFs). Although specific details about these results are different than those discussed above, the broad qualitative findings are robust to a different class of transformations. In particular, the magnitude and economic significance of the empirical results from Aghion et al. (2016) are not robust to reasonable CDF transformations. The core findings of Nunn and Wantchekon (2011) are again largely robust to reasonable CDF transformations. Finally, consistent with the results from Bond and Lang (2013), the growth in the black-white test score gap between kindergarten and third grade is highly dependent on the form of CDF transformations.

A second limitation of this analysis is in the assumptions associated with the choice of limiting the infinite set of all possible monotonic increasing transformations to a finite set of all reasonable transformations. The results of this analysis will of course be sensitive to this choice. As discussed by Bond and Lang (forthcoming), however, if a consensus were to form around such domain restrictions it may be possible to perform valid cardinal statistical analysis of ordinal variables. This paper aims to take a practical first step in developing this consensus. Future work could focus on generalizing this methodology and further developing a consensus for limiting the domain of transformations.

Finally, this paper abstracts from an additional complication of ordinal scales: cross-sectional heterogeneity. Following essentially every previous empirical study that uses ordinal variables, throughout this paper it is assumed that the reporting function is fixed across all observations. This, it can be argued, is a rather unrealistic assumption. As previously noted, existing work has developed a survey methodology, namely anchoring vignettes, that can help control for issues of

reporting function heterogeneity (see King et al. 2004; King and Wand 2007; Hopkins and King 2010). Although, these methods are useful, not all ordinal variables are enumerated via a survey. Therefore, future work could make an important contribution by examining robustness of empirical results to cross-sectional heterogeneity in reporting functions.

5 Conclusion

This paper builds off recent theoretical contributions of Schröder and Yitzhaki (2017) and Bond and Lang (forthcoming) on the appropriateness of the cardinal statistical treatment of ordinal variables. To perform this task, I examine the work of Aghion et al. (2016), Nunn and Wantchekon (2011), and Bond and Lang (2013) who use cardinal statistical methods to empirically analyze subjective well-being, trust, and early elementary test scores, respectively. This methodology first makes an assumption about the plausible domain of monotonic increasing transformations and then runs Monte Carlo simulations by randomly picking transformations within this domain. This leads to results that provide practical insights into the robustness of empirical results to monotonic increasing transformations.

In three empirical illustrations, I find that the valid cardinal treatment of ordinal variables requires methodologically sound justification, since theoretically such treatment may produce biased or inconsistent estimates. In practice, robustness to monotonic increasing transformations depends on the specific details of individual specifications. In particular, I find that most of the variables of interest in Aghion et al. (2016), fail the existing theoretical tests for the valid cardinal treatment of ordinal variables. This finding, on the surface, is rather troubling for the robustness of the original findings by Aghion et al. (2016) to monotonic increasing transformations of the Gallup SWB scale. However, since there are infinitely many monotonic increasing transformations, existence of at least one such transformation does not necessarily imply that the transformation is reasonable. Limiting the domain of all possible transformations to a finite set of plausible transformations and a simulation analysis suggests that although transformations that can change the sign exist, such transformations are relatively unlikely and could perhaps be argued to be unreasonable.

Meanwhile, the empirical findings of Nunn and Wantchekon (2011) and Bond and Lang (2013) largely pass existing theoretical tests, however the empirical results for the former are much more robust to monotonic increasing transformations than the latter. While the core qualitative results, in terms of effect size and statistical significance, of Nunn and Wantchekon (2011) largely persist for

all reasonable transformations, consistent with the key insight of Bond and Lang (2013), empirical analysis investigating the evolution of the black-white test score gap between kindergarten and third grade are quite fragile when exposed to monotonic increasing transformations. This highlights that passing existing theoretical tests by Schröder and Yitzhaki (2017) and Bond and Lang (forthcoming) does not necessarily imply results are robust to monotonic increasing transformations of the ordinal scale.

This research has implications for future empirical research which necessitates the use of ordinal variables. Inevitably, as economic research extends itself into realms of society and the economy where factors cannot be quantitatively measured or directly observed, the need to use ordinal variables becomes increasing frequent. This situation presents a challenge to researchers regarding empirical methodology and statistical model selection. The research builds of the recent work of Schröder and Yitzhaki (2017) and Bond and Lang (forthcoming), who claim that it is no longer valid to assume that the ordinality or cardinality of ordinal variables makes no qualitative difference (as in Ferrer-i-Carbonell and Frijters, 2004). However, in the present analysis I find that just because there is a monotonic increasing transformation that can change the sign of linear regression coefficients this transformation need not be plausible. At the same time, even if a given specification passes existing theoretical tests for the valid cardinal treatment of ordinal variables, the size and statistical significance of an estimated coefficient may not be robust. Therefore, in the presence of an ordinal dependent variable, the choice of an empirical methodology and statistical model requires sound methodology and careful thought. Although some empirical findings may be robust to monotonic increasing transformations, many will not be.

References

Acemoglu, D., Johnson, S., and Robinson, J. **(2001)** "The Colonial Origins of Comparative Development: An Empirical Investigation" *American Economic Review* 91 (5) pp. 1369-1401.

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

Aitchison, J. and Silvey, S.D. **(1957)** "The Generalization of Probit Analysis to the Case of Multiple Responses" *Biometrika*, 44 pp. 131-140.

Baker, F. **(2001)** *The Basics of Item Response Theory*, College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Becker, W. and Kennedy, P. **(1992)** "A Graphical Exposition of the Ordered Probit" *Econometric Theory*, 8 (1) pp. 127-131.

Baird, S., de Hoop, J., and Oxler, B. **(2013)** "Income Shocks and Adolescent Mental Health" *Journal of Human Resources*, 48 (2), pp. 370-403.

Bloem, J., Boughton, D. Htoo, K., Hein, A., and Payongayong, E. **(2018)** "Measuring Hope: A Quantitative Approach with Validation in Rural Myanmar" *Journal of Development Studies*, 54 (11), pp. 2078-2094.

Bond, T. and Lang, K. **(2013)** "The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results" *The Review of Economics and Statistics*, 95 (5) pp. 1468-1479.

Bond, T. and Lang, K. **(forthcoming)** "The Sad Truth About Happiness Scales" *Journal of Political Economy.* (Available online as NBER Working Paper No. 19950).

Borghans, L., Duckworth, A.L., Heckman, J., and ter Weel, B. **(2008)** "The Economics and Psychology of Personality Traits" *Journal of Human Resources*, 43 (4), pp. 972-1059.

Box, G. and Cox, D. **(1964)** "An Analysis of Transformations" *Journal of the Royal Statistical Society* 26 (2) pp. 211-252.

Bryson, A. and MacKerron, G. **(2017)** "Are Your Happy While You Work?" *The Economic Journal* 127 (599) pp. 106-125.

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

Clark, A. and Oswald, A. **(1996)** "Satisfaction and Comparison Income" *Journal of Public Economics* 61 (3) pp. 359-381.

Clark, A. and Oswald, A. **(1994)** "Unhappiness and Unemployment" *The Economic Journal* 104 (424) pp. 648-659.

Cornaglia, F., Feldman, N.E., Leigh, A. **(2014)** "Crime and Mental Well-Being" *Journal of Human Resources*, 49 (1), pp. 110-140.

Davis, S., Haltiwanger, J., and Schuh, S. **(1996)** "Small Business and Job Creation: Dissecting the Myth and Reassessing the Facts" *Small Business Economics* 8 (4) pp. 297-315.

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

Deaton, A. **(2008)** "Income, Health, and Well-Being around the World: Evidence from the Gallup World Poll" *Journal of Economic Perspectives*, 22 (2), pp. 1-31.

Di Tella, R., and MacCulloch, R. **(2008)** "Gross national happiness as an answer to the Easterlin Paradox?" *Journal of Development Economics* 86 (1) pp. 22-42.

Diener, E., and Seligman, M. **(2004)** "Beyond money: Toward an Economy of Well-Being" *Psychological Science in the Public Interest*, 5 (1) pp. 1-31.

Diener, E., Suh, M., Lucas, R., and Smith, H. **(1999)** "Subjective Well-Being: Three Decades of Progress" *Psychological Bulletin* 124 pp. 276-302.

Dolan, P., Peasgood, T., and White, M. **(2008)** "Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being" *The journal of Economic Psychology* 29 pp. 94-122.

Easterlin, R. **(1974)** "Does Economic Growth Improve the Human Lot? Some Empirical Evidence" David, P., and Reder, M. (Eds) *Nations and Households in Economic Growth*, pp. 89-125, New York: Academic Press.

Easterlin, R. **(1995)** "Will raising the incomes of all increase the happiness of all?" *Journal of Economic Behavior and Organization*, 27 pp. 1-34.

Ferrer-i-Carbonell, A. and Frijters, P. **(2004)** "How important is methodology for the estimates of the determinants of happiness?" *Economic Journal*, 114 pp. 641-659.

Frijters, P., Haisken-DeNew, J.P., and Shields, M.A. **(2004)** "Money Does Matter! Evidence from Increasing Real Income and Life Satisfaction in East Germany Following Reunification" *American Economic Review*, 94 (3) pp. 730-740.

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.

Fryer, R. and Levitt, S. **(2006)** "The Black-White Test Score Gap Through Third Grad" *American*

Law and Economics Review, 8 pp. 248-281.

Glewwe, P. **(1997)** "Estimating the Impact of Peer Group Effects on Socioeconomic Outcomes: Does the Distribution of Peer Group Characteristics Matter?" *Economics of Education Review*, 16 (1), pp. 39-43.

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*, 53 (2), pp. 330-355.

Graham, C., Egger, A., and Sukhtanker, S. **(2004)** "Does happiness pay? An exploration based on panel data from Russia?" *Journal of Economic Behavior and Organization*, 55 (3) pp. 319-342.

Greene, W. **(2012)** *Econometric Analysis* (Seventh ed.) Boston: Pearson Education.

Hadar, J. and Russell, W.R. **(1969)** "Rules for Ordering Uncertain Prospects" *American Economic Review*, 49 (1) pp. 25-34

Hawthorne, W. **(2003)** *Planting Rice and Harvesting Slaves: Transformations along the Guinea-Bissau Coast, 1400-1900.* Porsmouth, NH: Heinemann.

Heckman, J.J. **(1981)** "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," In: Manski, C., McFadden, D. (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA.

Hopkins, D. and King, G. **(2010)** "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability" *Public Opinion Quarterly*, pp. 1-22.

King, G. Murray, C. Salomon, J., Tandon, A. **(2004)** "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research" *American Political Science Review*, 98 (1) pp. 191-207.

King, G. and Wand, J. **(2007)** "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes" *Political Analysis*, 15 pp. 46-66.

Koelle, S.W. **(1854)** *Polyglotta Africana: Or a Comparative Vocabulary of Nearly Three Hundred Words and Phrases, in More than One Hundred Distinct African Languages.* London: Church Missionary House.

Krueger, A.B., Kahneman, D., Schkade, D., Schwarz, N., and Stone, A., **(2009)** "National Time Accounting: The Currency of Life" In *Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*, (ed.) Krueger, A.B., University of Chicago Press.

Krueger, A.B. **(2017)** "Where Have All the Workers Gone? An Inquiry into the Decline of the U.S. Labor Force Participation Rate" *Brookings Papers on Economic Activity*, fall 2017, pp. 1-87.

Jacob, B. and Rothstein, J. **(2016)** "The Measurement of Student Ability in Modern Assessment Systems" *Journal of Economic Perspectives*, 30 (3) pp. 85-108.

Jencks, C. and Phillips, M. **(1998)** "The Black-White Test Score Gap: An Introduction" in Jencks, C. and Phillips, M. (eds.) *The Black-White Test Score Gap* Washington DC: Brookings Institution Press.

Lancaster **(2000)** "The incidental parameter problem since 1948" *Journal of Econometrics*, 95 pp. 391-413.

Lang, K. **(2010)** "Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member" *The Journal of Economic Perspectives*, 24 (3) pp. 167-181.

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

Lorenz, M.O. **(1905)** "Methods of Measuring the Concentration of Wealth" *Journal of the American Statistical Association*, No. 70 pp. 209-219.

Luechinger, S., Meier, S., and Stutzer, A. **(2010)** "Why Does Unemployment Hurt the Employed? Evidence from the Life Satisfaction Gap Between the Public and the Private Sector" *Journal of Human Resources*, 45 (4), pp. 998-1045.

Manski, C. **(2003)** *Partial Identification of Probability Distributions*, Springer Series in Statistics, New York: Springer.

Marks, B., and Flemming, N. **(1999)** "Influences and consequences of well-being among Australian young people 1980-1995" *Social Indicators Research* 46 pp. 301-323.

McKelvey, R. and Zavoina, W. **(1975)** "A Statistical Model for the Analysis of Ordered Level Dependent Variables" *Journal of Mathematical Sociology*, 4 pp. 103-120.

Neyman, J. and Scott, E. **(1948)** "Consistent estimation from partially consistent observations" *Econometrica*, 16, pp. 1-32.

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, A.J. **(2008)** "On the curvature of the reporting function from objective reality to subjective feelings" *Economics Letters* 100 pp. 369-372.

Piot, C. **(1996)** "Of Slaves and the Gift: Kabre Sale of Kin during the Era of the Slave Trade" *Journal of African History*, 37 (1): pp. 31-49.

Putnam, R. **(2001)** *Bowling Alone: The Collapse and Revival of American Community*, Simon & Schuster, 1st edition, New York: NY.

Riedl, M. and Geishecker, I. **(2014)** "Keep it simple: Estimation Strategies for Ordered Response Models with Fixed Effects", *Journal of Applied Statistics*, vol. 41. (11), pp. 2358-2374.

Ritter, J. and Anker, R. **(2002)** "Good jobs, bad jobs: Workers' evaluations in five countries" *International Labor Review* 141 (4) pp. 331-358.

Schaffer, M.E. **(2015)** "GINIREG: Stata Module for Gini Regression" *Available online: https://EconPapers.repec.o*

Schröder, C. and Yitzhaki, S. **(2016**) "PISA Country Rankings and the Difficulty of Exams", Working Paper, Available at SSRN.

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

Snell **(1964)** "A Scaling Procedure for Ordered Categorical Data" *Biometrics*, 20 pp. 592-607.

Stevens, S. **(1946)** "On the Theory of Scales of Measurement" *Science*, vol. 10, pp. 667-680.

Stevenson, B. and Wolfers, J. **(2008)** "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox" *Brookings Paper on Economic Activity*, pp. 1-87.

Stevenson, B. and Wolfers, J. **(2013)** "Subjective Well-Being and Income: Is There Any Evidence of Satiation?" *American Economic Review: Papers and Proceedings*, 103 (3) pp. 598-604.

Tamer, E. **(2010)** "Partial Identification in Econometrics" *The Annual Review of Economics*, 2 pp. 167-195.

Thorndike, R. **(1966)** "Intellectual Status and Intellectual Growth" *Journal of Educational Psychology*, vol 57. pp. 121-127.

Wang, S. and Zhou, W. **(2018)** "Do Siblings Make Us Happy?" *Economic Development and Cultural Change*, vol. 66, no. 4. pp. 827-840.

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.

Yitzhaki, S. and Schechteman, E. **(2012)** "Identifying monotonic and non-monotonic relationships" *Economics Letters*, 116 (1) pp. 23-25.

Yizhaki, S. and Schechteman, E. **(2013)** *The Gini Methodology: A Primer on a Statistical Methodology.* Springer. New York.

Supplemental Appendix

A1 A Sampling of the Cardinal Use of Ordinal Variables

Table A1 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 A1: Examples of the Cardinal Treatment of Ordinal Variables

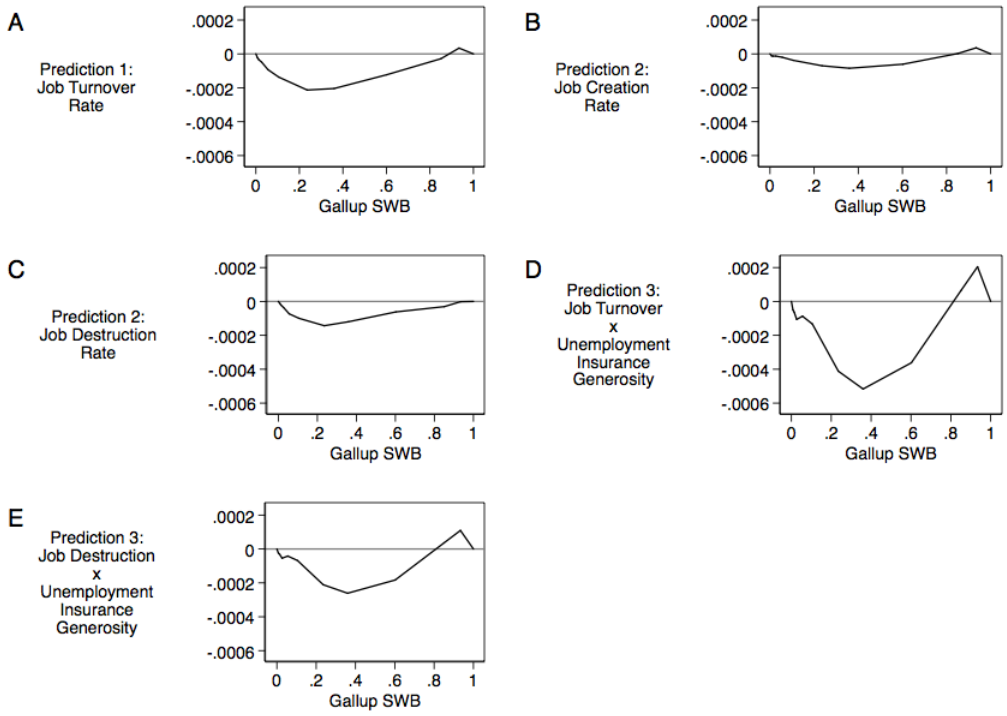
Citation	Journal	Dependent Variable	Method
Aghion et al. (2016)	*American Economic Review*	Subjective well-being	OLS with fixed effects
Alatas et al. (2012)	*American Economic Review*	Satisfaction	OLS
Ashraf et al. (2014)	*American Economic Review*	Subjective well-being	OLS
Bandiera et al. (2017)	*Quarterly 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.

A2 LMA Curves

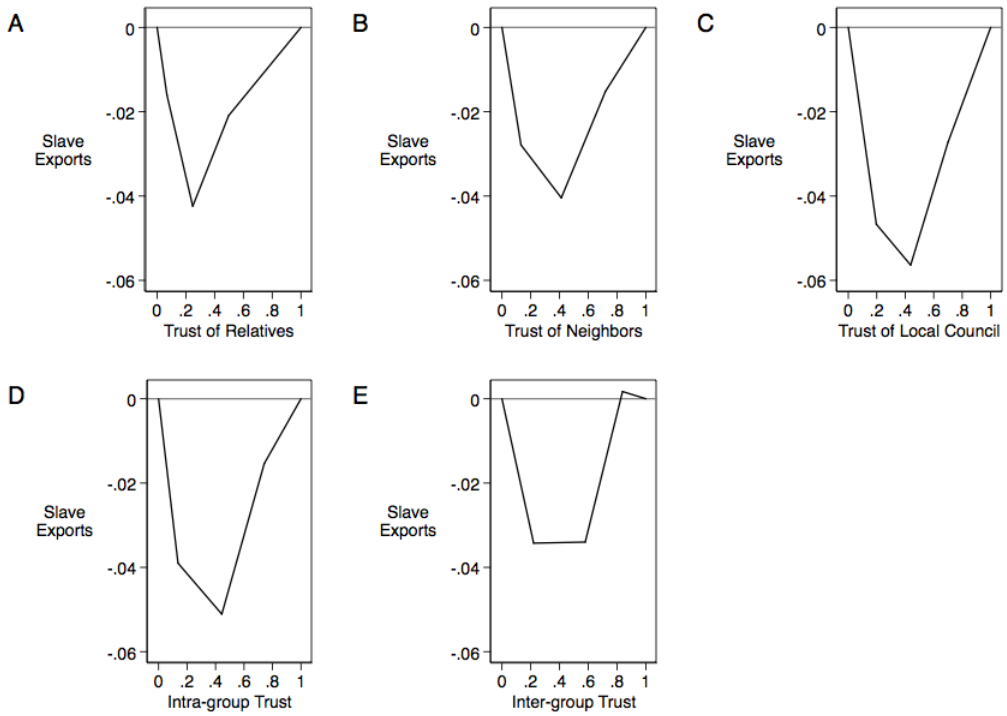
The LMA curves for the three empirical investigations are shown in the following figures. Figure A1 shows the LMA curves for the results from Aghion et al. (2016) testing the relationship between creative destruction and subjective well-being. Figure A2 shows the LMA curves for the results from Nunn and Wantchekon (2011) examining the effect of the slave trade on trust in sub-Saharan Africa. Finally, Figure A3 shows the LMA curves for the "fragile" results from Bond and Lang (2013) on the black-white test score gap in kindergarten through third grade. Specific details about how these LMA curves are constructed can be found in Section 2.2 of the main manuscript.

Figure A1: LMA Curves with Gallup Current Ladder SWB and Creative Destruction



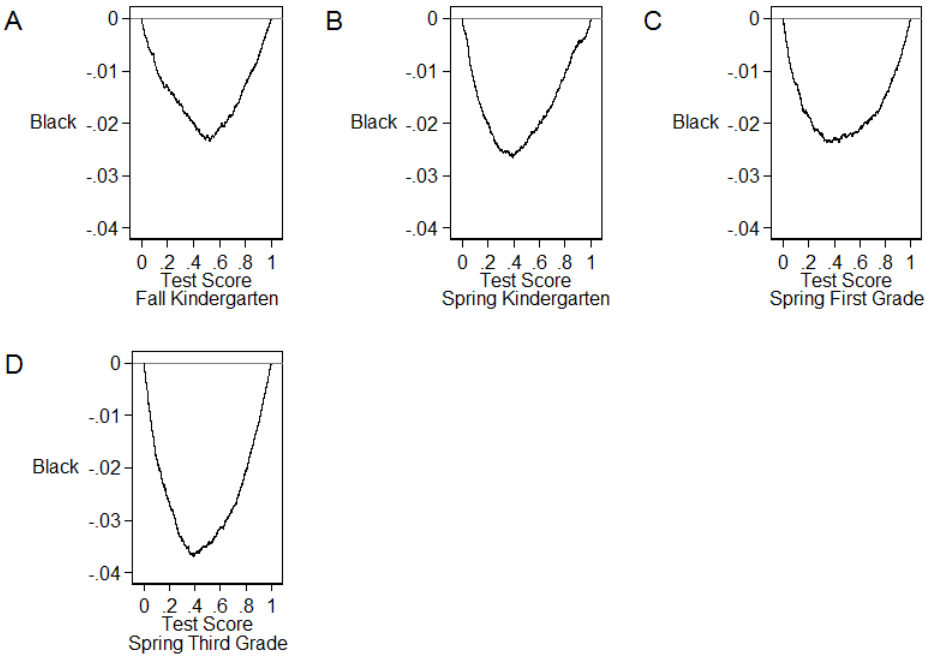
Notes: This figure shows LMA curves between the Gallup "ladder of life" SWB variable and the various variable of interest for each of the first three predictions tested in Aghion et al. (2016). The y-axis is fixed between all graphs.

Figure A2: LMA Curves with Afrobarometer Measures of Trust and the Slave Trade



Notes: This figure shows LMA curves between the five measures of trust gathered via the Afrobarometer survey and the natural log of slave exports normalized by land area (Nunn and Wantchekon, 2011). The y-axis is fixed between all graphs.

Figure A3: LMA Curves with ECLS Test Scores and Racial Status



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.

A3 Predictions Two and Three from Aghion et al. (2016)

In the theoretical framework section of their paper, Aghion et al. (2016) provide theoretical predictions that motivate their empirical strategies. Prediction one was presented in the main manuscript of this paper. Prediction two states: A higher job creation rate increases well-being, whereas a higher job destruction rate decreases well-being. Prediction three states: A higher turnover rate increases well-being more, whereas a higher job destruction rate decreases well-being less the more generous the unemployment benefits. Aghion et al. (2016) find empirical support for these predictions. In this section I test the robustness of these empirical findings to reasonable monotonic increasing transformations. Recall from Figure A1, that the LMA curves for the job creation coefficient in prediction two and both coefficients of interest in prediction three cross the horizontal axis. This suggests that it is theoretically possible for a monotonic increasing transformation to change the sign of these coefficients. At the same time, the economic significance of the empirical findings testing predictions two and three from Aghion et al. (2016) may not be robust to such transformations.

Figures A4 and A5 show the simulation results for these theoretical predictions, respectively. The first insight from both of these figures is, even though the LMA curves do cross the horizontal axis, in practice there is no reasonable monotonic increasing transformation that changes the sign on any of these coefficients. Therefore, empirical findings supporting theoretical predictions two and three qualitatively persist. Despite this finding, the magnitude of the effects and the statistical significance do change meaningfully over all reasonable transformations.

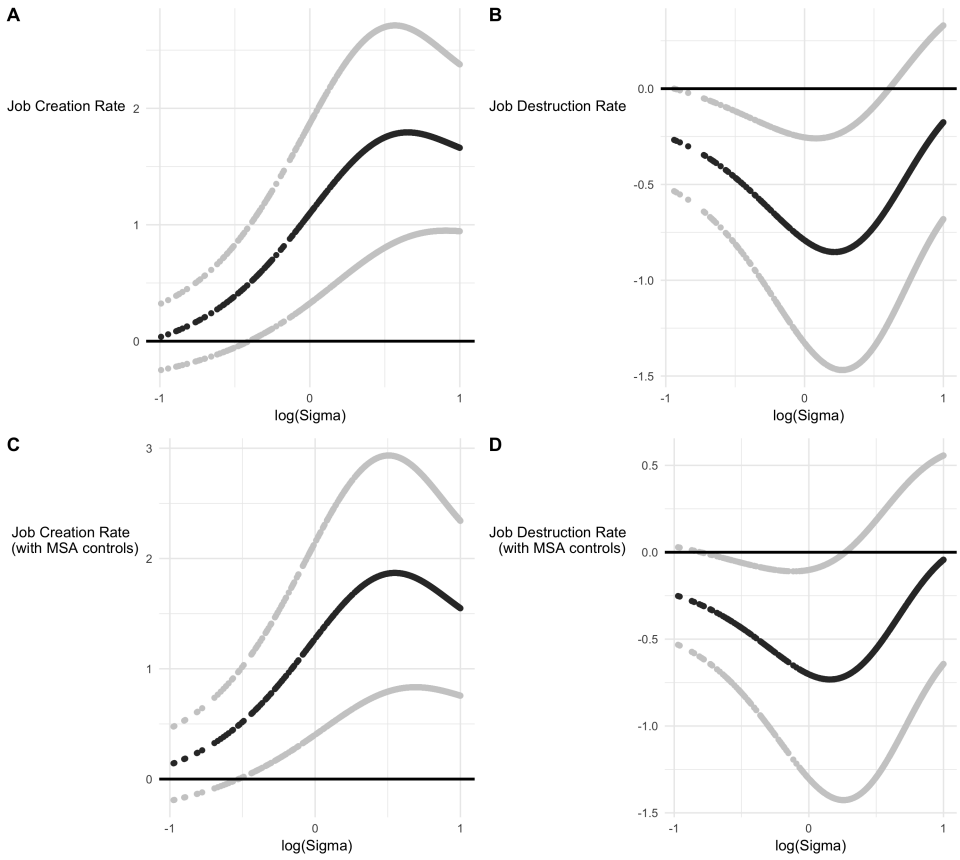
In particular, Panels A and C in Figure A4 show the coefficients on the job creation rate for prediction 2, without and with additional MSA-level controls, respectively. These simulation results show the coefficient ranging from between just above zero to just below two. Additionally, with the exception of some concave transformations, these effects are statistically significant. Note that the magnitudes of effects presented by Aghion et al. (2016) suggest that a one standard deviation increase in the job creation rate is associated with an increase in SWB of about 0.12 standard deviations. The simulation results in Panel C of Figure A4 suggest that a one standard deviation increase in the job creation rate is associated with an increase in SWB of between 0.01 and 0.17 standard deviations. Therefore, the economic significance of this effect is not robust to reasonable transformations.

Panels B and D in Figure A4 show the coefficients on the job destruction rate for prediction

2, without and with additional MSA-level controls, respectively. The simulation results show the coefficient ranging from -0.043 to -0.73. In terms of statistical significance, reasonable transformations cause the effect to become statistically insignificant more often than not. This is especially the case when additional MSA-controls are added into the regression, shown in Panel D of Figure A4. In fact, assuming a reporting function that is largely linear, which is represented by transformations with σ values around one, is the only case when the coefficient on the job destruction rate is statistically significant. This is enough to suggest that this result is not robust to reasonable monotonic increasing transformations.

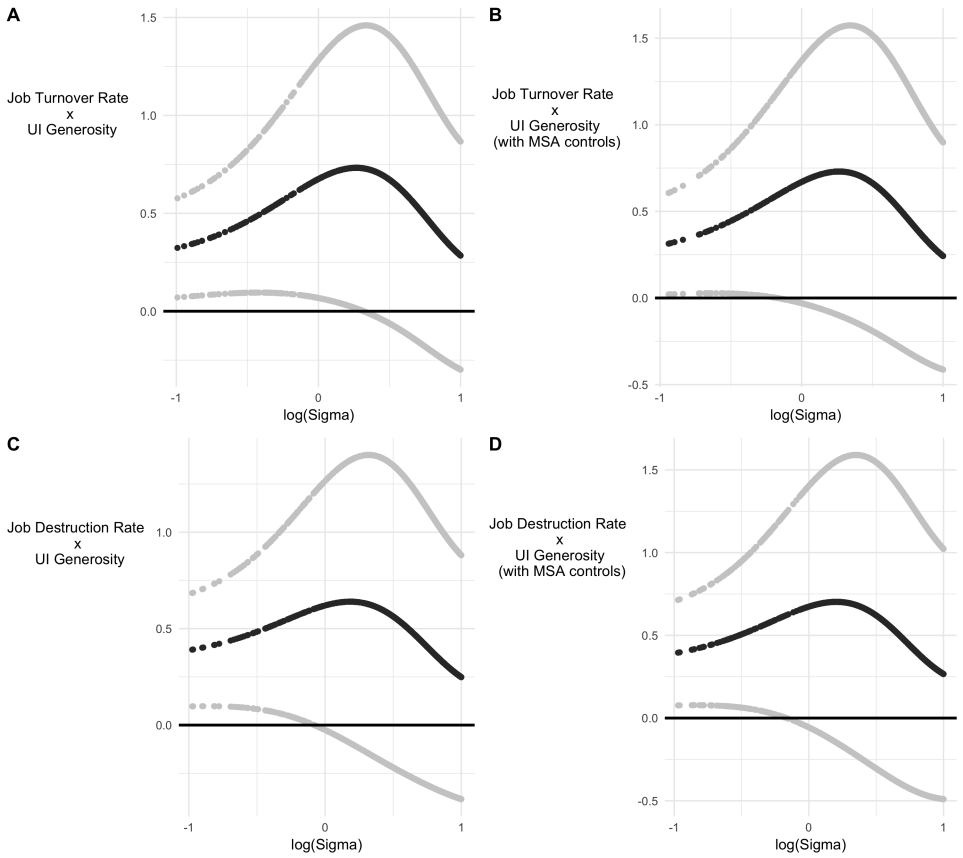
Similar results follow for tests of prediction three in Figure A5. The most notable features of these simulation results is the lack of robustness, in terms of statistical significance, of these results to monotonic increasing transformations. The empirical results presented by Aghion et al. (2016) are only statistically significant at the 10% level, when additional MSA-control variables are included. Each of the Panels in Figure A5 show, however, that for most reasonable transformations these results are statistically insignificant. Although it is worth noting that the coefficient estimates are slightly more stable in prediction three, these results still are not robust to reasonable monotonic increasing transformations.

Figure A4: Simulation Results for Prediction 2 in Aghion et al. (2016)



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 MSA-level. Each panel refers to a different specification used to test prediction 2. Panel A refers to the coefficient on the job creation rate in column (1) of prediction 2, which intentionally omits additional MSA-level controls. Panel B refers to the coefficient on the job destruction rate in column (1) of prediction 2, which again intentionally omits additional MSA-level controls. Panel C refers to the coefficient on the job creation rate in column (2) of prediction 2, which includes MSA-level controls. Finally, panel D refers to the coefficient on the job destruction rate in column (2) of prediction 2, which again includes MSA-level controls.

Figure A5: Simulation Results for Prediction 3 in Aghion et al. (2016)



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 MSA-level. Each panel refers to a different specification used to test prediction 3 in Aghion et al. (2016). Panel A refers to the coefficient on job turnover × unemployment insurance generosity in column (1), which intentionally omits additional MSA-level controls. Panel B refers to the coefficient on job turnover × unemployment insurance generosity in column (2), which includes additional MSA-level controls. Panel C refers to the coefficient on job destruction × unemployment insurance generosity in column (3), which intentionally omits additional MSA-level controls. Finally, panel D refers to the coefficient on job destruction × unemployment insurance generosity in column (4), which includes MSA-level controls.

A4 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 (Cameron and Trivedi 2010) and express the marginal effect in terms of the original linear ordinal scale. Table A2 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 A2: Marginal Effects in Terms of Transformed and Linear SWB Scales

Dep. Variable: Gallup SWB	(1) $\log(\sigma) = 0$	(2) $\log(\sigma) = -1$	(3) $\log(\sigma) = 1$
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

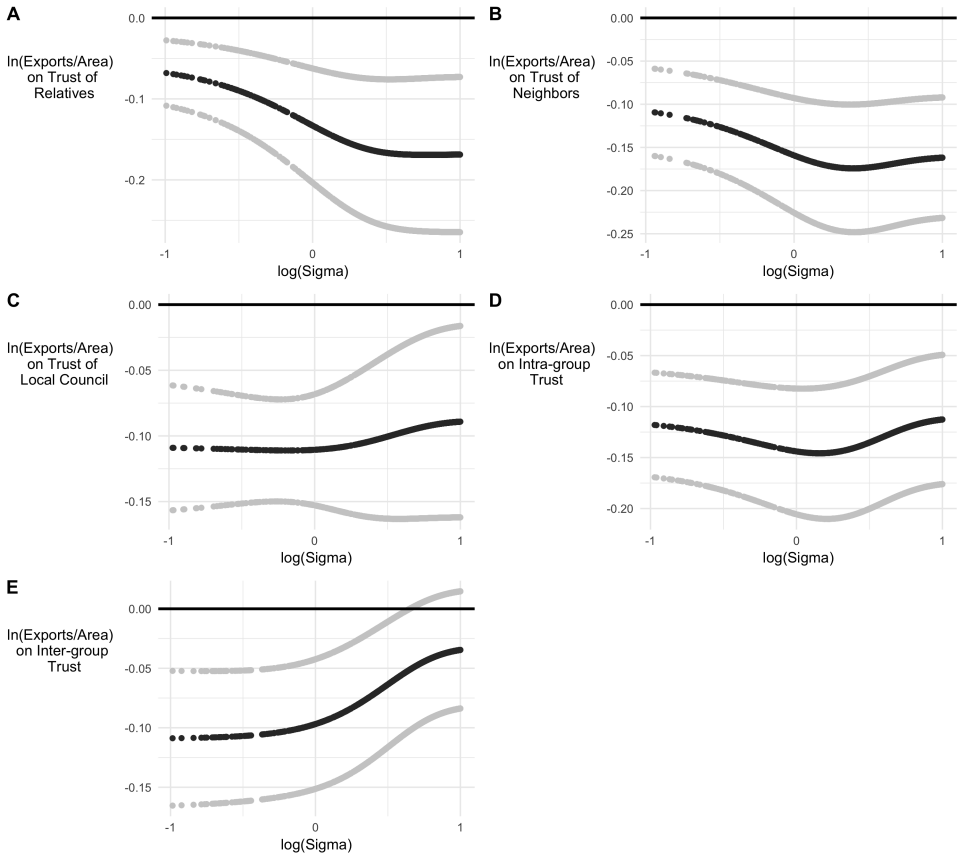
Notes: Within each panel, row (i) shows the raw marginal effect given the discrete σ value and row (ii) shows the marginal effect given the discrete σ value that is transformed back into terms of the original zero through ten linear ordinal SWB scale. Standard errors are shown in parentheses. In rows (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.

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 reasonable monotonic increasing transformations. These results are shown in Figure A6. In general the core finding from the simulations of 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.

For the sake of comparison with the instrumental variable results, recall that a statistic measuring robustness of from Panel B in Figure 5 is 0.096. Now consider the analogous result from Panel B in Figure A6. The overall change in the test score gap is 0.06 standard deviations for all reasonable transformations. The confidence interval around these coefficient estimates has a maximum difference of 0.15. Taking the ratio of these two numbers provides a statistic measuring robustness of a specific empirical result of 0.4 (= 0.06/0.15). Although the instrumental variable results are more robust than the OLS results, both robustness statistics are well below a value of one, indicating that both results are relatively robust to monotonic increasing transformations.

Figure A6: Simulation Results 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.

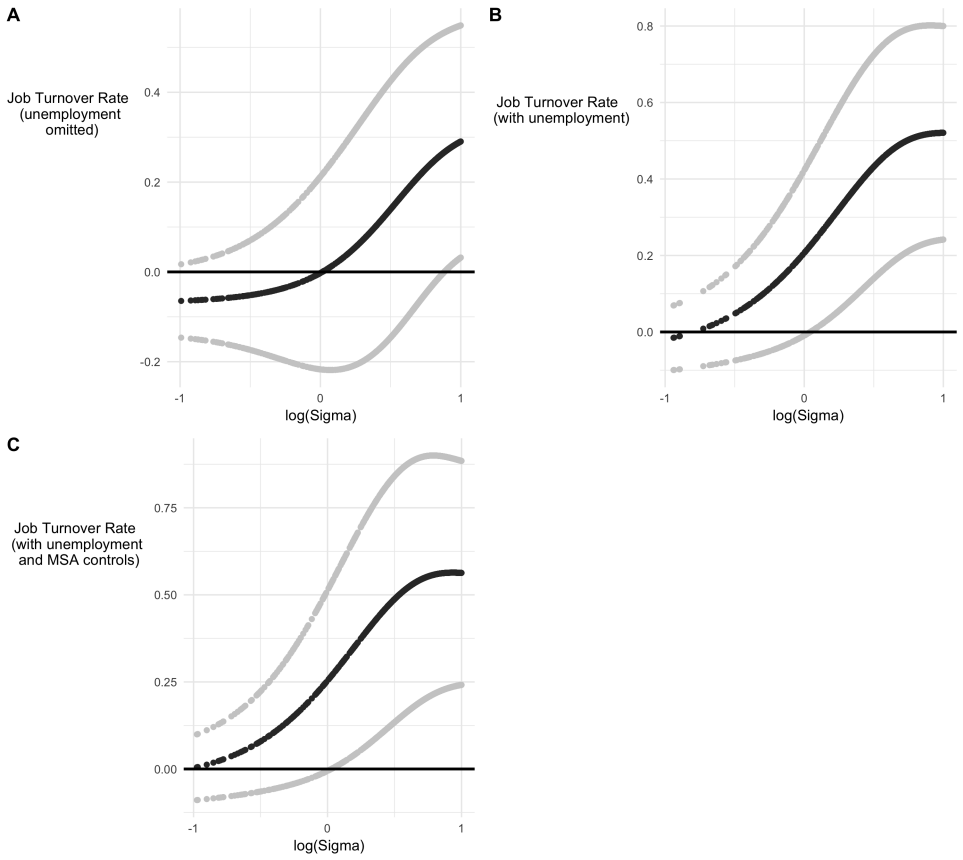
A6 Prediction One from Aghion et al. (2016), Zero - Five Scale

Comparing the simulation results from Aghion et al. (2016) with those from Nunn and Wantchekon (2011) leads to the question: Does the number of categories included on an ordinal scale impact the robustness of results to monotonic increasing transformations? To test this idea, I redefine the scale used to measure SWB in Aghion et al. (2016) as being defined by six categories, rather than 11. This nearly cuts the number of categories in half. To do this I re-coded responses of 1 and 2 to be 1, 3 and 4 to be 2, 5 and 6 to be 3 and so on. This leads to a scale that is defined from zero through five, rather than from zero through 10. Next, I re-run the exact same simulations on the empirical specifications testing prediction one from Aghion et al. (2016).

These results are presented in Figure A7 Several insights require discussion. First, similar to the use of the zero through ten scale, these results only change sign for relatively few cases and when additional MSA-level control variables are included in the specification the result never changes sign. Second, although the magnitude continues to change depending on the functional form of the transformation, the range of the coefficient estimates is roughly half the range of coefficient estimates when using the original zero through ten scale. In particular, the coefficient estimate varies from close to zero to just above 1, in Panel C of Figure 2, and this same coefficient varies from close to zero to just above 0.5, in Panel C of Figure A7. This result is theoretically reasonable, since re-coding the scale by cutting the number of categories roughly in half will effectively cut the overall variation in the variable roughly in half.

This result suggests that, although there may be an implicit trade-off between the number of categories on an ordinal scale and robustness to monotonic increasing transformations, limiting the number of categories is not a panacea. Even ordinal scales with relatively few categories may produce empirical findings that are relatively fragile to monotonic increasing transformations.

Figure A7: Simulation Results for Prediction 1 in Aghion et al. (2016) with Zero - Five Scale



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 MSA-level. Each panel refers to a different specification used to test prediction 1. Panel A intentionally omits the unemployment rate and additional MSA-level controls. Panel B includes the unemployment rate but intentionally omits additional MSA-level controls. Finally, panel C includes the unemployment rate and additional MSA-level controls.

A7 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.

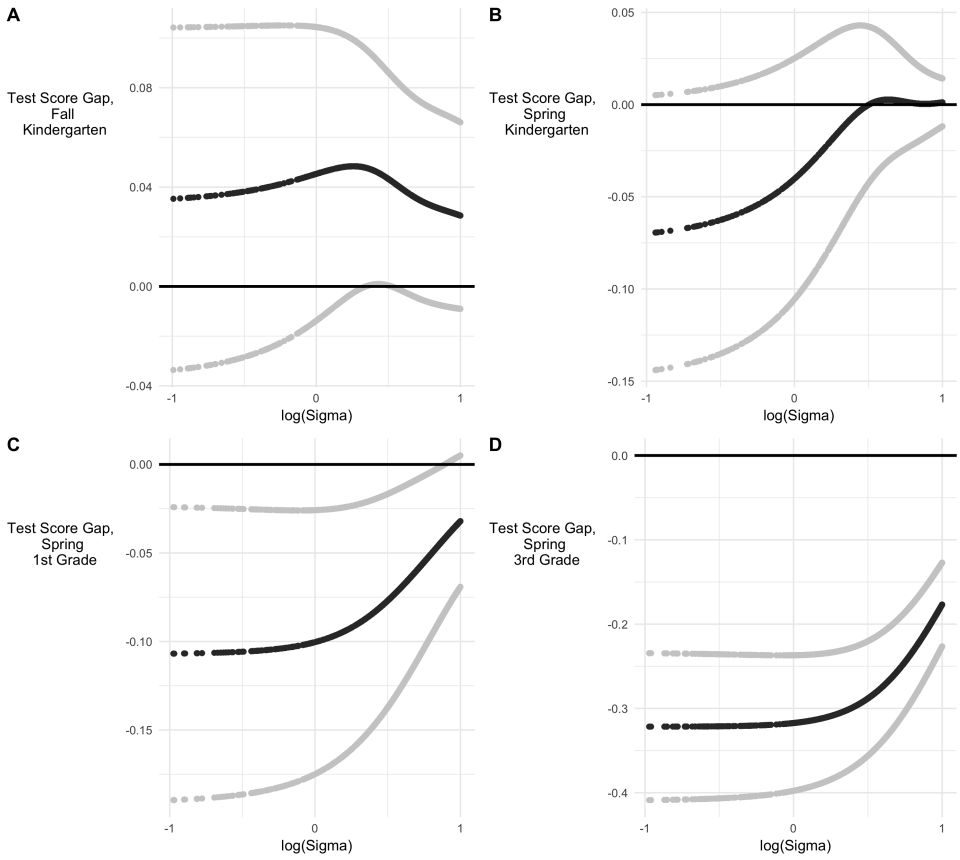
The first illustration is to perform the same analysis as shown in Figure 3, but control for socioeconomic factors that may explain some of the early elementary racial test score gap. This more closely examines the result from Fryer and Levitt (2004) suggesting that the black-white test score gap in 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 A8. Although the racial test score gap is not statistically significant, in the fall of kindergarten, the coefficient estimate is largely robust to reasonable 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 A8 where the test score gap ranges from between 0.17 and 0.32 standard deviations for reasonable 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 A9 largely replicates this result, with a range of the gap between a statistically insignificant 0.06 and a statistically significant -.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 A9, for the most part, replicates this finding with a racial test score gap ranging between 0.15 and 0.61 standard deviations. The results in Panels C and D in Figure A9 are the most different from those presented in Table 3 by Bond and Lang (2013). Nevertheless, the implications are qualitatively similar.

In addition to similar results presented in the main manuscript, these results lend credence to

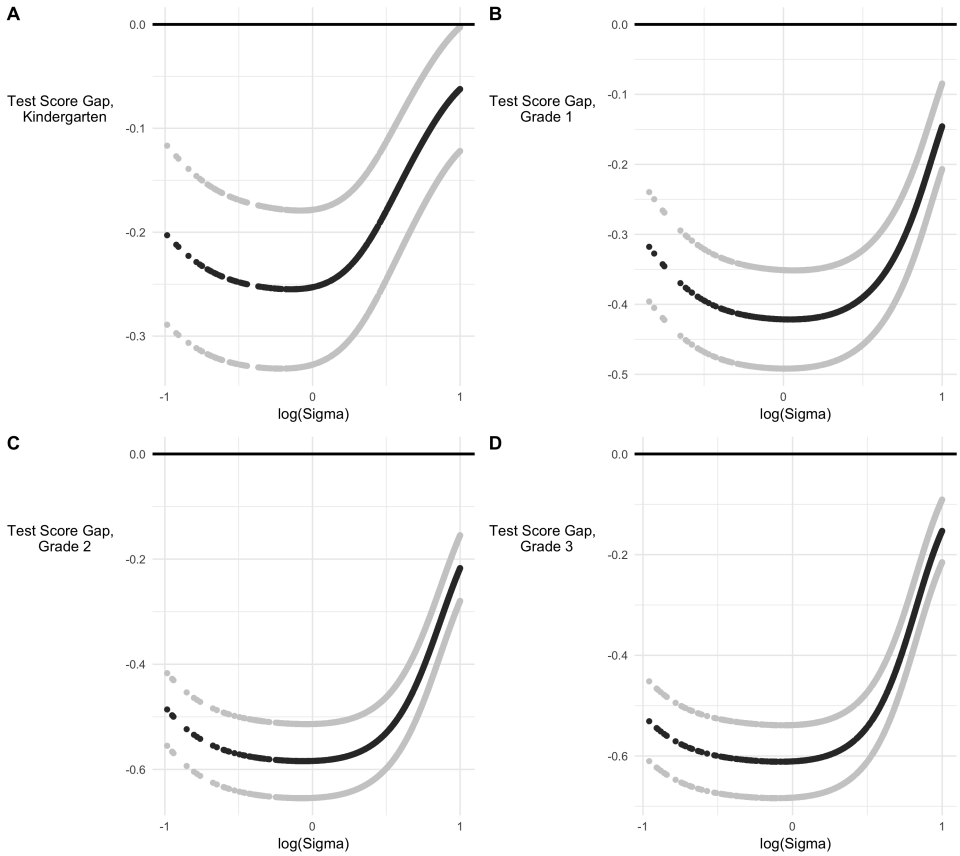
the credibility of the methodology developed in this paper. This is highlighted by the fact that the results from column 2 and 3 in Table 5 of Bond and Lang (2013) can be found within the range of results shown in Figure A8. Additionally, the results from Table 3 in Bond and Lang (2013) are for the most part replicated in Figure A9.

Figure A8: Simulation Results 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 A9: Simulation Results 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.

A8 Transformations with an Inflection Point

An alternative class of transformations are those with an inflection point. Rather than being either concave or convex, this class of transformations are convex below and concave above an inflection point. The motivation for this class of transformation builds off of the intuition of Oswald (2008), where people are reluctant to report the highest values of some ordinal scale. If the "true" reporting function is one with an inflection point, this would imply that people are also reluctant to report the lowest values on the scale. Said differently, it takes a relatively larger marginal gain or loss to move an individual off the mid-point of some scale.

One way to define such a class of reporting functions is to transform various cumulative distribution functions as follows:

$$T(Y) = Y_{Max} \times F\Big(\frac{X - Y_{Mid}}{\sigma}\Big) \tag{A1}$$

In equation (A1), $F(\cdot)$ is the cumulative distribution function (CDF) with a mean of Y_{Mid} and a standard deviation of σ. The domain of σ is dependent by the range of the ordinal scale. In general, if the scale is zero through Y_{Max} and $\sigma = Y_{Mid}$, then $F(\cdot)$ will essentially be linear. Note that although this line will be linear, it will not exactly replicate empirical results that assume a linear reporting function. In practice, CDF transformations do not preserve endpoints at zero and Y_{Max}, respectively. Indeed this is one reason for preferring the class of transformations used in the main manuscript. If σ is relatively close to zero, then $F(\cdot)$ will look increasingly like a step function, with the step at Y_{Mid}. These details are visualized in Figure A10 for a zero through ten scale.

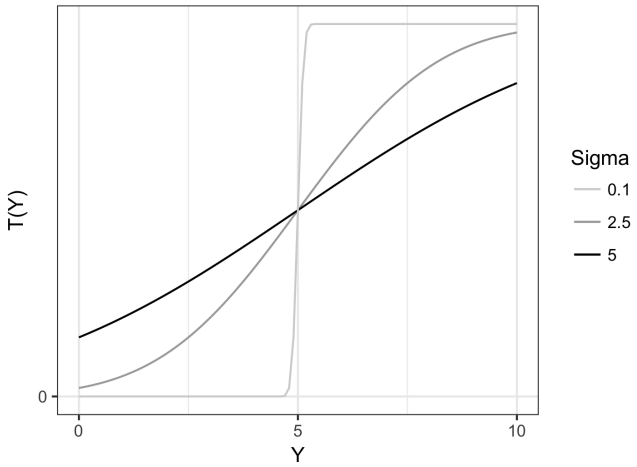
The following figures use equation (A1) to define a class of transformations used in the Monte Carlo simulations for the results presented in the main manuscript. Figure A11 shows simulation results for prediction 1 of Aghion et al. (2016). Panel A shows results corresponding to column 1 in Table 2 from Aghion et al. (2016). This shows that the coefficient is positive for linear transformations, with σ close to five, but becomes negative for transformations that approach a step-function, with σ close to zero. Coefficient estimates are statistically insignificant for all transformations. Panel B and C, show simulation results corresponding to columns 2 and 3 in Table 2 from Aghion et al. (2016). For linear transformations the coefficient estimate is positive and statistically significant. Additionally, similar to the empirical results reported by Aghion et al. (2016), the coefficient estimate is larger in Panels B and C than the estimate in Panel

A. This suggests that the qualitative result, that creative destruction increases SWB more when the unemployment rate is accounted for, persists for all transformations. Similar to the results presented in the main manuscript of this paper, however, the quantitative result does change depending on the form of the transformation. Most notably, for transformations approaching a step-function, with σ values close to zero, the coefficient estimates in Panels B and C become statistically insignificant. Although, it is worth pointing out that the coefficient estimates themselves remain relatively constant for the class of transformations defined by CDFs.

Figure A12 shows simulation results for the instrumental variable estimates of the effect of the slave trade on trust in sub-Saharan Africa, from Nunn and Wantchekon (2011). Similar to the results presented in the main manuscript of this paper, the empirical findings of Nunn and Wantchekon (2011) are relatively robust to a class of transformations defined by CDFs. For all transformations, none of the coefficient estimates change sign. This is consistent with the theoretical conditions of Schröder and Yitzhaki (2017). With the exception Panel A, referring to the effect on trust with neighbors, all coefficient estimates are statistically significant for all CDF transformations. In Panel A, the estimate becomes statistically insignificant for σ values close to zero, when the transformation resembles a step-function and is essentially a binary indicator variable identifying if respondents trust their relatives or not. Finally the coefficient estimates themselves are also relatively robust to reasonable transformations, this is again similar to the simulation results presented in the main manuscript.

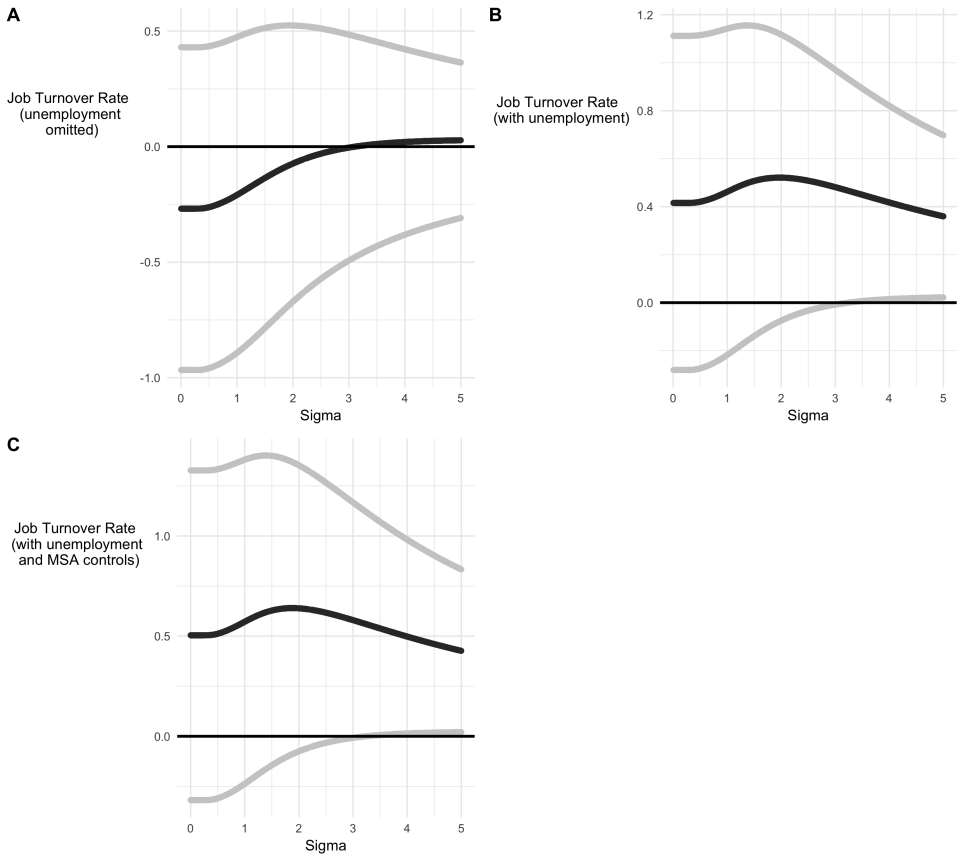
Finally, Figure A13 shows simulation results for the results from Bond and Lang (2013) on the black-white test score gap between kindergarten and third grade. Although the motivation for a reporting function with an inflection point may seem less credible in the case of test scores, I find results similar to those presented in the main manuscript. In short, the empirical findings of Bond and Lang (2013) are qualitatively replicated. In particular, the growth of the test score gap could be relatively large and meaningful with a growth of about 0.3 standard deviations or relatively small and insignificant with a growth of less than 0.1 standard deviations. Therefore, empirical results on the black-white test score gap in kindergarten through third grade are also not robust to a class of transformations defined by CDFs.

Figure A10: Specific Parameter Values of CDF Transformation



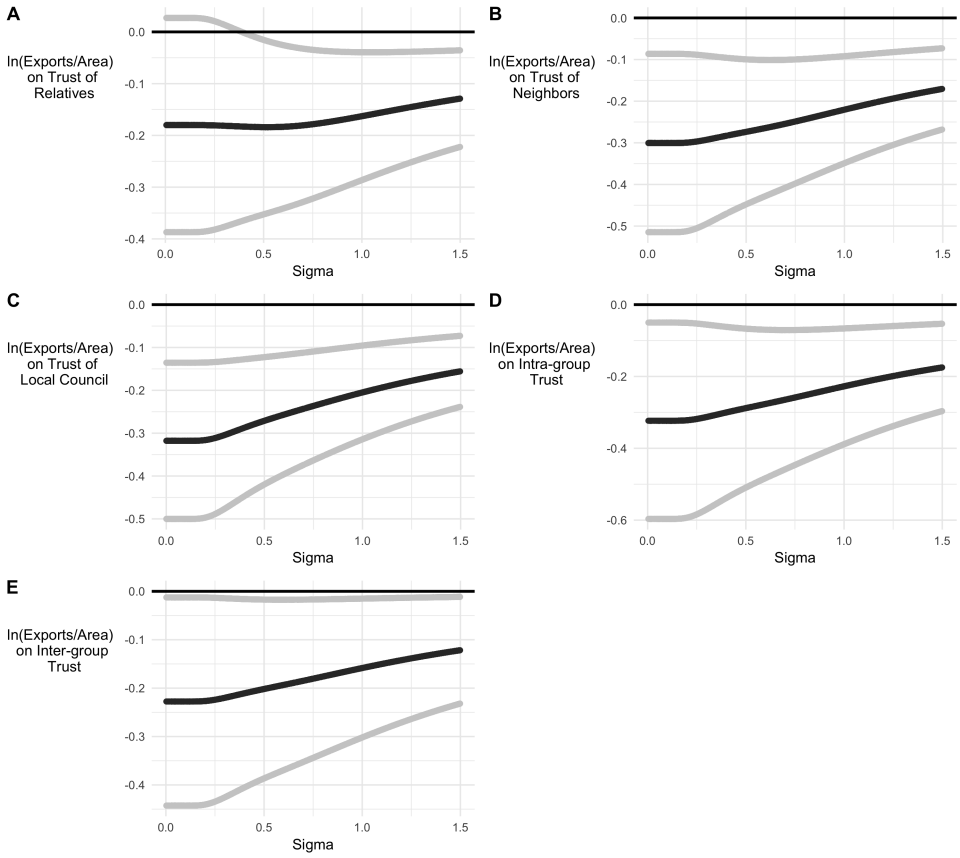
Notes: This figure shows various transformation functions, given specific parameter values. The functions map the original variable, Y, into a transformed ordinal variable, T(Y). In this figure the scale is assumed to run from 0 - 10.

Figure A11: Simulation Results for Prediction 1 in Aghion et al. (2016), CDF Transformation



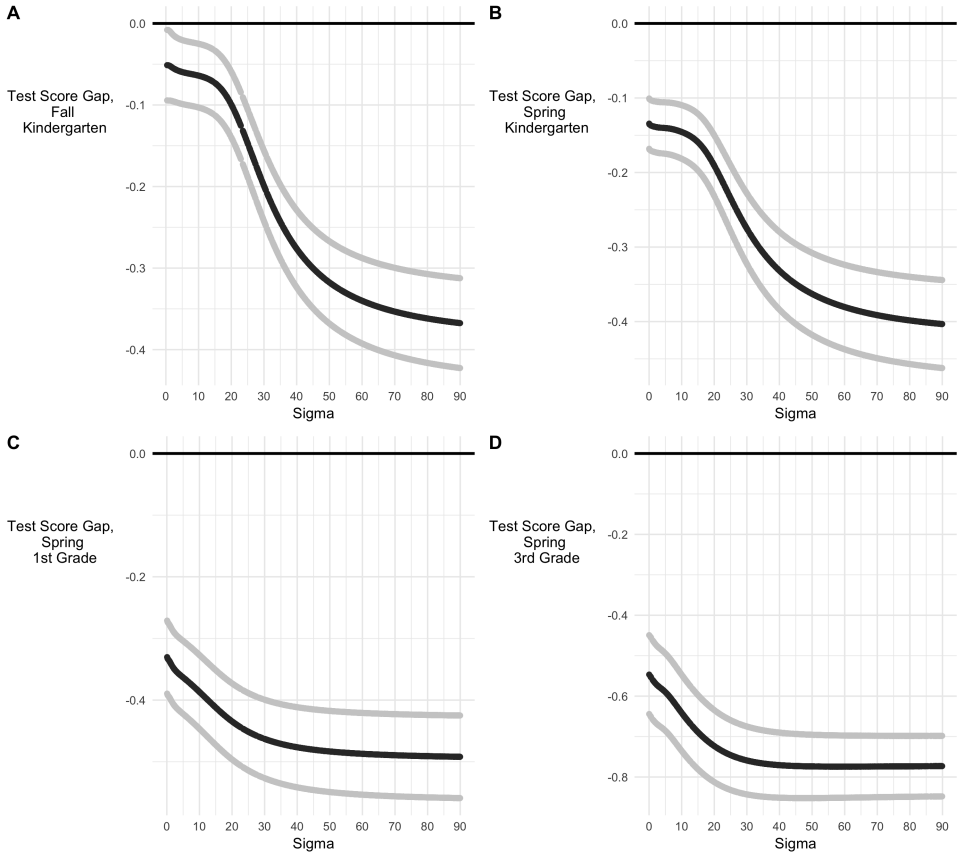
Notes: The dark lines represent the point estimates for a given specification with the corresponding sigma value. Lighter lines represent 95% confidence interval calculated with standard errors clustered by MSA-level. Each panel refers to a different specification used to test prediction 1. Panel A intentionally omits the unemployment rate and additional MSA-level controls. Panel B includes the unemployment rate but intentionally omits additional MSA-level controls. Finally, panel C includes the unemployment rate and additional MSA-level controls.

Figure A12: Simulation Results for Nunn and Wantchekon (2011), CDF Transformations



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 5 of Nunn and Wantchekon (2011) presenting IV estimation results. 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.

Figure A13: Simulation Results for Bond and Lang (2013), CDF Transformations



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.

Supplemental Appendix 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.

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.