

Kaiser, Caspar F.; Lepinteur, Anthony

Working Paper

Measuring the Unmeasurable? Systematic Evidence on Scale Transformations in Subjective Survey Data

IZA Discussion Papers, No. 18029

Provided in Cooperation with:

IZA – Institute of Labor Economics

Suggested Citation: Kaiser, Caspar F.; Lepinteur, Anthony (2025) : Measuring the Unmeasurable? Systematic Evidence on Scale Transformations in Subjective Survey Data, IZA Discussion Papers, No. 18029, Institute of Labor Economics (IZA), Bonn

This Version is available at:

<https://hdl.handle.net/10419/325087>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

DISCUSSION PAPER SERIES

IZA DP No. 18029

**Measuring the Unmeasurable? Systematic
Evidence on Scale Transformations in
Subjective Survey Data**

Caspar Kaiser
Anthony Lepinteur

JULY 2025

DISCUSSION PAPER SERIES

IZA DP No. 18029

Measuring the Unmeasurable? Systematic Evidence on Scale Transformations in Subjective Survey Data

Caspar Kaiser

*Warwick Business School, University of Warwick,
Wellbeing Research Centre and University of Oxford*

Anthony Lepinteur

University of Luxembourg and IZA

JULY 2025

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Measuring the Unmeasurable? Systematic Evidence on Scale Transformations in Subjective Survey Data*

Economists routinely use survey measures of, for example, risk preferences, trust, political attitudes, or wellbeing. The literature generally treats numerical response categories as if they represent equal psychological intervals. We provide the first systematic test of this assumption, developing a general framework to quantify how easily results can be overturned when this linearity assumption is relaxed. Using original experimental data, we show that respondents interpret survey scales in ways that do deviate from linearity, but only mildly. Focusing on wellbeing research, we then replicate 30,000+ coefficient estimates across more than 80 papers published in top economics journals. Replicated coefficient signs are remarkably robust to mild departures from linear scale-use. However, statistical inference and estimates of relative effect magnitudes become unreliable, even under modest departures from linearity. This is especially problematic for policy applications. We show that these concerns generalise to many other widely used survey-based constructs.

JEL Classification: I31, C18, C87

Keywords: Likert scales, ordinal scales, wellbeing, life satisfaction, survey methods

Corresponding author:

Anthony Lepinteur
University of Luxembourg
2 Av. de l'Université
4365 Esch-sur-Alzette
Luxembourg
E-mail: anthony.lepinteur@uni.lu

* We thank Dan Benjamin, Anita Braga, Matthew Cashman, Elena Fumagalli, Leonard Goff, Ori Heffetz, Martijn Hendricks, Richard Heys, Christos Makridis, Giorgia Menta, Andrew Oswald, Alberto Prati, Marco Ranaldi, Carsten Schröder, Claudia Senik, Robert Stüber, Mattie Toma, as well as seminar and conference participants at the London School of Economics, University of Leeds, Brno, University of Groningen, Warwick Business School, DIW, Freie Universität Berlin, General Conference of ISQOLS Rotterdam, Regional Conference of ISQOLS Johannesburg, General Conference IARIW 2024, Nanyang Technological University, University of Alcalá, Measuring Progress Workshop (Statec), for their helpful comments and suggestions on earlier drafts of the paper. All errors remain our own.

1 Introduction

Ordered response scales are ubiquitous in economics, but their interpretation rests on an untested assumption: that numerical labels reflect equal psychological intervals. The contribution of this paper is to provide the first systematic test of this assumption, developing a general framework to quantify how easily empirical results can be overturned when this linearity assumption is relaxed. Using original experimental data, we show that respondents use survey scales in ways that deviate from linearity, but only mildly so. We then replicate 30,000+ coefficient estimates across more than 80 papers published in top economics journals. Coefficient signs are remarkably robust to the mild departures from linear scale-use we document experimentally. However, estimates of relative effect magnitudes — which are crucial for policy applications — are highly unreliable even under these modest non-linearities.

Across the social sciences, ordered response scales, or ‘*Likert scales*’, are the default instrument for measuring latent constructs like political preferences, risk attitudes, wellbeing, trust, etc. These scales are easy to administer and, for many disciplines, have proved pivotal for answering questions that cannot otherwise be answered with behavioral data. Yet, some scepticism, especially among economists, remains over the validity and use of Likert scale measures. Three concerns underlie such scepticism. The **first** concern focuses on whether commonly used survey items really do capture the underlying constructs of interest — such as attributes of utility functions (e.g. risk aversion) or utility itself (e.g., subjective wellbeing as a proxy for flow-utility). See e.g. Bertrand and Mullainathan (2001) or Benjamin et al. (2023a). The **second** concern asks whether responses can be compared meaningfully between people and over time: does a reported “6 out of 10” mean the same for you as for me, or for me today as for me a year ago? See e.g. Angelini et al. (2014), Fabian (2022), Kaiser (2022), or Prati and Senik (2025). The **third** concern involves the relationship between the numerical labels that researchers attach to ordered response categories (i.e., “1”, “2”, “3”, etc.) and how these map onto the unobserved latent variable that researchers are trying to measure.

We focus on the third concern. The core issue is this: *we do not know the functional form of the relationship between reported scale values and the underlying latent variable*. Even if all respondents use the scale in approximately the same way, does a one-unit difference on the response scale represent the same magnitude of change in the latent variable across all parts of the scale? Or is this relationship non-linear, with differences between certain response categories representing larger gaps in the underlying construct than others?

Although this problem, of course, applies to any construct measured with Likert scales, much of the methodological work focused on wellbeing. Among the first to address this issue

is Ferrer-i Carbonell and Frijters (2004), who showed that coefficients estimated from an ordered logit or probit regression are very similar to estimated based on OLS regressions. Nevertheless, Oswald (2008) highlighted how a potentially non-linear “reporting function” (i.e. the mapping from underlying states to survey responses) could distort estimates of non-linear effects, such as estimates of the curvature of the income-to-wellbeing relationship. That paper also provided some empirical evidence to suggest that the reporting function is close to linear. Focusing on signs of coefficient estimates, Schröder and Yitzhaki (2017) provided conditions under which single-covariate regression results can be sign-reversed when allowing for a non-linear reporting functions. They also showed that such sign reversal can indeed occur in practice; as did Bloem (2022) who broadened the analysis to a broader broader class of non-linear functions.

Bond and Lang (2019) generalized these concerns. They demonstrated that virtually all empirical findings based on Likert scales can be reversed via appropriate monotonic transformations of the response scale. They argued that without strong assumptions about the distribution of the latent concept within response categories and about the functional form of the reporting function, it is impossible to draw definitive conclusions about the sign of differences between groups. In turn, Kaiser and Vendrik (2023) identified effect heterogeneities across the distribution of wellbeing as the underlying mechanism that drives potential sign reversals. They derived a general condition under which coefficients in OLS regressions with multiple covariates are reversible and applied this condition to a selected set of covariates.¹

However, we currently lack systematic evidence on how serious these concerns really are. Existing studies have only analysed a small number of *selected* datasets and variables. When results can be reversed in principle, we have no measure of how ‘easy’ it is to obtain such reversals, and thus how concerned we should be in practice. We also have surprisingly little direct evidence on how respondents actually interpret survey scales. This makes it difficult to assess which transformations are empirically plausible. Finally, while much attention has focused on coefficient signs, we know little about how non-linear transformations affect statistical significance or the relative magnitudes of estimates.

We address these gaps. To do so, we first introduce a ‘cost’ function C to quantify the extent to which any scale transformation departs from linearity. This cost function has a

¹All of these papers focus on the question of how a non-linear transformation of a linear reporting function would affect estimates of the conditional mean of underlying wellbeing. A few papers - specifically Chen et al. (2022) and Bloem and Oswald (2022) - have noted that estimates of the conditional median are invariant to such non-linear transformations, and in that sense more robust. This is valuable. However, since the (conditional) mean, rather than median, is the primary quantity of interest for many economic and policy applications, we will focus on it.

natural interpretation, with $C = 0$ indicating linear scale use, and $C = 1$ indicating (in a certain sense) ‘maximally’ non-linear scale use. This cost function enables us to numerically determine the “least non-linear” transformation capable of reversing regression results in terms of sign, significance, and relative magnitudes. The statistical machinery we develop is general: it applies to *any* bounded ordered scale.

Among such scales, subjective wellbeing measures stand out as both highly influential and highly scrutinized, making them a natural proving ground for our approach. Indeed, over the past half-century economists have built a sizeable literature that relies on Likert scales to study life satisfaction and happiness. Early contributions by Easterlin (1974) and Van Praag (1971) paved the way. The 1990s then saw a surge of papers linking self-reported wellbeing to income, unemployment and macroeconomic conditions (Clark and Oswald 1994; Oswald 1997; Blanchflower and Oswald 2004). Today, governments and international organizations collect, publish and even base cost–benefit calculations on wellbeing scales — see the UK Treasury’s 2022 *Green Book* update for instance (UK HMRC Treasury 2021). At the same time, the most recent and prominent methodological critiques — by Schröder and Yitzhaki (2017) and Bond and Lang (2019) — also targeted the wellbeing literature, making it the domain where addressing these concerns appears most urgent.

Using new experimental data, we first offer novel empirical evidence on how non-linear respondents’ scale use is in practice. We then reproduce the quasi-universe of wellbeing literature published in top-tier economics journals over the past fifteen years, creating an extensive database we call *WellBase*. In that section, we reproduce 72 papers, 1,601 regressions, with 3,163 coefficients of interest (and 28,513 coefficients overall). Using this dataset, we systematically assess the vulnerability of published findings to non-linear transformations.

Our results are as follows. Respondents, on average, interpret and use wellbeing scales in a manner that does deviate from linearity, but only mildly so. Our upper bound estimate serves as a benchmark for what we call *plausible* scale use. The relationship between the ‘cost’ of deviating from linearity and the risk of sign reversal is, as one might expect, concave. Approximately 20% of results published in leading economic journals are reversed with some transformation that has a *plausible* cost. Restricting ourselves to interpreting wellbeing data as merely ordinal (i.e. allowing for any departure from linear scale use), increases this share to about 60%.

We also show that the risk and cost of reversal are not merely random noise, but systematically related to identifiable features of research design. Certain design choices — in particular, leveraging exogenous sources of variation, such as natural experiments — are associated with substantially lower risk of sign reversal. Finally, while not itself a design choice, the level of statistical significance can serve as a useful signal of robustness: estimates with

higher significance levels are much less prone to reversals under *plausible* transformations.

We also examine risks of ‘significance reversals’. Estimates originally significant at the 0.1% level prove highly robust: roughly 94% remain significant at the 5% level even under a purely ordinal interpretation. However, estimates with p-values between 0.01 and 0.05 are highly vulnerable even under *plausible* transformations. The potential for non-linear scale use therefore makes reliable statistical inference considerably more challenging. Turning to relative magnitudes, we focus on unemployment and income as key determinants studied across multiple papers in our database. While coefficient signs for these are fairly robust, their relative magnitudes are highly sensitive to scale use assumptions. Marginal rates of substitution between unemployment and income can vary by an order of magnitude under plausible deviations from linearity.

Finally, we show that our results apply beyond wellbeing scales. We re-estimate 411 regressions from 14 published papers in top-five economics journals that use Likert-scale measures for, among others, risk aversion, social trust, and political preferences. In these replications, the prevalence and predictors of sign reversals closely mirror our wellbeing results.

The rest of the paper is structured as follows. The next section will provide the methodological background and introduces our cost-function approach. Section 3 empirically assesses respondents’ scale interpretations. Section 4 describes *WellBase* and present our results based on it. Section 5 concludes. The appendices provide proofs, additional discussion, and further results. Early versions of Stata routines for our proposed robustness analyses are available at: [link](#).

2 Analytical approach

This section provides the theoretical framework for our empirical analyses. We first note conditions under which regression coefficients maintain their sign across all monotonic transformations of the response scale and discuss how ratios of coefficients can be bounded. Versions of propositions 1-3 previously appeared in the working paper of Kaiser and Vendrik (2023). We here state them in our notation and provide several extensions and corrections. We then introduce a cost function to quantify departures from a linear response scale. This enables us to determine the minimal non-linearity required to reverse signs, change statistical significance, or alter relative magnitudes of coefficients.

2.1 Set-up and intuition

Consider a dataset containing responses to a survey question. For each individual i , responses are recorded using ordered categories: $r_i \in \{1, 2, \dots, k, \dots, K\}$. We also observe a vector of covariates \mathbf{X}_i .

These responses measure an underlying but unobservable state s_i .² We assume that respondents use the scale identically and that higher values of r_i correspond to higher levels of s_i . However, the functional relationship between r_i and s_i is otherwise unknown. This uncertainty motivates our analysis. We could transform r_i using any positive monotonic function f to obtain $\tilde{r}_i = f(r_i)$. Different transformations yield different interpretations of the response scale. The identity function $f(r) = r$ treats the scale as cardinal. Non-linear transformations alter the assumed ‘distances’ between response categories. Following Oswald (2008), we can interpret f as the inverse of a ‘reporting function’ that maps underlying states to survey responses.³

Throughout, we will be concerned with estimates from OLS⁴ regressions of \tilde{r}_i on \mathbf{X}_i :

$$\tilde{r}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(\tilde{r})} + e_i, \quad (1)$$

where e_i denotes the residuals. We use superscripts to distinguish coefficients from different transformations: $\hat{\beta}_m^{(\tilde{r})}$ denotes the coefficient on X_{im} from regressing \tilde{r}_i , while $\hat{\beta}_m^{(r)}$ denotes the coefficient from the standard cardinal specification.

We are interested in the stability of these coefficients across possible transformations f . A purely ordinal interpretation permits all positive monotonic transformations and deems them as equally viable. As we will show, in some instances, coefficient signs can be determined and relative magnitudes can be bounded even under this purely ordinal interpretation. However, in many instances, very little can be said under a purely ordinal interpretation. We therefore introduce a cost function C that quantifies how non-linear a given transformation of the response scale is. This function takes values between 0 (linear transformation) and 1 (maximally non-linear transformation). It thereby allows us to take an intermediate position between purely cardinal and purely ordinal interpretations of survey response data.

²For example, for a question about happiness, that underlying state would be the level of happiness the respondent is experiencing. In a question about trust, this state would be the subjectively ‘felt’ level of trust.

³If Oswald’s reporting function is $g : s \mapsto r$, then some transformation f satisfies $f = g^{-1}$.

⁴As Bond and Lang (2019) show for discrete and Kaiser and Vendrik (2023) for continuous covariates, the same concerns apply in principle to ordered probit/logit models.

2.2 Sign reversals

We now establish conditions under which the signs of estimates $\hat{\beta}_m^{(\tilde{r})}$ remain invariant across all positive monotonic transformations. Substantively, if the sign of $\hat{\beta}_m^{(\tilde{r})}$ does not change under any transformation, then how we would code survey responses would not affect our estimates of the sign of their association with X_{im} . To do so, define a new variable $d_{ki} \equiv \mathbf{1}(r_i \leq k)$ that dichotomises r_i at every response category. The following proposition then holds:

Proposition 1 (Non-reversal condition). *The sign of $\hat{\beta}_m$ is invariant under all positive monotonic transformations of r_i if and only if the estimates $\hat{\beta}_{km}^{(d)}$ on X_{im} from OLS regressions of d_{ki} on \mathbf{X}_i share the same sign for all $k = 1, \dots, K - 1$.*

The proof appears in Appendix A.1. This condition can be read as establishing whether first-order stochastic dominance of r_i with respect to some (possibly continuous) variable X_i holds. As we show in Appendices A.1.2 to A.1.4 this result extends to continuous outcomes, fixed effects, and two-stage least squares estimation.

Intuitively, this proposition shows that sign reversals require heterogeneities in the association of a covariate across the distribution of observed responses. An association is ‘heterogeneous’ in this sense when the signs of $\hat{\beta}_{km}^{(d)}$ are positive at some dichotomizations, but negative at others. In this case, variation in variable X_{im} pushes respondents up at some parts of the scale while pushing them down at others. Monotonic transformations can arbitrarily stretch or compress different parts of the scale to emphasize these opposing effects. Effectively, this allows us to ‘choose’ the sign of the average association.

Proposition 1 is a mechanical statement about the behavior of OLS regression coefficients and does not require any further assumptions. To connect estimates $\hat{\beta}_m^{(\tilde{r})}$ from regressions of \tilde{r}_i to underlying states s_i , we must make two additional assumptions: One assumption about the relationship between s_i and \mathbf{X}_i and one assumption on the relationship between \tilde{r}_i and s_i . Regarding the former, we assume a linear relationship between the underlying state and covariates:

Assumption 1 (Linear model). *The underlying state s_i is linear in X_i : $s_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$ with $(\varepsilon_i X_i) = 0$. The underlying state s_i is linear in X_i : $s_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$ with $E(\varepsilon_i X_i) = 0$.*

This assumption is not special to survey-based research and follows Angrist and Pischke (2009). We will not focus on it. Regarding the latter, we require that the measurement error from using a discrete response scale is reasonably well-behaved:

Assumption 2 (Favourable within-category heterogeneity). *For some $f(r_i) = \tilde{r}_i$, we have $s_i = \tilde{r}_i + \zeta_i$, where $\zeta_i = \mathbf{X}_i\boldsymbol{\gamma} + \vartheta_i$ with $E(\vartheta_i X_i) = 0$. For coefficient γ_m corresponding to X_{im} , either $\text{sgn}(\beta_m) \neq \text{sgn}(\gamma_m)$ or $\text{sgn}(\beta_m) = \text{sgn}(\gamma_m)$ and $|\beta_m| > |\gamma_m|$.*

We can think of ζ_i as a measurement error associated with discretizing continuous s_i to the discrete levels of \tilde{r}_i . The reason why we label Assumption 2 “favourable within-category heterogeneity” is because the coefficients on the measurement error ζ_i indicate how the underlying state varies across individuals within response categories. Substantively, we require that this within-category variation is either weaker than the corresponding variation across categories, or of the same direction as across categories.⁵ Section 3.3 and Appendix C provide empirical support for this assumption. With these assumptions in place, we can now state the following:

Proposition 2 (Non-reversal for underlying satisfaction). *Under Assumptions 1 and 2, when the condition of Proposition 1 holds, the sign of $\hat{\beta}_m$ from any transformation \tilde{r}_i consistently estimates the sign of β_m .*

See Appendix A.2 for the proof. Proposition 2 tells us that the sign of the association of some variable X_{mi} with underlying satisfaction s_i can be identified with data on r_i whenever the measurement error due to discretizing s_i is sufficiently well-behaved. If we do not maintain Assumption 2, i.e. when we are unwilling to place suitable restrictions on within-category heterogeneity in s_i , then estimates based on observed data on r_i and X_i can almost *always* fail to yield the correct sign for the direction in which s_i varies with X_i . Although not framed in those terms, this was previously pointed out by Bond and Lang (2019).

2.3 Coefficient ratios

Beyond coefficient signs, researchers often focus on the ratios $\hat{\beta}_m^{(\tilde{r})}/\hat{\beta}_n^{(\tilde{r})}$ of estimated coefficients corresponding to different covariates. Such ratios are frequently interpreted as marginal rates of substitution and are central to policy applications that derive monetary valuations from survey data (Frijters and Krekel 2021). Generally, the absolute magnitudes

⁵To gain some intuition on this assumption, consider a binary treatment $X_{im} \in \{0, 1\}$ where the true average treatment effect on the underlying state s_i is negative (i.e. $\beta_m < 0$). Now suppose that the treatment nevertheless contains two opposing effects: (1) it increases s_i for a few individuals such that they are shifted to a higher response category r_i , while (2) also lowering the underlying state s_i of most individuals within each category (who do **not** switch categories). In this case, the within-category measurement error ζ_i would be negatively correlated with the treatment (i.e. $\gamma_m < 0$), while the regression of r_i on X_{im} will show a positive association due to the positive between-category effect (i.e. $\hat{\beta}_m > 0$). Since both β_m and γ_m are negative, Assumption 2 is violated. Here, the estimate $\hat{\beta}_m$ (which is only based on between-category variation) would incorrectly indicate a positive treatment effect even though the true effect is negative.

of coefficients are meaningless. they can be freely changed by an arbitrary *linear* transformation of the response scale. Ratios of coefficients, in contrast, in virtue of being unaffected by linear transformations of the response scale, do provide a meaningful measure of the relative magnitude of a variable’s association with the the outcome of interest.

Unfortunately, however, coefficient ratios are generally affected by non-linear transformations. The only exception occurs when the corresponding ratios $\hat{\beta}_{km}^{(d)}/\hat{\beta}_{kn}^{(d)}$ from regressions of dichotomized variables d_{ki} are constant across all k . Empirically, this is never the case. However, whenever the coefficient in the denominator is not reversible, we can establish bounds on this ratio:

Proposition 3 (Bounded coefficient ratios). *If and only if $\hat{\beta}_n^{(\tilde{r})}$ in the denominator is not reversible across all positive monotonic transformations of r_i , the ratio $\hat{\beta}_m^{(\tilde{r})}/\hat{\beta}_n^{(\tilde{r})}$ is bounded by the minimum and maximum values of $\hat{\beta}_{km}^{(d)}/\hat{\beta}_{kn}^{(d)}$ across all $k = 1, \dots, K - 1$.*

The proof is provided in Appendix A.4.⁶ Unfortunately, these bounds will often turn out to be impractically wide (c.f. section 4.2.4). This, in part, motivates the material of the next section.

2.4 Quantifying non-linear scale use

Thus far, we were concerned with the possible behavior of estimated coefficients when treating any transformation $f(r) = \tilde{r}$ of r as an equally viable interpretation of the response scale. However, while some degree of non-linearity in response scales seems plausible, extreme transformations strain credulity. Consider a transformation that compresses categories 1-10 into a tiny interval while stretching category 11 across most of the scale. Such a transformation, while mathematically valid, seems to be, at best, an unusual assumption about how respondents use survey scales. We therefore need a principled way to quantify how “extreme” a transformation is — that is, how far it departs from the standard assumption of linearity. We can then identify the minimal departure from linearity needed to overturn empirical results. Intuitively, if reversing a finding requires only a minor adjustment to category spacing, then a result is fragile. If a reversal demands an extreme transformation that finding is more robust.

To implement this idea, we need some additional notation. Let l_k denote the (real) value assigned to response category k in the original coding of r_i , with $l_k = k$ in the standard rank-order coding. Similarly, let \tilde{l}_k denote the value assigned to category k in some transformed coding $\tilde{r}_i = f(r_i)$, where $f(l_k) = \tilde{l}_k$. To quantify deviations from linearity, we propose a cost

⁶Kaiser and Vendrik (2023) make an error in their corresponding proposition, since they fail to consider the case where $\hat{\beta}_n^{(\tilde{r})}$ has a reversible sign.

function that measures how much the differences between adjacent response categories vary. Specifically, let $\Delta\tilde{\mathbf{I}}$ capture these differences, with $\Delta\tilde{\mathbf{I}} \equiv [\tilde{l}_2 - \tilde{l}_1, \tilde{l}_3 - \tilde{l}_2, \dots, \tilde{l}_K - \tilde{l}_{K-1}]$.

We now propose a family of cost functions of the form:

$$C_\alpha(\tilde{\mathbf{I}}) = \left(\frac{\text{Var}(\Delta\tilde{\mathbf{I}})}{\max\text{Var}(\Delta\tilde{\mathbf{I}})} \right)^{1/\alpha}, \quad (2)$$

where $\text{Var}(\Delta\tilde{\mathbf{I}})$ denotes the variance of the differences in labels, while $\max\text{Var}(\Delta\tilde{\mathbf{I}})$ represents the maximum possible variance of these differences. In Appendix A.5, we show that $\max\text{Var}(\Delta\tilde{\mathbf{I}}) = \left(\frac{1}{K-1} - \frac{1}{(K-1)^2} \right) (l_K - l_1)^2$.

Any $\alpha > 0$ yields a cost function that is bounded between 0 and 1, with 0 representing perfect linearity and 1 representing maximal non-linearity (i.e., a single ‘‘jump’’).⁷ Generally, smaller values for α make the cost function more lenient, allowing for stronger non-linearities at lower cost values. For $\alpha = 2$, this cost function gives the ratio of the standard deviation to the maximum standard deviation of differences between adjacent labels. As respectively illustrated in Figures 1 and A21, this setting for α yields reasonable transformations across levels of C in the case of commonly used 11-point and 7-point scales. We will use this setting for α in the empirical sections and there drop the α subscript.⁸

We can now use this cost function to quantify the robustness of empirical findings. The general approach is to find transformations that minimize C_α subject to a set of appropriate constraints. Two constraints for this optimization problem are shared across all applications:

- 1. Normalization:** The ‘length’ of the scale must be preserved: $l_K - l_1 = \tilde{l}_K - \tilde{l}_1$.
- 2. Monotonicity:** Transformed labels must be strictly increasing: $\tilde{l}_k - \tilde{l}_{k-1} > 0 \forall k \geq 2$.

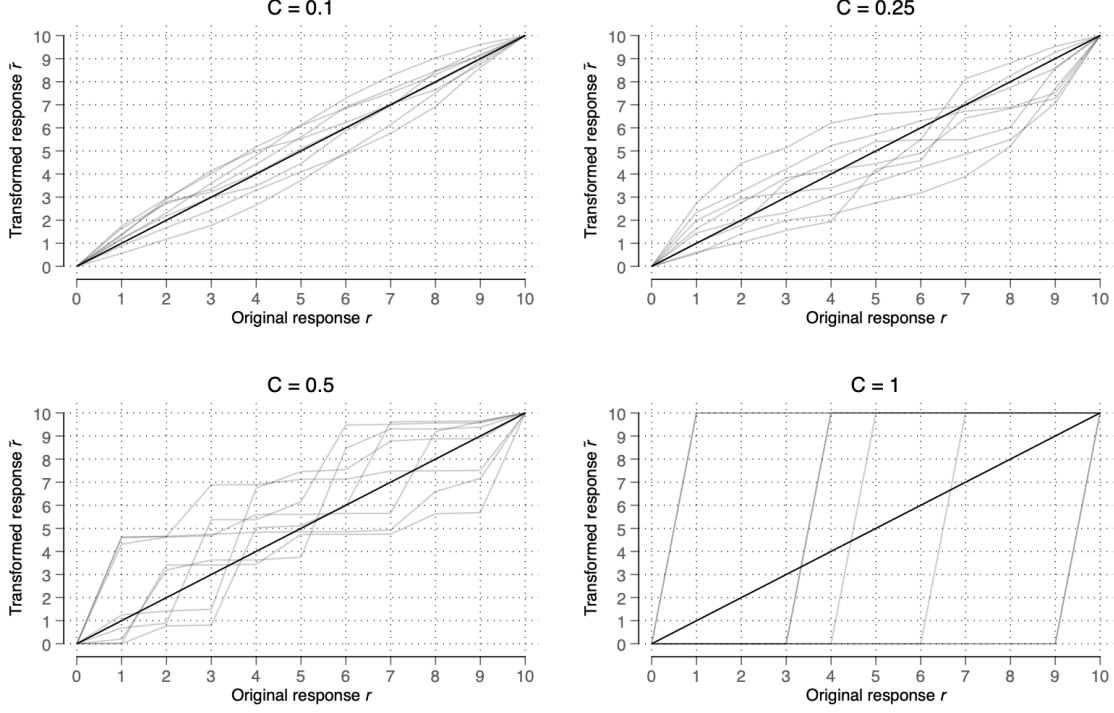
Here, the **Normalization** constraint ensures that transformations preserve the overall range of the scale.⁹ This prevents arbitrary stretching or compression that would make

⁷While it might here seem natural to adopt some standard inequality measure, such as the Gini coefficient or Atkinson index, these do not take a value of 1 under single-jump transformations when the number of response categories is finite. Instead, their maximum depends on the number of available response options K . Among the standard inequality measures we are aware of, only a normalized Theil index would avoid this limitation, but we prefer our variance-based family for its more intuitive interpretation.

⁸However, when the number of categories becomes large (e.g., 100 categories when approximating a continuous scale), a fixed value for α becomes problematic. In such cases, it becomes possible to achieve visually strong non-linearities even for small values of C . As we derive in Appendix B, this occurs because, as the number of response options K increases, the variance of differences between adjacent labels scales by a factor $\frac{1}{(K-1)^2}$ for any fixed (smooth) transformation function. To render the extent of non-linearity comparable across scales with varying numbers of response options, we propose setting $\alpha = 2 \log_{10}(K - 1)$. Apart from the general advantages associated with allowing α to depend on the log of K , this particular adjustment has the advantage of yielding $\alpha = 2$ for the common 11-point scale.

⁹Some papers study potential stretching of the scale across respondents while maintaining the linearity

Figure 1: Examples of scale transformations with different costs $C_{\alpha=2}$.



Notes: The figure shows different ways how respondents might interpret response scales. Specifically, each panel shows several randomly selected ways to transform an 11-point response scale. Within each panel, the displayed transformations all satisfy a given cost $C_{\alpha=2}$ displayed at the top of each panel. The horizontal axis represents the original scale r . The vertical axis shows the transformed scale $f(r) = \tilde{r}$. The straight 45-degree line in each panel represents linear scale use, i.e. the standard assumption that the difference between choosing “3” versus “4” means the same as choosing “7” versus “8”. As our cost $C_{\alpha=2}$ increases from 0 to 1, transformations increasingly depart from this linear benchmark. At the extreme of $C = 1$, the scale collapses to a single jump. Here, all response options below some threshold represent the same mean level of the underlying state, while all above represent another level.

comparisons meaningless. The **Monotonicity** constraint forces that only positive monotonic transformations are considered. We thereby ensure that higher response categories always map to higher transformed values.

We then need a third constraint that depends on our application. For example, if we are interested in reversing coefficient signs, we need the sign of $\hat{\beta}_m^{(\tilde{r})}$ to be different from $\hat{\beta}_m^{(r)}$:

3a. Sign Reversal: $\text{sgn}(\hat{\beta}_m^{(\tilde{r})}) \neq \text{sgn}(\hat{\beta}_m^{(r)})$.

On the other hand, for coefficient ratios, we should constrain ourselves to achieving some assumption. See e.g. Benjamin et al. (2023b) or Fabian (2022). A fruitful avenue for future work is to combine these research streams.

target ratio within the bounds identified by Proposition 3:

3b. Fixed Ratio: $\hat{\beta}_m^{(\tilde{r})}/\hat{\beta}_n^{(\tilde{r})} = \rho$ for some target ratio ρ .

For statistical inference, our constraint would require the p-value $p(\hat{\beta}_m^{(\tilde{r})})$ to cross a chosen significance level. This is outlined in more detail in the next section. For any application, we then solve:

$$\tilde{\mathbf{I}}^* = \operatorname{argmin}_{\tilde{\mathbf{I}}} C_\alpha(\tilde{\mathbf{I}}), \quad (3)$$

subject to the relevant constraints. In general, there may not be a unique solution to this optimization problem. However, for any solution, $C_\alpha(\tilde{\mathbf{I}}^*)$ quantifies the minimal departure from linearity required to achieve the specified objective; be that a sign reversal, a ‘significance’ reversal, or achieving a given relative effect magnitude.

2.5 Statistical inference

In most cases we are not only interested in the signs and ratios of estimated coefficients, but also in their statistical significance. To assess how significance levels change under monotonic transformations, we need the variance-covariance matrix of $\hat{\beta}^{(\tilde{r})}$ from regressions of any transformed variable \tilde{r}_i . The variance-covariance matrix takes the standard form:

$$\operatorname{Var}(\hat{\beta}^{(\tilde{r})}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \quad (4)$$

where $\hat{\Omega}$ is an estimate of the covariance matrix of the residuals. The form of $\hat{\Omega}$ depends on the assumed error structure, but in all cases it depends only on the residuals $\tilde{\mathbf{e}}$ and (for clustered errors) the design matrix \mathbf{X} . Usefully, the residuals from a regression of any \tilde{r}_i on \mathbf{X}_i can be expressed as a weighted combination of residuals from the corresponding dichotomized regressions of d_{ki} . As shown in Appendix A.3.1 we have:

$$\tilde{\mathbf{e}} = \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{e}_{dk}, \quad (5)$$

where \mathbf{e}_{dk} denotes the vector of residuals from regressing \mathbf{d}_k on \mathbf{X} .

This decomposition allows us to compute $\hat{\Omega}$ for any transformation using only results from the $K - 1$ dichotomized regressions of d_{ki} . Appendix A.3.2 provides explicit expressions for homoskedastic, heteroskedasticity-robust, and clustered standard errors in terms of these weighted residuals. Once we have the variance-covariance matrix, expressions for standard errors and p-values for any coefficient under any transformation follow immediately.

We can now make use of the cost function framework of the previous section. First, we can obtain bounds on p-values $p(\hat{\beta}_m^{(\tilde{r})})$ associated with any coefficient $\hat{\beta}_m^{(\tilde{r})}$ for *any* positive monotonic transformation of r .¹⁰ We can do so by numerically maximizing (for an upper bound) or numerically minimizing (for a lower bound) $p(\hat{\beta}_m^{(\tilde{r})})$ subject to constraints (1)-(2) of the previous section. With such bounds in hand, it is possible to specify some fixed p-value as a constraint on the optimization problem we previously specified:

3c. Fixed P-value: $p(\hat{\beta}_m^{(\tilde{r})}) = \pi$.

We then solve the optimization problem of Equation 3 subject to constraints (1)-(3c). By choosing π appropriately (e.g. $\pi = 0.05$), this allows us to assess how non-linear we require transformations of r_i to be in order to turn a statistically significant result into a statistically insignificant one, and vice versa.

3 How are response options interpreted?

Our cost function approach of the previous section is based on the idea that more extreme departures from a linear interpretation of the response scale are increasingly unlikely. The case of $C = 1$, where there is only a single ‘jump’ in the underlying state for some two adjacent response categories, and no differences in the underlying state for all other response categories, is an example of a clearly unnatural interpretation of the response options.

That said, there is only limited work on how respondents use response options in empirical practice. Any evidence that we do have is rather indirect. For example, there is some work from psychophysics on how individuals subjectively interpret numerical stimuli (Banks and Coleman 1981; Banks and Hill 1974; Schneider et al. 1974). These studies were concerned with how the subjective intensity of numbers relates to their objective magnitudes. For bounded intervals – which seem analogous to bounded survey scales – the relationship between objective numerical values and their subjective interpretations is approximately linear.

Evidence specifically relating to bounded response scales in survey data was previously discussed in Van Praag and Van der Sar (1988), Van Praag (1991) and Van Praag et al. (1999). They broadly conclude that scale use is likely to be fairly linear. Similarly, Kaiser and Oswald (2022) find that the relationship between reported satisfaction (in the domains of jobs, relationships, housing, and health) and the subsequent probability of taking a quitting

¹⁰Note here that is not the case that the p-value $p(\hat{\beta}_m^{(\tilde{r})})$ associated with some estimated coefficient $\hat{\beta}_m^{(\tilde{r})}$ is bounded by the smallest and largest p-values obtained from corresponding regressions of d_{ki} . For example, when $\hat{\beta}_m$ is reversible, we can find some transformation where $p(\hat{\beta}_m^{(\tilde{r})}) = 1$ despite $p(\hat{\beta}_{km}^{(d)}) < 1 \forall k$.

action (i.e. switching jobs, divorce, changing flats, going to hospital) is close to linear. They are able to replicate this result across multiple datasets from the UK, Germany, and Australia. Under the assumption that the association between underlying satisfaction and the probability of taking an action is linear (for which they give some arguments), this is also evidence of a close-to-linear relationship between *reported* and *underlying* satisfaction.

However, none of the previous studies provide a clear upper bound for our cost C . In the material below, we therefore attempt to find an upper bound for C . In section 3.3 we additionally present novel evidence on whether Assumption 2 is likely to hold.

3.1 Data and approach

We rely on original data collected from a sample of $N = 1,268$ participants recruited via Prolific. We sought for this sample to be nationally representative of the adult population of the United Kingdom. See Appendix Table A7 for further details on data collection and Appendix Table A8 for descriptive statistics. With this data we implement four different methods to estimate C . Given that our primary interest in our replication effort of section 4 is on overall life satisfaction, our attempts at bounding C also tend to be specific to overall life satisfaction. As will become apparent, these methods disagree in their substantive conclusions about the particular *shape* of respondents’ scale use. But they do agree on the likely extent to which scale use is non-linear.

3.1.1 Linear prompting

For our first method, we randomized participants into two conditions. One half of participants is given a standard life satisfaction question: ‘*Overall, how satisfied are you with your life nowadays?*’.¹¹ The other half received the same question, but we added the following prompt: ‘*Please treat the scale below as linear. For example, the difference in satisfaction between options “4” and “5” should be treated as just as large as the difference between options “6” and “7”.*’. Thus, in the second group, we directly ask respondents to use the scale in a linear fashion. In both conditions, after respondents gave their discrete answer, they were also asked about their satisfaction level *within* the chosen category. We therefore obtain both a discrete and continuous measurement of r . See Appendix D for screenshots of how the relevant survey questions were presented to respondents.

To make an inference about deviations from linear scale use in the unprompted case, we need two assumptions. We state these informally. First, we assume that respondents

¹¹This phrasing exactly follows the phrasing used by the UK’s Office of National Statistics in the Annual Population Survey. See: [link](#).

adhere to our linearity prompt. Second, we assume that the distribution of underlying satisfaction is the same across both groups. Given randomization, the latter assumption is reasonable. With these assumptions in place, we proceed as follows. For every value $r_{un}^{(disc)} \in \{0, 1, \dots, k, \dots, K\}$ of the unprompted discrete satisfaction data, we find that value $r_{lin}^{*(cont)}$ from continuous data in the linearly prompted group which satisfies $F_{un}(r_{un}^{(disc)} = k) = F_{lin}(r_{lin}^{*(cont)})$. Here, F_{un} and F_{lin} respectively denoted the cumulative distribution functions of $r_{un}^{(disc)}$ and $r_{lin}^{(cont)}$. If it were the case that scale use is unaffected by the prompt – i.e. if respondents were using the scale in a linear fashion without being prompted to do so – then we should observe a linear relationship between $r_{un}^{(disc)}$ and $r_{lin}^{*(cont)}$. Any deviations from such a linear relationship, in turn, are indicative of non-linear scale use.

3.1.2 Objective-subjective questions

Our second and third methods replicate and extend a method first proposed by Oswald (2008). Towards the start of the survey, we ask respondents to subjectively rate both their height and weight on a scale from 0 to 10. Specifically, we ask ‘*How tall are you?*’ (‘*How heavy are you?*’), with extremes labelled as 0=‘*Extremely short (light)*’ and 10=‘*Extremely tall (heavy)*’. These scales are made to look identical to the scales for our life satisfaction question (see Appendix Figure A14). Towards the end of the survey – after all subjective questions are answered – we then ask respondents about their ‘objective’¹² height (in feet and inches) and weight (in stone). Using these data, we can in turn compute the mean objective height and weight within each response category. From these, we read off in how far, expressed in terms of our cost C , respondents’ average scale use deviated from linearity. Under the assumption that scale use for questions on height and weight is comparable to scale use for questions on life satisfaction (c.f Benjamin et al. (2023b)), these estimates are in turn informative about non-linearities in scale use for life satisfaction questions.

3.1.3 Interactive sliders

Our fourth method relies on an interactive online application. We directly ask how respondents interpret the scale.¹³ We do so in three steps (see Appendix Figures A15-A18 for screenshots and this [link](#) for an interactive demo). In the first step, we explain to respondents that scale use might be non-linear. In the second step, as a comprehension check, we ask respondents to graphically indicate, using a set of interactive sliders, a particular pre-specified type of non-linear scale use (specifically a case in which the difference between a ‘3’

¹²Of course this is still self-reported, but is objective in the sense that no subjective scale is used.

¹³We only asked this method to participants that were not prompted to use a linear scale.

and a ‘4’ is larger than the difference between a ‘7’ and an ‘8’). We only proceed with those respondents who succeed in this comprehension test (82%). In the third and final step, we then ask respondents to indicate their own scale use with the same set of interactive sliders. To ease the cognitive burden on respondents, we provide respondents with several presets (incl. concave, convex, logistic, and inverse logistic scale use). Finally, whenever respondents do not move the sliders (implying linear scale use), we ask respondents to verify that they really did mean to indicate that their scale use was linear.

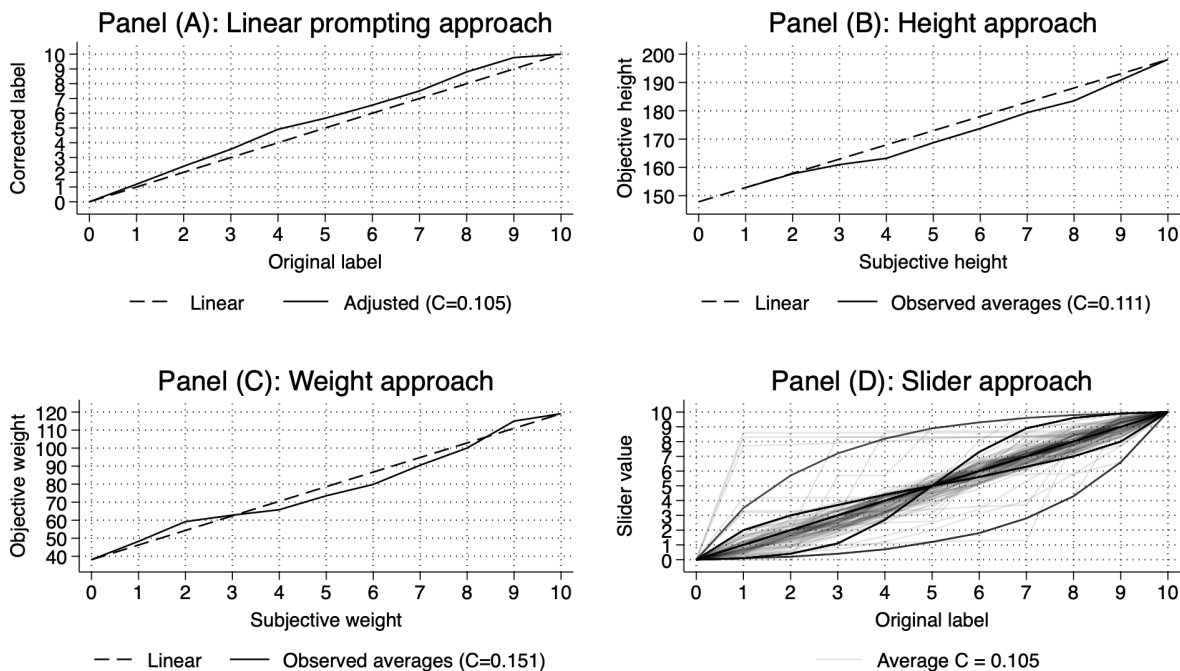
3.2 Results on scale use

Results are displayed in Figure 2. In each panel, the vertical axis represents the unadjusted data – either on life satisfaction (panels (A) and (D)), or on subjective height (panel (B)) or weight (panel (C)). In panel (A), the vertical axis gives $r_{lin}^{*(cont)}$ for each level of $r_{un}^{(disc)}$. For panels (B) and (C) the vertical axis respectively denotes objective height (converted to cm) and weight (converted to kg). Finally, the vertical axis in panel (D) shows the position of the slider for each response category of $r_{un}^{(disc)}$.

Across all methods we observe deviations from linearity. We use bootstrapping with 500 replications to obtain confidence intervals. The linear prompting approach gives evidence to imply that lower response categories – i.e between 0 and 4 – cover a slightly wider satisfaction range than the subsequent categories. Here, we obtain $C = 0.105$ (95% CI: 0.078 – 0.153). In the height approach, categories 3 and 4 cover a relatively smaller range, while categories 8-10 cover a wider range. This in turn yields $C = 0.111$ (95% CI: 0.102 – 0.178). The weight approach yields broadly similar, though more pronounced, results ($C = 0.151$; 95% CI: 0.115 – 0.229). Finally, the sliders approach yields a substantial share of individuals who state that their scale use is linear (42%). Among the remaining 58%, some selected the concave (11% of total), convex (9% of total) or other presets (9% of total). About a third of respondents (30% of total) were idiosyncratic in their self-reported scale use. Taking the average C across all respondents, we obtain $C = 0.105$ (95% CI: 0.095 – 0.115).

Hence, across methods, our point estimates for C range between 0.105 (sliders and linear prompting) and 0.151 (weight). All estimates differ statistically significantly from zero at any conventional level (with $p < 0.01$). However, these approaches yield **inconsistent** results regarding how individuals interpret the relative differences between response options: The solid lines in each panel have markedly different shapes, indicating disagreement about the specific form of non-linearity. This disagreement reflects both methodological differences (height and weight questions measure different constructs than life satisfaction) and the inherent difficulty of eliciting subjective scale interpretations. Yet, despite this disagreement

Figure 2: How do people use response scales? Converging evidence of mild departures from linear scale use across methods.



Notes: Four different approaches to measuring scale non-linearity all point to similar conclusions. Panel (A) is based on a randomized experiment: half our sample answered a standard satisfaction question, while the other half received explicit instructions to treat the scale linearly. The solid line shows how we would need to adjust response labels in the standard group to match the distribution of the linearly-prompted group. In Panel (B) we first asked respondents to subjectively rate their height (0=“extremely short” to 10=“extremely tall”) and then asked for their actual ‘objective’ height. The graph displays the average objective height within each subjective category. Panel (C) repeats this exercise for weight. In Panel (D) respondents were given interactive sliders and asked to indicate how they personally interpret the gaps between satisfaction scale points. Each gray line represents one respondent’s interpretation. Across all four methods, we find that people interpret scales in ways that deviate from perfect linearity, but only mildly so. The ‘cost’ C , which quantifies departure from linearity (where 0=perfectly linear and 1=maximally non-linear), ranges from 0.105 to 0.151 across methods. Based on a nationally representative sample of $N \approx 1,200$ UK residents recruited via Prolific.

about *shape*, we do obtain convergent evidence to suggest that the *extent* of non-linearity in scale use is, at most, modest. No method suggests departures from linearity anywhere near the more extreme transformations shown in the bottom panels of Figure 1. For strongly non-linear scale use (say, $C > 0.3$) to be viable, all four of our quite different approaches would need to be systematically biased toward linearity. While we cannot rule this out entirely, it seems unlikely that diverse methods would all err in the same direction. On this basis, we

conclude that reporting functions substantially more non-linear than allowed by $C = 0.3$ – twice our maximum observed estimate – are unlikely.

This upper bound provides an empirical anchor for labelling scale transformations in the analyses that follow. We will synonymously call transformations with $0 \leq C \leq 0.15$ “*plausible*” or “*mild*”. Transformations with $0.15 \leq C \leq 0.30$ will be called “*marginal*” or “*conservatively-plausible*”. Transformations with $0.30 \leq C \leq 1.00$ will be qualitatively labelled “*implausible*” or “*unlikely*”. These names are, of course, tentative, and should be revised against future evidence.

3.3 How does satisfaction vary within response options?

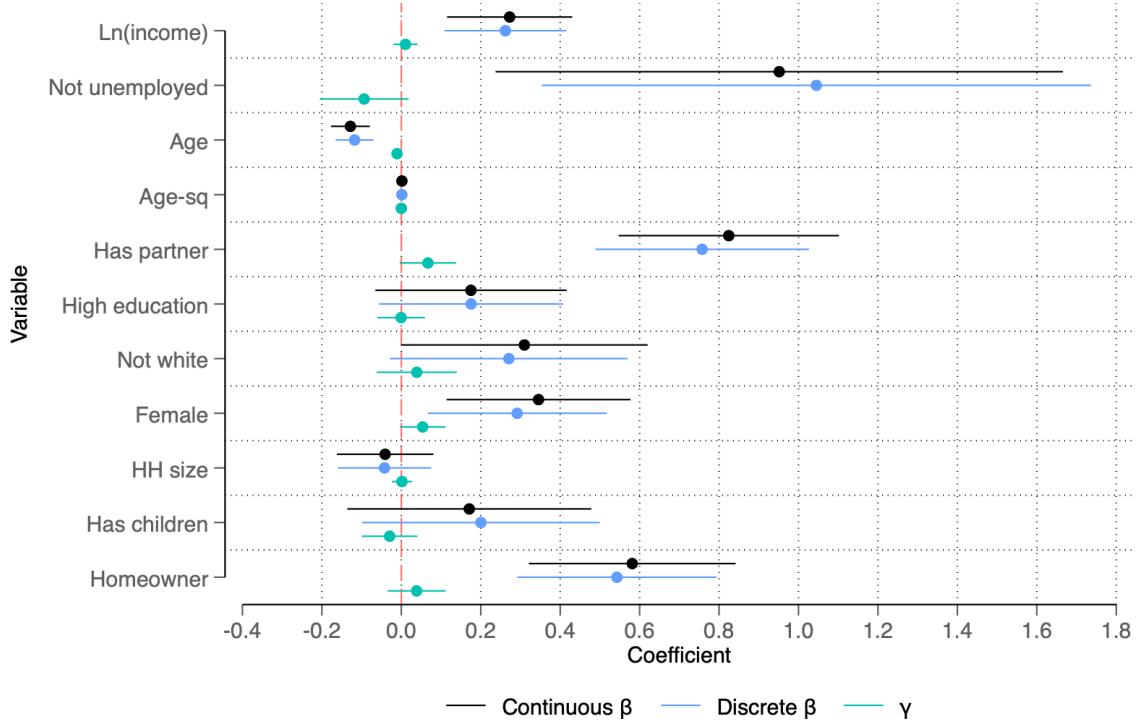
The robustness of the empirical literature – to be assessed in section 4 – depends both on plausible values for C and on whether Assumption 2 is met. This assumption is concerned with potential complications arising from discretizing the response scale, rather than with uncertainty over the choice of f and over what cost C is permissible. Here, we take both continuous and discrete measurements and compare results. This allows to assess whether discretizing poses any special problem. We obtain the required data by first asking respondents about their discrete satisfaction, and then asking a follow-up question about their satisfaction level *within* the chosen category. See Appendix Figure A13 for an illustration of how this follow-up question was presented to respondents.

With this data, we formally evaluate what γ_m (which is key to Assumption 2) would be for each covariate m if scale use were linear (i.e., if $C = 0$). We rely on the following argument: The coefficient γ_m is intended to capture systematic within-category heterogeneity in underlying satisfaction as it relates to covariate X_{im} .¹⁴ When satisfaction is measured on an increasingly granular scale, there is minimal scope for such heterogeneity to emerge. This implies that γ_m should approach zero. Therefore, when we assume that $C = 0$, i.e. that scale use is linear, we may consider a (quasi-)continuous measurement of satisfaction r_i^{cont} to be a reasonable proxy for underlying satisfaction, i.e., $r_i^{cont} \approx s_i$. On this basis, we estimate two regressions: one using discrete measurements r_i^{disc} and another using continuous measurements r_i^{cont} . The difference in estimated coefficients, $\hat{\beta}_m^{cont} - \hat{\beta}_m^{disc}$, gives us an estimate of γ_m .¹⁵

¹⁴More specifically recall that in Assumption 2, $\zeta_i = \mathbf{X}_i\gamma + \vartheta_i$ represents the measurement error associated with discretizing continuous satisfaction. While other sources of measurement error may exist (e.g., misunderstanding questions, momentary distractions), we assume these are uncorrelated with our covariates X_{im} (i.e., classical measurement error), and therefore do not systematically bias our coefficient estimates beyond the discretization error we are explicitly modeling.

¹⁵To see this more formally, note that since we assume $r_i^{cont} \approx s_i$ for $C = 0$, we have $r_i^{cont} = \mathbf{X}_i\beta + \varepsilon_i$. For the discrete measure, we have $r_i^{disc} = \mathbf{X}_i(\beta - \gamma) + \varepsilon_i - \vartheta_i$. Thus, the difference in coefficients between the continuous and discrete regressions yields $\beta_m^{cont} - \beta_m^{disc} = \gamma_m$.

Figure 3: Discrete and continuous measures of satisfaction yield nearly identical estimates



Notes: Comparison of regression coefficients using either a continuous (black dots) or a discrete (blue dots) 11-point measure of satisfaction. The differences between these estimates (γ_m ; teal dots), represent measurement errors from discretization. Whiskers indicate 95% confidence intervals. Across all covariates, the γ_m estimates are close to zero and statistically insignificant. It makes little difference whether satisfaction is measured on a continuous or a discrete scale. This provides empirical support for Assumption 2. The coefficient patterns themselves align with the wider literature: satisfaction follows a U-shape with age (negative linear term, positive squared term), unemployment strongly reduces satisfaction, higher household income increases it, and having a partner is beneficial. Women report slightly higher satisfaction than men.

Figure 3 presents our results. The figure shows estimates $\hat{\beta}_m^{cont}$, $\hat{\beta}_m^{disc}$, and $\hat{\gamma}_m$ for a large set of standard socio-economic characteristics. We see that $\hat{\gamma}_m$ is close to zero and statistically insignificant (at the 5% level) in all cases. Moreover, for most variables, $\hat{\gamma}_m$ takes the same sign as $\hat{\beta}_m^{cont}$, causing the estimate of $\hat{\beta}_m^{disc}$ to be biased towards zero. The only case in which $\hat{\gamma}_m$ takes on a different sign than $\hat{\beta}_m^{cont}$ (which is necessary but not a sufficient condition for violating Assumption 2) occurs for unemployment and having children. However, for none of the variables in this analysis does Assumption 2 look to be violated – or indeed anywhere close to being violated. Overall, this is evidence in favour of Assumption 2.

To verify the robustness of this result, we replicated Figure 3 using three alternative datasets. In each of these datasets, we again observe a continuous and a discrete measurement

of either respondents’ satisfaction (in the case of the Benjamin et al. and the Prati & Kaiser dataset) or happiness (in the LISS dataset). Additional details about these datasets are given in Appendix Table A7. Descriptive statistics are given in Appendix tables A9-A11. The main methodological difference in these datasets compared to our own data is that answers to the continuous and the discrete question were given at different times in the survey. As a consequence, respondents were not forced to give their continuous answer as being located within a given discrete category. These additional results are given in Appendix Figures A5-A7. We reach broadly the same conclusions: In almost all cases, γ_m is statistically insignificant and of the same sign as β_m^{disc} – implying that Assumption 2 is satisfied. Across 42 coefficients in total, we only observe evidence for a violation of Assumption 2 twice: in the case of the higher education dummy in the Kaiser & Prati data and for the gender dummy in the LISS data. In both instances β_m^{disc} is positive, β_m^{cont} is negative, and γ_m is statistically significant at the 5% level. Thus, we again get strong evidence in favour of Assumption 2.

However, although it looks as though $\gamma_m \approx 0$ if scale use were linear (i.e., for $C = 0$), it remains unclear how γ_m would behave for non-linear scale use (i.e. $C > 0$). Given the evidence of section 3, we are especially interested in the case of $0 < C < 0.15$. It is not feasible to estimate γ for all possible transformations f that satisfy $C < 0.15$ (recall that any specific value of C picks out a *family* of transformations, and not one particular transformation). However, it *is* possible to perform a worst-case analysis with our data. We conduct this analysis in Appendix C, where we search for transformations that yield, for a given value for C , minimal and maximal coefficient values for either our continuous or our discrete measure. The results show no clear evidence for violations of Assumption 2. Nevertheless, these worst-case analyses do indicate that continuous measures of satisfaction are generally more susceptible to sign reversals than discrete scales.

4 Systematic Evidence from *WellBase*

Subjective wellbeing is becoming increasingly central to policy. Among constructs measured using ordered response scales, it also often is the focus of methodological critiques. Wellbeing therefore is the ideal proving ground for our framework. Drawing on our replication database we call *WellBase*, this section takes the first systematic look at how robust the empirical economics of subjective wellbeing really are. *WellBase* includes 72 papers, 1,601 regressions and 28,513 coefficients.

We use these replications to quantify three kinds of risks that can arise when analysts assume the response scale to be linear: (i) the risk that a coefficient’s *sign* changes after a positive monotonic transformation of the scale, (ii) the risk that its *statistical significance*

changes, and (iii) the extent to which such transformations can alter the *relative magnitudes* of point estimates. Because the same Likert-style measurement issues may affect other constructs in economics, we also benchmark wellbeing against scales for, among others, risk, trust and political preferences. In that analysis, we replicate 2,903 coefficients across 14 papers.

4.1 Data

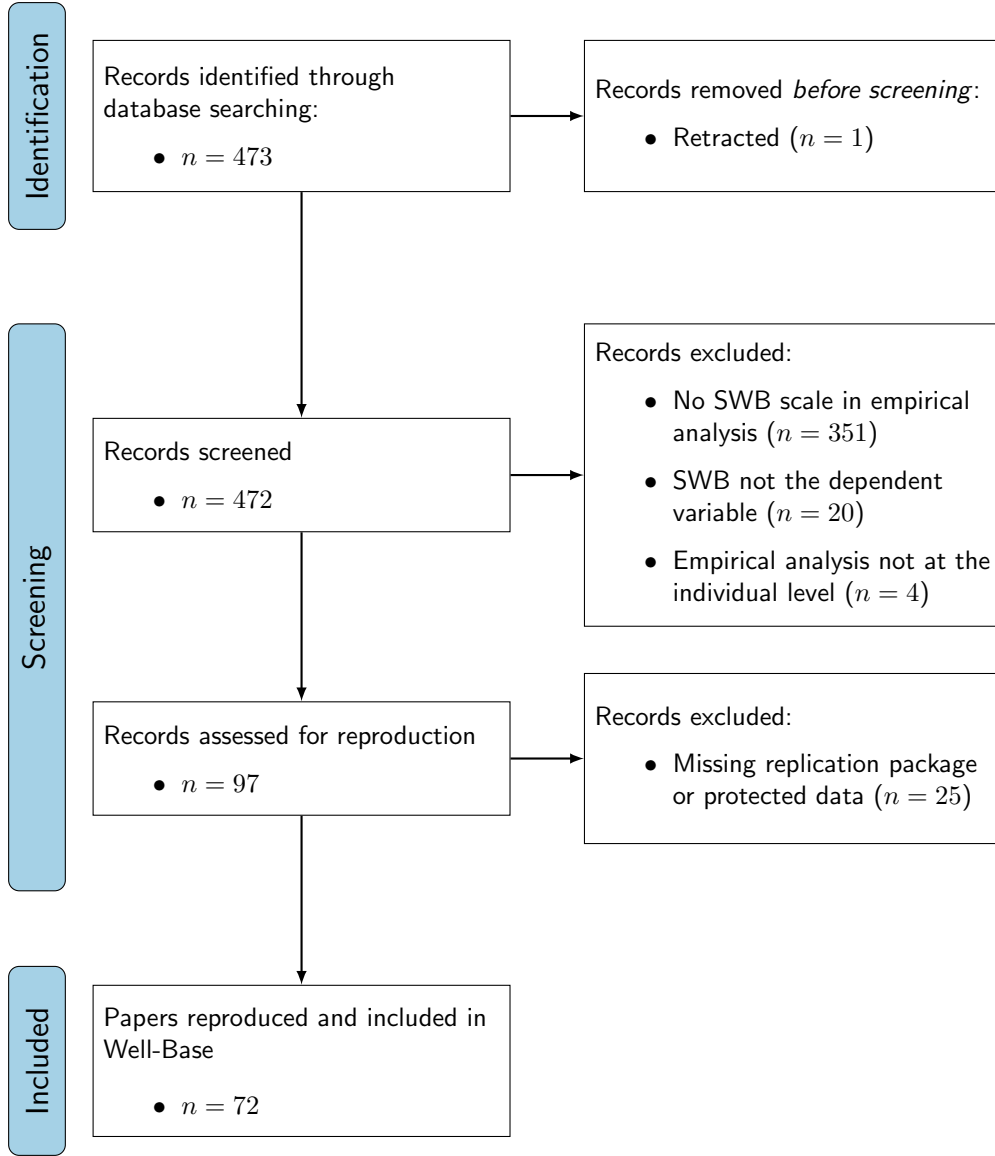
Our goal was to replicate the universe of empirical research on subjective wellbeing published in top economics journals. We had three inclusion criteria. First, we only included articles published in economics journals ranked among the Top 30 on RePEc (as of June 2022). These journals tend to enforce data- and code- sharing policies more stringently, rendering it more likely that a paper can be reproduced. Second, we only included papers published between January 2010 and May 2025. Third, we focus on papers that use a cognitive measure of subjective wellbeing as dependent variable in an individual-level analysis. Our search, conducted via Google Scholar, yielded 473 records based on the following keywords: “Life Satisfaction”, “Cantril Ladder”, “Subjective well-being”, and “Subjective wellbeing”. We chose “Life Satisfaction” and “Cantril Ladder” because they are among the most commonly used scales to measure cognitive subjective wellbeing. We added “Subjective well-being” and “Subjective wellbeing” to capture any papers employing less frequent cognitive wellbeing measures.¹⁶

See Figure 4 for a summary of our selection process. Approximately 75% of the initial records did not include a wellbeing scale in the empirical analysis (instead merely referring to subjective wellbeing in e.g. the literature review). 24 papers included cognitive measures of wellbeing but used them neither as dependent variables nor in an individual-level analysis. Overall, 97 of the initially identified papers fulfilled our inclusion criteria. Insufficient replication materials prevented us from reproducing 25 of these studies.

We therefore reproduced 72 articles. Among these, we successfully reproduced all of the 1,601 relevant regressions in both the main manuscripts and any associated appendices (printed or online). Less than 1% of these regressions (spread across five articles) were not using a linear estimator, but were using an ordered probit approach instead. Additionally, 3% percent of regressions (across two papers) were estimated using probit-adjusted OLS. To make these regressions comparable and to apply the methods of section 2, we reproduced these regressions using OLS. In all such cases, the results, in terms of sign and statistical significance, remained the same. Similarly, about 2% of the regressions (spread across three

¹⁶Such less frequently used measures may for example include questions on respondents *overall* wellbeing across their life (used in e.g. Clark and Senik (2010)).

Figure 4: PRISMA Chart



Note: PRISMA flowchart summarizing our selection process to produce *WellBase* (Page et al. 2021).

articles) used a binary dummy for high wellbeing as dependent variable. We re-estimated these regressions using the underlying 7- and 11-point versions of the wellbeing measure. Again, the results of these estimations remained the same.

Our replication effort yielded two categories of estimates: (1) published coefficients that form the core of each paper's analysis and (2) unpublished coefficients that typically serve as control variables mentioned only in table/figure notes. In total, we replicated 5,313 published estimates and 23,200 unpublished estimates. Table A1 provides a complete list of

all reproduced articles, along with a number of details such as the type of wellbeing scale used in the empirical analysis.

For each replicated regression, we record a set of additional regression characteristics. These include: measures of fit, number of control variables, the use and type of fixed effects. We also record sample characteristics, including sample size, average age, gender composition, and country composition, as well as distributional properties of the wellbeing measure. In addition, we collect detailed information on each independent variable: its type (e.g., indicator, categorical, or continuous), its distribution, the associated coefficient and standard error, whether it is instrumented, whether it represents socio-demographic characteristics, whether it is tied to a natural experiment, a randomized control trial, a macroeconomic indicator, a placebo, or some space or time fixed effects. In what follows, we will refer to this replication database as *WellBase*.

Table 1 provides an overview of the characteristics of the estimates we have replicated. About 6% of these estimates can be found directly in the published manuscripts. An additional 12% are reported in the appendices. The majority, constituting 81%, are coefficients on unprinted control variables not shown in the printed articles.¹⁷ About 4% relate to quasi-natural experiments (e.g., centralization reforms in Switzerland, the London Olympics, or RCTs), while another 4% are macroeconomic factors (like economic growth or inflation rates). Approximately 25% of coefficients relate to time-invariant characteristics (e.g., biological gender). Likewise, 25% of estimates are based on a continuous covariate (e.g., income or age). See Appendix Tables A3 and A4 for descriptive statistics at the regression level and paper level, respectively.

Appendix Table A2 focuses on the 27 papers in *WellBase* for which at least half of the printed regressions use a wellbeing scale as dependent variable. In these studies, the main objective is to uncover the drivers of subjective wellbeing.¹⁸ For each of these, Table A2 summarizes the hypotheses tested, and records the sign and significance of the main coefficients. A large number of studies, including Bertrand (2013); Layard et al. (2014); Vendrik (2013); Clark and Senik (2010); Frijters et al. (2014); Gerritsen (2016) find that economic resources (e.g. household income or labour earnings) are associated with higher levels of reported wellbeing. Reported wellbeing systematically declines following major adverse life events, including physical violence (Johnston et al. 2018), exposure to the Chernobyl disaster (Danzer and Danzer 2016), widowhood (Odermatt and Stutzer 2019), or falling into poverty

¹⁷Most of these unprinted control variables are standard sociodemographic characteristics that researchers include in regressions, such as age, gender, race, religion, marital status, family size, employment status, job characteristics, income, health, and childhood characteristics.

¹⁸By contrast, the remaining papers in *WellBase* generally do not treat subjective wellbeing as a primary outcome of interest, but rather include it only peripherally, e.g. in robustness checks.

Table 1: Descriptive Statistics of *WellBase* at the estimate level

	Mean	SD	Min	Max
About the wellbeing scales:				
<i>Number of response categories:</i>				
3-points scale	0.00		0	1
4-points scale	0.27		0	1
5-points scale	0.14		0	1
6-points scale	0.00		0	1
7-points scale	0.00		0	1
10-points scale	0.22		0	1
11-points scale	0.36		0	1
More than 11-points scale	0.00		0	1
<i>Type of question:</i>				
Life Satisfaction	0.77		0	1
Cantril Ladder	0.05		0	1
Happiness Question	0.18		0	1
About the estimation samples:				
Number of observations	158,089.14	368,570.79	59	2,471,360
Number of observations (logged)	9.98	2.17	4.08	14.72
About the econometric models:				
Number of controls	34.08	30.02	1	191
Individual FE	0.14		0	1
About the independent variables:				
Printed in manuscript	0.06		0	1
Printed in appendix	0.12		0	1
Not printed	0.81		0	1
Continuous variable	0.25		0	1
Time-varying variable	0.75		0	1
Two-stage least square	0.01		0	1
Individual-specific	0.91		0	1
Natural experiment, RCT and policy reform	0.04		0	1
Macroeconomic indicator	0.04		0	1
Absolute t-statistics	5.11	13.76	0.00	474.74
Absolute t-statistics (logged)	0.45	1.53	-9.08	6.16
<hr/>				
Total number of estimates:	28,513			
Total number of regressions:	1,601			
Total number of papers:	72			

Note: These numbers refer to the sample of 28,513 estimates included in *WellBase*.

(Clark et al. 2016). Several papers analyse changes in policy or the shared environment, such as centralization reforms in Switzerland (Flèche 2021), income transparency reforms in Norway (Perez-Truglia 2020), or the London Olympic Games (Dolan et al. 2019).

4.2 Results on Wellbeing Scales

This section presents a series of results systematically assessing the sensitivity of estimates reproduced in *WellBase* to the assumption that scale use is linear. We do so with the help of the cost function C defined in Section 2.4. Recall that when $C = 0$, this corresponds to the near-universally adopted assumption that scale use is linear in underlying wellbeing. As C increases, the transformed scale increasingly departs from this assumption. When C may take on any value on the unit interval, the assumption of cardinality is replaced by a purely ordinal interpretation. As noted earlier, we refer to transformations with $0 \leq C \leq 0.15$ as “*plausible*” or “*mild*”, those with $0.15 < C \leq 0.30$ as “*conservatively plausible*” or “*marginal*”, and those with $0.30 < C \leq 1.00$ as “*implausible*” or “*unlikely*”. This is based on the evidence reported in Section 3. Three regions are shaded in the figures: a green one for “*plausible*” transformations, an orange one for “*conservatively plausible*” transformations, and a red one for “*implausible*” transformations.

4.2.1 On sign reversals: Documenting the risk of reversal

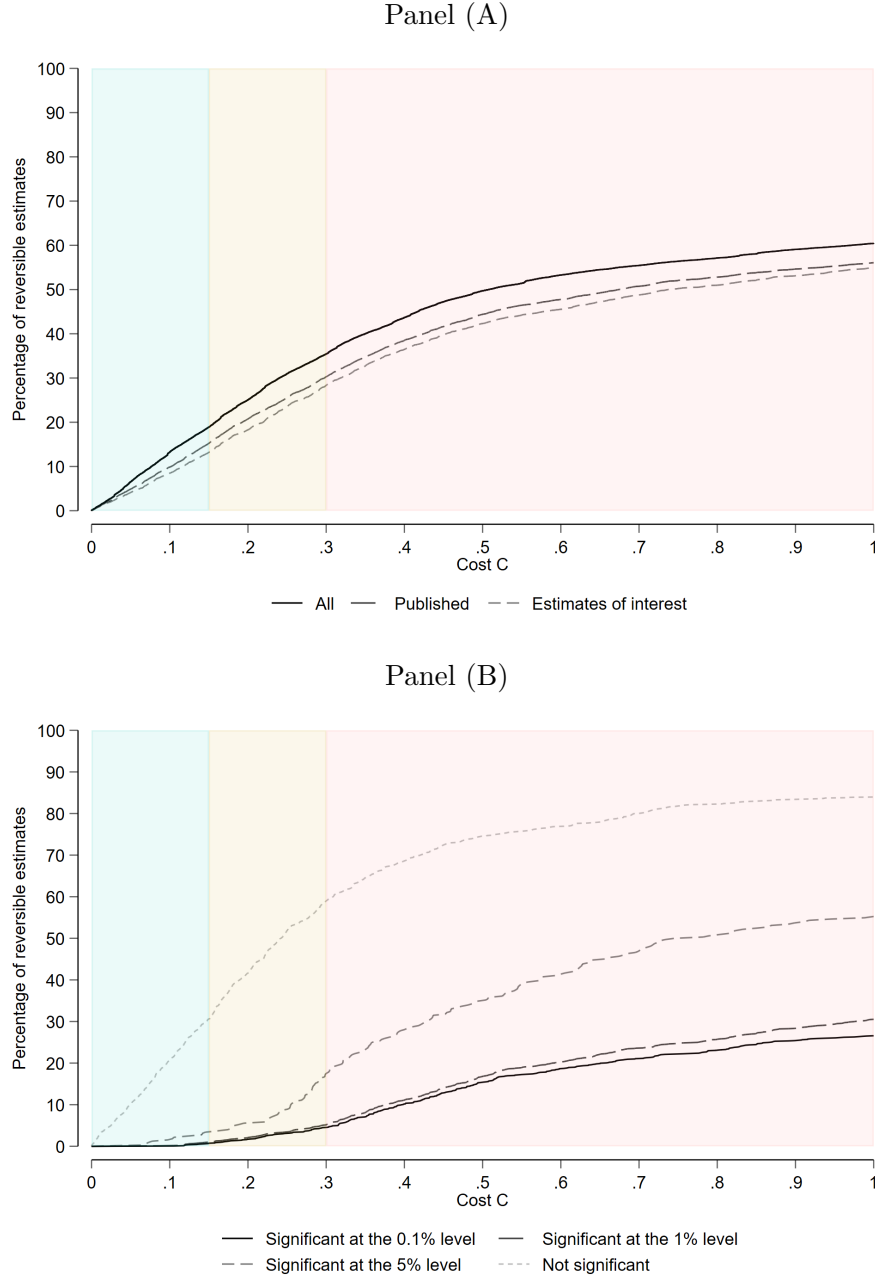
Figure 5 shows the share of point estimates whose sign can be reversed by applying some positive monotonic transformation of the response scale with a cost of at most C .

We report three lines in Panel (A). The solid dark line shows the share of sign reversals among all point estimates in *WellBase*. The remaining two lines present the same statistic for *printed estimates* and for *estimates of interest*. Here, an *estimate of interest* refers to estimates explicitly discussed in the text of the manuscript, and on which the conclusions of the included papers are based. The lines in Panel (A) all exhibit a concave relationship between the cost C and the percentage of sign reversals. About 60% of all replicated estimates can be sign-reversed via at least one positive monotonic transformation of the wellbeing scale when allowing for any cost C . However, focusing on “*plausible*” transformations only (i.e. $C < 0.15$), the risk of sign reversals drops to 18% of all point estimates in *WellBase*. *Printed estimates* and *estimates of interest* specifically exhibit even lower risks of sign reversal.¹⁹

Panel (B) focuses only on *estimates of interest* and displays a further breakdown by estimates’ original level of statistical significance. There is clear gradient between the original level of significance and the possibility of sign reversal: the less significant an estimate at $C = 0$, the greater the chance that there is at least one transformation changing its sign. Sign reversals are virtually non-existent under any “*plausible*” transformation among coefficients that meet the 5% significance threshold.

¹⁹It is here also worth re-emphasizing that not *all* transformations with a given C are sign-reversing; it only means that *at least one* such transformation can do so.

Figure 5: Cumulative sign-reversal shares for different values of C in WellBase



Notes: Cumulative shares of coefficients for which the sign can be reversed by at least one positive monotonic transformation of the response scale with at most cost C . When $C = 0$, this corresponds to the standard linearity assumption on scale use. When C may take on any value on the unit interval, shown on the far right of the graphs, any monotonic transformation of the original scale is permissible and the assumption of cardinality is thereby replaced by a purely ordinal interpretation. Based on the scale-use evidence presented in Section 3, shaded regions indicate “plausible” (green), “conservatively plausible” (orange), and “implausible” (red) degrees of non-linearity. Panel (A) confirms that the risk of reversal is an empirical threat: 60% of all replicated estimates can be sign-reversed by some positive monotonic transformation of the response scale. This risk drops to 18% when restricting attention to “likely” transformations. Panel (B) shows that it is much harder to reverse the sign of coefficients that are originally significant at the 5% level or below.

Finally, in Appendix Table A2 we restrict attention to studies whose main objective is to uncover the determinants of subjective wellbeing. The last two columns of this table indicate whether a reversal is possible and, if so, what minimal deviation from linearity (i.e., cost C) is required to produce such a reversal. About half of the coefficients reported in Appendix Table A2 are reversible. However, the risk is again much lower among the statistically significant coefficients (at the 5% level): about 33% of these can be sign-reversed, and in 95% of cases doing so would require a cost $C > 0.15$.

Overall, these results indicate that although sign reversals are often possible in principle, reversals under *plausible* (i.e $C < 0.15$) transformations are not. This is especially true for results that were highly statistically significant in their original form.

4.2.2 On sign reversals: Predicting the risk of reversal

We now investigate whether the risk of sign reversal can be predicted by observable features of the research design. To address this question, we estimate a linear probability model of the form:

$$\text{Rev}_{mpr} = \beta_0 + \beta_1 \ln(\# \text{Observations})_{pr} + \beta_2 \mathbf{Model}_{pr} + \beta_3 \mathbf{Estimate}_{mpr} + \beta_4 \mathbf{X}_{pr} + \varepsilon_{mpr}, \quad (6)$$

where the dependent variable, Rev_{mpr} , is a dummy equal to one if there exists at least one positive monotonic transformation of the wellbeing scale capable of reversing the sign of estimate m from regression r in paper p , and zero otherwise.²⁰

The term $\ln(\# \text{Observations})_{pr}$ gives the logged number of observations in each regression r in each paper p . The vector \mathbf{Model}_{pr} includes the logged number of control variables and a dummy for regressions that include individual fixed-effects. These features reflect standard practices through which researchers attempt to limit omitted variable bias, either by conditioning on observed covariates or accounting for unobserved time-invariant heterogeneity. The vector $\mathbf{Estimate}_{mpr}$ captures characteristics specific to the covariate m . It includes dummies for whether the covariate is continuous (as opposed to categorical or binary), time-varying, or instrumented via two-stage least squares. It also includes a categorical variable classifying whether the covariate corresponds to an individual socio-demographic characteristic (the reference category), a natural experiment (e.g., policy reform or randomized controlled trial), a placebo, or a macroeconomic indicator. Finally, the vector \mathbf{X}_{pr} comprises control variables: a dummy indicating whether the wellbeing scale includes at least seven

²⁰We also estimate a probit model to assess the robustness of our findings. Marginal effects are reported in Table A5. Conclusions are the same.

categories, and a categorical variable differentiating among life satisfaction questions, the Cantril Ladder, and happiness questions (see Appendix Table A1 where we report the wording of these questions). We estimate two versions of Equation (6): one without and one with the logged t-statistic. We treat the t-statistic differently because, unlike the other variables — which reflect researchers’ design choices — it is an outcome of those choices that is not directly controlled. We include it to test whether the observed negative association between statistical significance and reversal risk (Panel B, Figure 5) holds in a multivariate setting.

Conditional upon the possibility of a sign reversal for a given estimate, we further estimate the following via OLS:

$$\text{Cost}_{mpr} = \beta_0 + \beta_1 \ln(\# \text{Observations})_{pr} + \beta_2 \mathbf{Model}_{pr} + \beta_3 \mathbf{Estimate}_{mpr} + \beta_4 \mathbf{X}_{pr} + \varepsilon_{mpr}, \quad (7)$$

Equation (7) mirrors Equation (6) but uses the minimum cost C needed for a sign reversal as the dependent variable. Comparing Equations (6) and (7) enables us to assess whether the probability of reversal and the ease of achieving it share common determinants. In both types of regressions, we cluster standard errors at the regression–paper ($r \times p$) level. Continuous independent variables are standardized using the means and standard deviations reported in Table 1.

Table 2 reports predictors of reversal risk in Columns (1) and (2) and reversal cost in Columns (3) and (4). We highlight three main findings. First, the determinants of whether a reversal is possible and how costly it is are largely shared. Variables that lower the probability of reversal also increase the cost required to achieve a reversal. Second, the logged t-statistic is by far the strongest predictor of robustness: estimates with higher t-statistics are substantially less prone to reversal and more costly to reverse. This single variable alone explains much of the variation, raising the R^2 of the model from 17% to over 41% in Columns (1) and (2) and from 11% to over 51% in Columns (3) and (4). Last, a covariate’s source of variation matters: keeping the logged t-statistics constant, the sign of estimates exploiting arguably exogenous sources of variation (e.g., natural experiments or macroeconomic indicators) are both less likely to reverse and require larger departures from linearity.²¹

The risk and cost of sign reversal are not just random noise. They reflect identifiable features of research design, and are therefore within researchers’ control. Finally, signs of highly significant results are far more likely to persist across scale transformations.

²¹We conduct a series of robustness checks in Appendix Table A5. Specifically, we re-estimate Columns (2) and (4) of Table 2 while adding journal or paper fixed effects, and employing a probit model instead of a linear probability model. Our main conclusions remain robust across these specifications.

Table 2: Predictors of the Probability and Cost of Sign-reversal

	P(Sign-reversal)		Cost of sign-reversal	
	(1)	(2)	(3)	(4)
About the estimation sample:				
Number of observations (logged)	-0.105*** (0.007)	0.029*** (0.006)	0.016*** (0.003)	-0.038*** (0.003)
About the econometric model:				
Number of controls	0.019* (0.008)	0.001 (0.006)	-0.009** (0.003)	-0.006 (0.003)
Individual FE	0.084*** (0.021)	0.004 (0.016)	-0.045*** (0.009)	-0.010 (0.008)
About the independent variable:				
Continuous variable	-0.080*** (0.010)	-0.034*** (0.008)	-0.008 (0.006)	-0.002 (0.005)
Time-varying variable	-0.142*** (0.009)	-0.074*** (0.007)	0.065*** (0.005)	0.038*** (0.004)
Two-stage least square	-0.056 (0.033)	-0.013 (0.032)	0.048** (0.018)	0.016 (0.017)
Natural experiment	-0.167*** (0.017)	-0.090*** (0.014)	0.091*** (0.013)	0.034*** (0.009)
Macroeconomic indicator	0.145*** (0.015)	-0.026 (0.014)	-0.009 (0.009)	0.024** (0.008)
Absolute t-statistics (logged)		-0.275*** (0.004)		0.187*** (0.002)
Observations	28,513	28,513	17,240	17,240
R ²	0.168	0.411	0.105	0.512

Notes: The table shows the results from regressions assessing the risk and cost of sign reversal under positive monotonic transformations of the well-being scale. Specifically, Columns (1) and (2) report coefficients from an OLS model where the dependent variable equals one if at least one transformation reverses the sign of a coefficient m from a regression r reported in paper p . Conditional on a sign reversal being possible, Columns (3) and (4) report coefficients from an OLS model where the dependent variable is the minimum cost C required for reversal. All regressions control for a dummy indicating whether the well-being scale includes at least seven response categories and for the type of well-being measure (life satisfaction, Cantril Ladder, or happiness). Standard errors are clustered at the regression-paper level. Statistical significance is denoted as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.2.3 On significance reversals: Documenting the risk of reversal

We now quantify the risk of *significance* reversals.²² To this end, we first divide all *estimates of interest* in *WellBase* into two groups: those initially significant at the 5% level, and those not significant at this level. For all estimates, we compute the maximum and minimum attainable p-values under any monotonic transformation. We define significance reversals as instances where some transformation of the wellbeing scale cause the maximum attainable p-value for an originally significant estimate to exceed the 5% threshold, or conversely, where the minimum attainable p-value for an originally non-significant estimate drops below this threshold. Conditional upon the existence of a significance reversal, we then numerically search for the transformation that produces this reversal with the smallest deviation from linearity.

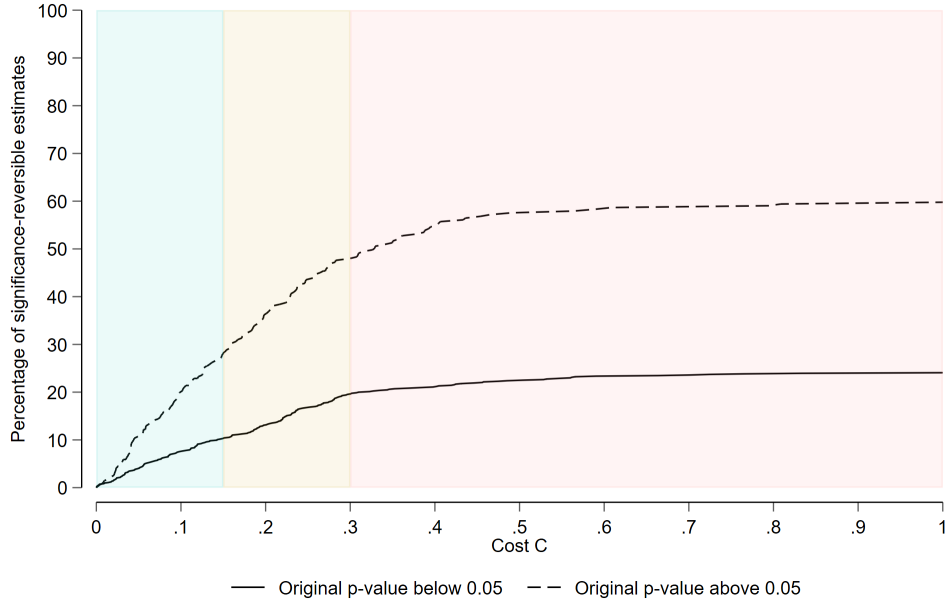
Figure 6 plots the share of significance reversal against the cost-of-reversal C . The solid black curve traces this share for coefficients originally significant at the 5% level. The dotted grey curve shows the corresponding share for insignificant coefficient crossing the significance threshold. The relationship between the cost C and the probability of significance reversals is, again, concave. The hazard of gaining significance is always greater than that of losing it: 60% of previously insignificant estimates can become significant with some positive monotonic transformation of the response scale. Only 24% of significant coefficients can be turned insignificant. Restricting attention to “plausible” transformations ($C < 0.15$) reduces these figures to 30% and 8%, respectively. Panel (B) restricts attention to initially significant coefficients. About 87% of coefficients with $0.01 < p \leq 0.05$ lose significance under some “mild” ($C < 0.15$) transformation. In contrast, coefficients that were already highly significant ($p < 0.001$) are almost immovable: 95% stay below a p-value of 0.05 under any positive monotonic transformation.

These results mirror those for sign reversals: significance reversals are a real concern, but their occurrence appears limited when restricting attention to “plausible” departures from linearity. This is especially true for highly significant estimates, which almost never become insignificant regardless of the transformation considered. However, as is intuitive, estimates just below the 5% threshold easily lose significance even under “mild” transformations.

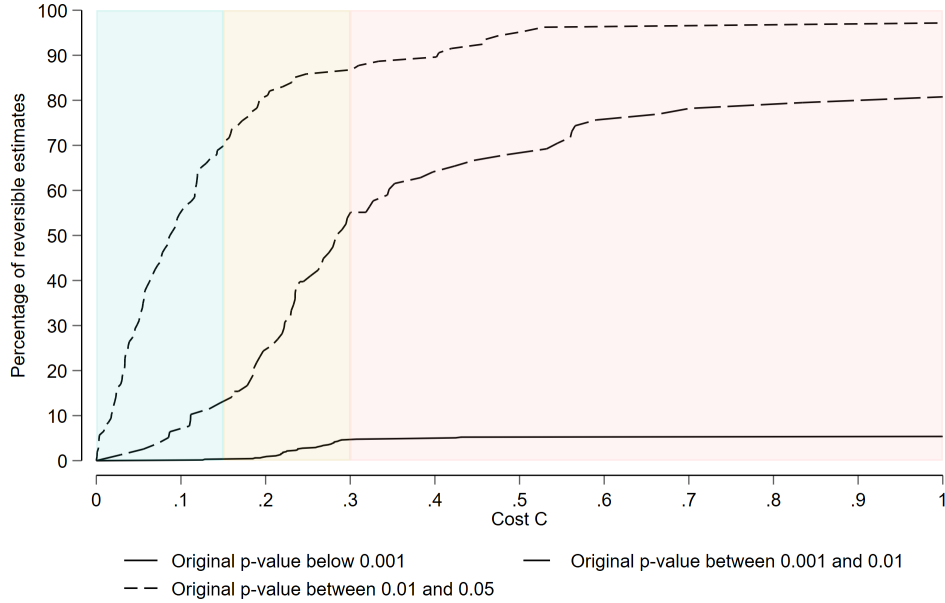
²²One might be tempted to view a coefficient that becomes (or ceases to be) statistically significant after a positive monotonic transformation of the wellbeing scale as Type II (false negative) or Type I (false positive) errors. This is not the case. Type I/II errors arise solely from sampling variability under a fixed coding of the data. The ‘significance reversals’ documented here, by contrast, are a deterministic consequence of re-labelling the ordered categories. Hence these risks only indicate how strongly inferences rely on the (implicit) cardinality assumptions built into the chosen coding of the scale.

Figure 6: Cumulative significance-reversal shares for different values of C in WellBase

Panel (A)



Panel (B)



Notes: Cumulative shares of coefficients for which *statistical significance* can be reversed by at least one positive monotonic transformation of the response scale with at most cost C . See notes of Figure 5 for more details about C and the shaded regions. Panel (A) shows that up to 24% of originally significant estimates may lose significance with at least one positive monotonic transformation, while approximately 60% of originally insignificant coefficients can be rendered significant. Panel(B) shows that originally highly significant coefficients ($p < 0.001$) are extremely robust, whereas marginally significant ones ($0.01 < p \leq 0.05$) are fragile even under “plausible” transformations.

4.2.4 On relative magnitudes: The case of unemployment and income

Turning to relative magnitudes, we focus on unemployment and income. These key determinants are studied across multiple papers in our database. The income–wellbeing relationship is especially central to policy-oriented work, because income is used as the numéraire in monetary valuations based on subjective wellbeing data (e.g. Dolan et al. 2019). Our analysis draws on the subset of nine studies in *WellBase* that simultaneously include both unemployment and household income in their regressions. To facilitate comparability across studies, we standardize each study’s wellbeing variable to mean zero and standard deviation one.

We first compute a paper-specific average point estimate for unemployment and for income weighted by the inverse of the standard error of the estimates.²³ The vertical markers in Figure 7 present such point estimates under the assumption of linearity (i.e. $C = 0$). Unemployment is indicated in blue. The red markers show income. On average, unemployment is associated with a decrease in wellbeing of roughly 0.39 standard deviations. A unit increase in log income is, on average, associated with a 0.18 SD increase in wellbeing.

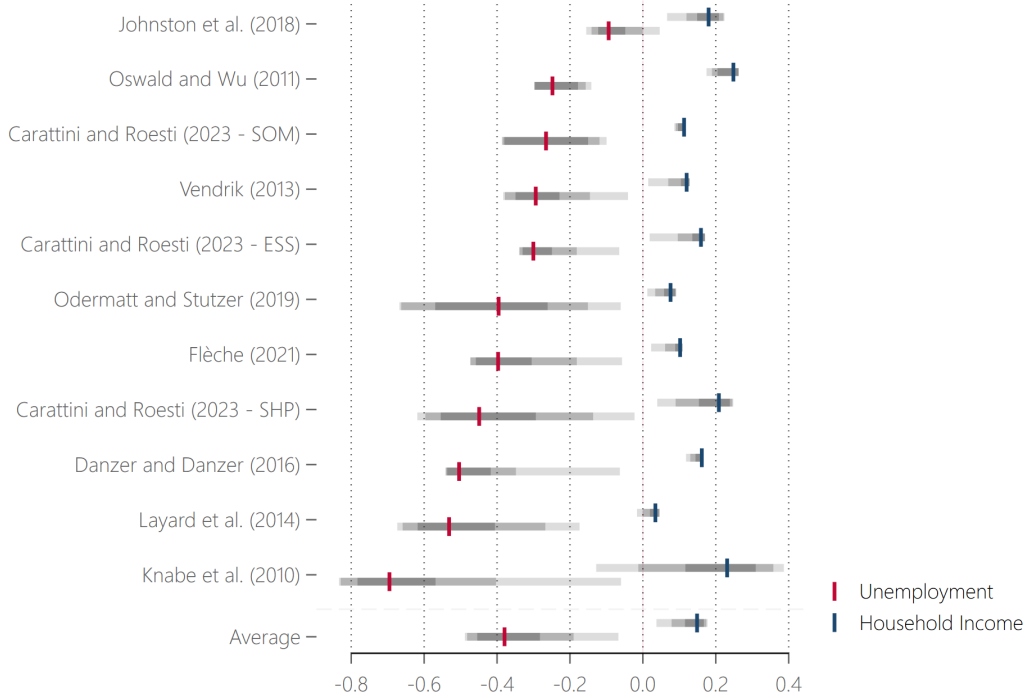
The grey bars in Figure 7 now show how these estimates may vary as we depart from linear scale use. The magnitude of estimates vary widely, even under “plausible” transformations. When taking the meta-analytic average across all studies, and for $C < 0.15$ (for $C < 0.30$), unemployment decreases wellbeing between 0.28 (0.19) and 0.45 (0.50) SDs. A unit increase in log income is correspondingly associated with an average increase between 0.11 (0.04) and 0.16 (0.18) SDs.

Given that there is no natural absolute scale for wellbeing (linear or not), the absolute magnitudes of coefficients are not meaningful. Ratios of coefficients, in contrast, do provide a meaningful relative measure. When interpreted as causal estimates, such ratios can be interpreted as marginal rates of substitution (MRS) between two variables. Figure 8 therefore shows ratios of the coefficient on unemployment to the coefficient on $\ln(\text{HH income})$ across different levels for C .²⁴ Each grey line in Panel (A) corresponds to a different paper. The black line shows the average ratio across all papers. We observe that this mean MRS can range from positive to as low as -18.25 . Under “likely” transformations ($C < 0.15$), the ranges are only slightly narrower, ranging from zero to -10 . Panel (B) disaggregates these MRS estimates by regression type. We do so because regressions with individual fixed effects produce income coefficients that are systematically closer to zero than cross-sectional regressions. Hence, we expect ratios using these coefficients as the numéraire to be more

²³The only exception is Carattini and Roesti (2023), who used three distinct datasets, where we treat each dataset from their paper as a unique observation.

²⁴For the computations to follow, we exclude regressions where the coefficient on income was reversible. According to Proposition 3, coefficients are unbounded in such a case. We had to exclude four regressions on that basis: one in Knabe et al. (2010) and three in Layard et al. (2014) (c.f. Figure 8).

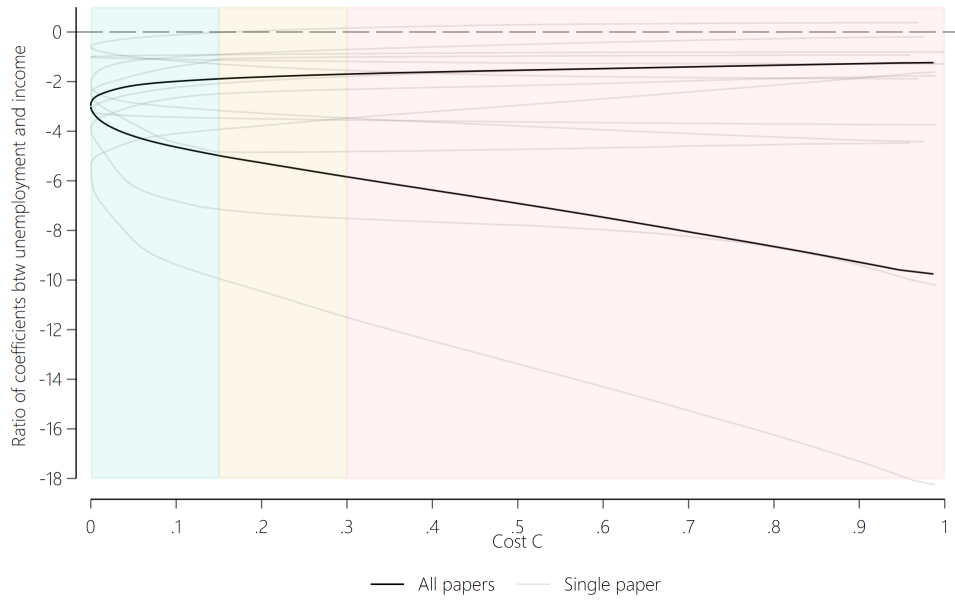
Figure 7: Forest plot showing the sensitivity of relative estimate magnitudes to transformations of the wellbeing scale



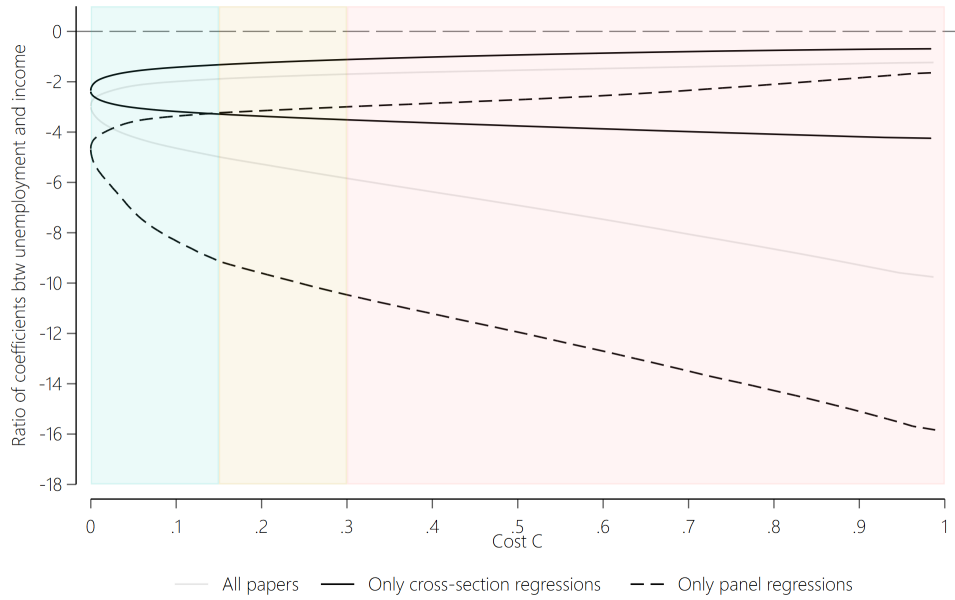
Notes: Standardized average point estimates for unemployment and the log of household income among papers included in *WellBase*. Papers are ranked by the average effect size of the unemployment coefficient. The overall average, weighted by the inverse of the standard error of the individual estimates and based on a meta-analytic fixed-effects model (Borenstein et al. 2010), is displayed at the bottom. Wellbeing scales are standardized (mean of zero and standard deviation of one), whereas estimates for unemployment and household income (expressed in logarithmic terms) are not standardized. Confidence intervals are replaced by grey bars, which indicate the range of point estimates after applying positive monotonic transformations of the wellbeing scales. There are three shades of grey, with the darkest corresponding to “plausible” transformations ($C < 0.15$), the middle to “conservatively plausible” transformations ($0.15 \leq C < 0.30$), and the lightest to “implausible” transformations ($0.30 \leq C$).

Figure 8: Ranges of unemployment-income ratios

Panel (A)



Panel (B)



Notes: Range of marginal rate of substitution (MRS) between unemployment and log household income. See notes of Figure 5 for more details about C and the shaded regions. Panel (A) plots the MRS between unemployment and log household income by paper (grey) and their average (black) across values of C . This reveals a wide range — from positive to -18.25 . Panel (B) disaggregates by regression type: MRS estimates from regressions with individual fixed effects are particularly unstable, while cross-sectional regressions are more robust.

sensitive. This is indeed what we observe: regressions that control for individual fixed effects are particularly sensitive, yielding MRS values that, even under “plausible” ($C < 0.15$) transformations of the wellbeing scale, span a wide range of (from -9.14 to -3.23). In contrast, cross-sectional regressions are more robust: the average MRS initially equals roughly -2.37 at linearity and broadens only moderately, ranging between -3.29 and -1.32 at the boundary of “plausible” transformations.

Thus, although the risk of sign and significance reversals appeared relatively small under “plausible” transformations of the wellbeing scale, the same cannot be said about the *magnitudes* of estimates. Applied to the case of unemployment and income — two key drivers of subjective wellbeing — both absolute and relative magnitudes vary widely, even under mild transformations.

4.3 Likert Scales for Attitudes, Preferences and Perceptions

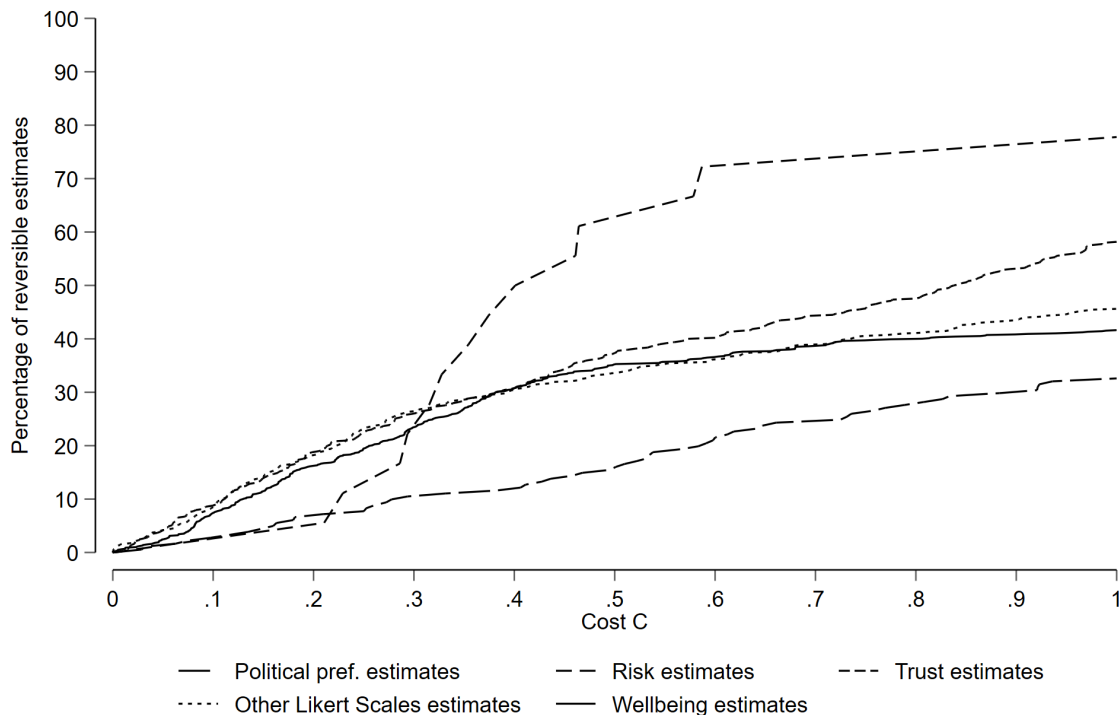
Although our focus has so far been on wellbeing scales, these are not the only constructs in economics measured using discrete and bounded response scales. Concepts such as risk aversion, trust, or political preferences are also routinely captured with such scales, and are broadly accepted within the discipline. To gauge whether concerns about the cardinal vs ordinal nature of Likert-style measurement ought to be unique to wellbeing, we now compare the reversal risks between these different types of measures.

To do so systematically, we screened every article that appeared between January 2010 and May 2025 in the five leading economics journals²⁵ and retained those whose full text contained the term “Likert scale” or whose title included at least one of the following expressions: “risk aversion,” “risk preferences,” “trust,” or “preferences for”. This search strategy is unlikely to cover all Likert-scale based research published in top 5 economics journals, but assembling a true census of all such published research is beyond the scope of this study.

As shown in figure A19, we reproduced 14 articles for a total of 411 regressions and 2,903 estimates (47.67% of which are printed coefficients). Of the included papers, three contained Likert-scale measure of trust (Acemoglu et al. 2020; Algan and Cahuc 2010; Falk et al. 2018), two contained a Likert-scale measure of political preferences (Kuziemko et al. 2015; Alesina et al. 2018), one contained a Likert-scale measure of risk aversion (Dohmen et al. 2010), and eight contained a Likert-scale measure of a other concepts including hiring interest, optimism, fear, political correctness attitudes, and work morale (Cohn et al. 2015; Kessler et al. 2019; Exley and Kessler 2022; Spenkuch et al. 2023; Braghieri 2024; Engelmann et al. 2024; Englmaier et al. 2024; Gagnon et al. 2025).

²⁵We count the following journals as part of the ‘top five’: *Quarterly Journal of Economics*, *American*

Figure 9: Comparing the risk of sign reversal between wellbeing scales and other Likert scales in Top-5 Economics journals



Notes: Cumulative shares of coefficients for which the sign can be reversed by at least one positive monotonic transformation of the response scale with at most cost C . Figure 9 shows that the risk of reversal is not unique to wellbeing. In many cases, results on constructs such as risk (77%), trust (58%), political preferences (32%), and other constructs (45%) are sign reversible.

Our results are shown in Figure 9. There we compare the sign reversal risk for estimates based on wellbeing scales published in top-five economics journals (solid line) with the corresponding risk for estimates derived from other types of Likert scales. The risk of sign reversal for wellbeing estimates in this subsample is around 41%. This is lower than that for risk and trust (77% and 58%, respectively), but larger than for political preferences (32%), and similar to that observed for other concepts measured with Likert scales (45%).

To explore why these risks vary across concepts, we replicated the analysis of Table 2. Figure A20 shows that the predictors of reversal risk are remarkably similar across types of measures: larger t-statistics reduce reversal risk, while more response categories increase it (see also Figure A4). Hence, the higher overall reversal risk for e.g. measures of risk-preferences reflects that these use scales with more categories and tend to have smaller t-

Economic Review, Journal of Political Economy, Review of Economic Studies, Econometrica.

statistics. In contrast, measures of trust or political preferences typically use fewer categories and exhibit larger t-statistics than wellbeing measures (see Appendix Table A6).

In sum, neither the level nor the determinants of reversal risk are unique to wellbeing. Any concept measured with Likert-type scales is similarly vulnerable.

5 Discussion

Economists increasingly rely on bounded survey scales to measure latent constructs like risk preferences, trust, political attitudes, and wellbeing. Standard practice treats these scales as cardinal measures, assuming without evidence that psychological distances between adjacent response categories remain constant across the entire scale. Our theoretical framework formalizes when this assumption matters and introduces a cost function to quantify the minimal deviation from linearity required to reverse the sign, to reverse significance, or to change the relative magnitude of estimated coefficients.

We gathered original experimental data to assess how individuals use response scales. Across a series of elicitation strategies, we find that respondents, on average, use such scales in a way that mildly deviates from linearity. Our estimates imply an upper bound on the cost of deviation from linearity at $C = 0.15$. We use this value as an empirical anchor for judging the plausibility of reversals.

We then ask to what extent wellbeing research published in top-ranked economics journals depends on the linearity assumption. To do so, we constructed *WellBase*, a database comprising the universe of replicable regressions using cognitive wellbeing as a dependent variable in the top 30 economics journals between January 2010 and May 2025. For each estimate, we assess whether its sign can be reversed by at least one positive monotonic transformation of the wellbeing scale and, if so, compute the minimal cost of such a transformation. Plausibility is defined based on the evidence we collected. We further examine whether research practices exist that are systematically associated with a lower risk of sign reversal. Finally, we use *WellBase* to document the rate of significance reversals and changes in coefficient ratios under positive monotonic transformations.

We find that the risk of sign reversal is concave in the cost of deviating from linearity. Plausible transformations of the wellbeing scale can reverse the sign of about 20% of the wellbeing research published in top-ranked economics journals. If linearity is entirely abandoned, this share increases to approximately 60%. The corresponding values for our sample of non-wellbeing Likert scales lies between 33% and 78%. Among wellbeing-based coefficients with p -values below 0.1 — the ones typically emphasized in published texts — the risk is negligible if we consider plausible transformations only. More generally, the risk

of sign reversal is not random: it can be predicted by observable features of the research design. One key finding is that estimates relying on arguably exogenous variation — such as natural experiments or macroeconomic shocks — are systematically less prone to reversals.

Regarding significance reversals, we again find a concave relationship: the marginal effect of relaxing linearity on the risk of significance reversal diminishes with cost. Among coefficients with an original p -value below 0.01, the risk drops sharply as the level of precision increases. If the linearity assumption were fully abandoned, roughly 86% of the estimates originally significant at the 1% level would remain robust at the 10% level. However, for estimates with p -values between 0.1 and 0.01, the risk of significance reversal escalates quickly — even under empirically plausible transformations of the wellbeing scale. Hence, the bar for statistical inference is higher than in the absence of concerns over non-linear scale use.

To assess the sensitivity of coefficient magnitudes and ratios, we restrict the analysis to papers that include both unemployment and income as covariates. Even small deviations from linearity substantially affect the absolute size of these coefficients and can easily alter their ratio — by an order of magnitude. Thus, while the direction of estimates tends to be stable, their relative sizes are highly sensitive to scale assumptions.

Some of our conclusions are nevertheless encouraging. The overall risk of sign reversal is limited under plausible deviations from linearity, and partially predictable based on research design. Likewise, the risk of significance reversal is small for estimates with high original precision. But other conclusions are more concerning. First, our results are not unique to wellbeing data: estimates based on other widely used Likert-type scales in economics — such as trust, risk preferences, or political attitudes — face similar risks. Potentially non-linear scale use is therefore a concern for a much broader segment of economic research than is widely recognised. Second, the risk of significance reversal is high for estimates with p -values between 0.1 and 0.01, even under modest departures from linearity. Finally, estimated magnitudes and coefficient ratios are highly unstable. Here, too, do minimal non-linearities in scale use suffice to reverse researchers’ substantive conclusions.

Our results do have some practical implications. It seems that researchers can keep current survey instruments largely unchanged: discrete and continuous response formats yield similar regressions coefficients, and explicit instructions on how respondents should use response scales appear to have negligible effects. What is needed, however, is a broader evidence-base on scale use. Our own tests, while indicative, are drawn from a single type of wellbeing scale and could not pin-down the precise functional form by which scale use departs from linearity. Until such evidence accumulates, we recommend that analysts routinely probe the robustness of their headline results to monotonic transformations of their outcome. One contribution of this paper, we hope, is to render such tests more tractable.

References

- Daron Acemoglu, Ali Cheema, Asim I Khwaja, and James A Robinson. Trust in state and nonstate actors: Evidence from dispute resolution in pakistan. *Journal of Political Economy*, 128(8):3090–3147, 2020.
- Achyuta Adhvaryu, Anant Nyshadham, and Huayu Xu. Hostel takeover: Living conditions, reference dependence, and the well-being of migrant workers. *Journal of Public Economics*, 226:104949, 2023.
- Philippe Aghion, Ufuk Akcigit, Angus Deaton, and Alexandra Roulet. Creative destruction and subjective well-being. *American Economic Review*, 106:3869–3897, 2016.
- Nicolás Ajzenman, Cevat Giray Aksoy, and Sergei Guriev. Exposure to transit migration: Public attitudes and entrepreneurship. *Journal of Development Economics*, 158:102899, 2022.
- Cevat Giray Aksoy and Semih Tumen. Local governance quality and the environmental cost of forced migration. *Journal of Development Economics*, 149:102603, 2021.
- Alberto Alesina, Stefanie Stantcheva, and Edoardo Teso. Intergenerational mobility and preferences for redistribution. *American Economic Review*, 108(2):521–554, 2018.
- Yann Algan and Pierre Cahuc. Inherited trust and growth. *American Economic Review*, 100(5):2060–2092, 2010.
- Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. The welfare effects of social media. *American economic review*, 110(3):629–676, 2020.
- Viola Angelini, Danilo Cavapozzi, Luca Corazzini, and Omar Paccagnella. Do danes and italians rate life satisfaction in the same way? using vignettes to correct for individual-specific scale biases. *Oxford Bulletin of Economics and Statistics*, 76(5):643–666, 2014.
- Manuela Angelucci and Daniel Bennett. The economic impact of depression treatment in india: Evidence from community-based provision of pharmacotherapy. *American economic review*, 114(1):169–198, 2024.
- Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist’s companion*. Princeton University Press, Princeton, NJ, 2009.
- Nava Ashraf, Erica Field, and Jean Lee. Household bargaining and excess fertility: an experimental study in zambia. *American Economic Review*, 104(7):2210–2237, 2014.

- Tijan L Bah, Catia Batista, Flore Gubert, and David McKenzie. Can information and alternatives to irregular migration reduce “backway” migration from the gambia? *Journal of Development Economics*, 165:103153, 2023.
- James Banks, Cormac o’Dea, and Zoë Oldfield. Cognitive function, numeracy and retirement saving trajectories. *The Economic Journal*, 120(548):F381–F410, 2010.
- William P. Banks and Mark J. Coleman. Two subjective scales of number. *Perception & Psychophysics*, 29(2):95–105, 1981. doi: 10.3758/BF03207272.
- William P. Banks and David K. Hill. The apparent magnitude of number scaled by random production. *Journal of Experimental Psychology*, 102(2):353, 1974.
- Daniel J. Benjamin, Kristen Cooper, Ori Heffetz, and Miles Kimball. From happiness data to economic conclusions. *Annual Review of Economics*, 16, 2023a. doi: 10.1146/annurev-economics-091322-034426.
- Daniel J Benjamin, Kristen Cooper, Ori Heffetz, Miles S Kimball, and Jiannan Zhou. Adjusting for scale-use heterogeneity in self-reported well-being. Technical report, National Bureau of Economic Research, 2023b.
- Marianne Bertrand. Career, family, and the well-being of college-educated women. *American Economic Review*, 103:244–250, 2013.
- Marianne Bertrand and Sendhil Mullainathan. Do people mean what they say? implications for subjective survey data. *American Economic Review*, 91(2):67–72, 2001.
- Pedro Bessone, Gautam Rao, Frank Schilbach, Heather Schofield, and Mattie Toma. The economic consequences of increasing sleep among the urban poor. *The Quarterly Journal of Economics*, 136(3):1887–1941, 2021.
- Carola Binder and Christos Makridis. Stuck in the seventies: gas prices and consumer sentiment. *Review of Economics and Statistics*, 104(2):293–305, 2022.
- Kjetil Bjorvatn, Denise Ferris, Selim Gulesci, Arne Nascowitz, Vincent Somville, and Lore Vandewalle. Childcare, labor supply, and business development: Experimental evidence from uganda. *American Economic Journal: Applied Economics*, 17(2):75–101, 2025.
- David Blakeslee, Ram Fishman, and Veena Srinivasan. Way down in the hole: Adaptation to long-term water loss in rural india. *American Economic Review*, 110(1):200–224, 2020.

- David G Blanchflower and Andrew J Oswald. Well-being over time in britain and the usa. *Journal of public economics*, 88(7-8):1359–1386, 2004.
- Christopher Blattman and Stefan Dercon. The impacts of industrial and entrepreneurial work on income and health: Experimental evidence from ethiopia. *American Economic Journal: Applied Economics*, 10(3):1–38, 2018.
- Jeffrey R. Bloem. How much does the cardinal treatment of ordinal variables matter? an empirical investigation. *Political Analysis*, 30(2):197–213, 2022. doi: 10.1017/pan.2020.55.
- Jeffrey R. Bloem and Andrew J. Oswald. The analysis of human feelings: a practical suggestion for a robustness test. *Review of Income and Wealth*, 68(3):689–710, 2022. doi: 10.1111/roiw.12542.
- Nicholas Bloom, James Liang, John Roberts, and Zhichun Jenny Ying. Does working from home work? evidence from a chinese experiment. *The Quarterly journal of economics*, 130(1):165–218, 2015.
- Joshua Blumenstock, Michael Callen, and Tarek Ghani. Why do defaults affect behavior? experimental evidence from afghanistan. *American Economic Review*, 108(10):2868–2901, 2018.
- Timothy N Bond and Kevin Lang. The sad truth about happiness scales. *Journal of Political Economy*, 127:1629–1640, 2019.
- Michael Borenstein, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2):97–111, 2010.
- Luca Braghieri. Political correctness, social image, and information transmission. *American Economic Review*, 114(12):3877–3904, 2024.
- Gharad Bryan, James J Choi, and Dean Karlan. Randomizing religion: the impact of protestant evangelism on economic outcomes. *The Quarterly Journal of Economics*, 136(1):293–380, 2021.
- Filipe Campante and David Yanagizawa-Drott. Does religion affect economic growth and happiness? Evidence from Ramadan. *The Quarterly Journal of Economics*, 130:615–658, 2015.
- Stefano Carattini and Moritz Roesti. Trust, happiness, and pro-social behavior. *Review of Economics and Statistics*, pages 1–45, 2023.

- Stefano Caria, Simon Franklin, and Marc Witte. Searching with friends. *Journal of Labor Economics*, 41(4):887–922, 2023.
- Le-Yu Chen, Elizaveta Oparina, Nattavudh Powdthavee, and Sorawoot Srisuma. Robust ranking of happiness outcomes: A median regression perspective. *Journal of Economic Behavior & Organization*, 200:672–686, 2022.
- Yi Chen and Hanming Fang. The long-term consequences of china’s “later, longer, fewer” campaign in old age. *Journal of Development Economics*, 151:102664, 2021.
- Terence C Cheng, Nattavudh Powdthavee, and Andrew J Oswald. Longitudinal evidence for a midlife nadir in human well-being: Results from four data sets. *The Economic Journal*, 127:126–142, 2017.
- Alberto Ciancio, Adeline Delavande, Hans-Peter Kohler, and Iliana V Kohler. Mortality risk information, survival expectations and sexual behaviours. *The Economic Journal*, 134(660):1431–1464, 2024.
- Andrew E Clark and Andrew J Oswald. Unhappiness and unemployment. *The economic journal*, 104(424):648–659, 1994.
- Andrew E Clark and Claudia Senik. Who compares to whom? The anatomy of income comparisons in Europe. *The Economic Journal*, 120:573–594, 2010.
- Andrew E Clark and Rong Zhu. Taking back control? quasi-experimental evidence on the impact of retirement on locus of control. *The Economic Journal*, 134(660):1465–1493, 2024.
- Andrew E Clark, Conchita D’Ambrosio, and Simone Ghislandi. Adaptation to poverty in long-run panel data. *Review of Economics and Statistics*, 98:591–600, 2016.
- Alain Cohn, Jan Engelmann, Ernst Fehr, and Michel André Maréchal. Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, 105(2):860–885, 2015.
- Charles Courtemanche, David Frisvold, David Jimenez-Gomez, Mariétou H Ouayogodé, and Michael K Price. Chain restaurant calorie posting laws, obesity, and consumer welfare. *Journal of the European Economic Association*, page jvaf004, 2025.
- Aidan Coville, Sebastian Galiani, Paul Gertler, and Susumu Yoshida. Financing municipal water and sanitation services in nairobi’s informal settlements. *Review of Economics and Statistics*, pages 1–48, 2023.

- Gordon B Dahl, Christina Felfe, Paul Frijters, and Helmut Rainer. Caught between cultures: Unintended consequences of improving opportunity for immigrant girls. *The Review of Economic Studies*, 89:2491–2528, 2022.
- Patricio S Dalton, Julius Rüschepöhler, Burak Uras, and Bilal Zia. Curating local knowledge: Experimental evidence from small retailers in indonesia. *Journal of the European Economic Association*, 19(5):2622–2657, 2021.
- Alexander M Danzer and Natalia Danzer. The long-run consequences of Chernobyl: Evidence on subjective well-being, mental health and welfare. *Journal of Public Economics*, 135: 47–60, 2016.
- Jan-Emmanuel De Neve, George Ward, Femke De Keulenaer, Bert Van Landeghem, Georgios Kavetsos, and Michael I Norton. The asymmetric experience of positive and negative economic growth: Global evidence using subjective well-being data. *Review of Economics and Statistics*, 100:362–375, 2018.
- Taryn Dinkelman and Sam Schulhofer-Wohl. Migration, congestion externalities, and the evaluation of spatial investments. *Journal of Development Economics*, 114:189–202, 2015.
- Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sunde. Are risk aversion and impatience related to cognitive ability? *American economic review*, 100(3):1238–1260, 2010.
- Paul Dolan, Georgios Kavetsos, Christian Krekel, Dimitris Mavridis, Robert Metcalfe, Claudia Senik, Stefan Szymanski, and Nicolas R Ziebarth. Quantifying the intangible impact of the Olympics using subjective well-being data. *Journal of Public Economics*, 177:104043, 2019.
- Richard A. Easterlin. Does economic growth improve the human lot? some empirical evidence. In Paul A. David and Melvin W. Reder, editors, *Nations and households in economic growth*, pages 89–125. Academic Press, New York, 1974.
- Eric Edmonds, Ben Feigenberg, and Jessica Leight. Advancing the agency of adolescent girls. *Review of Economics and Statistics*, 105(4):852–866, 2023.
- Jan B Engelmann, Maël Lebreton, Nahuel A Salem-Garcia, Peter Schwardmann, and Joël J van der Weele. Anticipatory anxiety and wishful thinking. *American Economic Review*, 114(4):926–960, 2024.

- Florian Englmaier, Stefan Grimm, Dominik Grothe, David Schindler, and Simeon Schudy. The effect of incentives in nonroutine analytical team tasks. *Journal of Political Economy*, 132(8):2695–2747, 2024.
- Christine L Exley and Judd B Kessler. The gender gap in self-promotion. *The Quarterly Journal of Economics*, 137(3):1345–1381, 2022.
- Mark Fabian. Scale norming undermines the use of life satisfaction scale data for welfare analysis. *Journal of Happiness Studies*, 23(4):1509–1541, 2022.
- Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. Global evidence on economic preferences. *The quarterly journal of economics*, 133(4):1645–1692, 2018.
- Ada Ferrer-i Carbonell and Paul Frijters. How important is methodology for the estimates of the determinants of happiness? *The economic journal*, 114(497):641–659, 2004.
- Hayley Fisher and Anna Zhu. The effect of changing financial incentives on repartnering. *The Economic Journal*, 129(623):2833–2866, 2019.
- Sarah Flèche. The welfare consequences of centralization: Evidence from a quasi-natural experiment in Switzerland. *Review of Economics and Statistics*, 103:621–635, 2021.
- Paul Frijters and Christian Krekel. *A handbook for wellbeing policy-making: History, theory, measurement, implementation, and examples*. Oxford University Press, 2021.
- Paul Frijters, David W Johnston, and Michael A Shields. Does childhood predict adult life satisfaction? Evidence from British cohort surveys. *The Economic Journal*, 124:F688–F719, 2014.
- Nickolas Gagnon, Kristof Bosmans, and Arno Riedl. The effect of gender discrimination on labor supply. *Journal of Political Economy*, 133(3):000–000, 2025.
- Jules Gazeaud, Nausheen Khan, Eric Mvukiyehe, and Olivier Sterck. With or without him? experimental evidence on cash grants and gender-sensitive trainings in tunisia. *Journal of Development Economics*, 165:103169, 2023.
- Aart Gerritsen. Optimal taxation when people do not maximize well-being. *Journal of Public Economics*, 144:122–139, 2016.

- Hélène Giacobino, Elise Huillery, Bastien Michel, and Mathilde Sage. Schoolgirls, not brides: Education as a shield against child marriage. *American Economic Journal: Applied Economics*, 16(4):109–143, 2024.
- Edward L Glaeser, Joshua D Gottlieb, and Oren Ziv. Unhappy cities. *Journal of labor economics*, 34:S129–S182, 2016.
- Michael Grimm, Sidiki Soubeiga, and Michael Weber. Supporting small firms in a fragile context: Comparing matching and cash grants in burkina faso. *Journal of Development Economics*, 171:103344, 2024.
- Sergei Guriev and Daniel Treisman. Informational autocrats. *Journal of economic perspectives*, 33(4):100–127, 2019.
- Johannes Haushofer and Jeremy Shapiro. The short-term impact of unconditional cash transfers to the poor: Experimental evidence from Kenya. *The Quarterly Journal of Economics*, 131:1973–2042, 2016.
- Johannes Haushofer, Matthieu Chemin, Channing Jang, and Justin Abraham. Economic and psychological effects of health insurance and cash transfers: Evidence from a randomized experiment in kenya. *Journal of Development Economics*, 144:102416, 2020.
- Ori Heffetz and Daniel B Reeves. Difficulty of reaching respondents and nonresponse bias: Evidence from large government surveys. *Review of Economics and Statistics*, 101(1): 176–191, 2019.
- Wei Huang, Xiaoyan Lei, and Ang Sun. Fertility restrictions and life cycle outcomes: Evidence from the one-child policy in china. *Review of Economics and Statistics*, 103(4): 694–710, 2021.
- David W Johnston, Michael A Shields, and Agne Suziedelyte. Victimization, well-being and compensation: Using panel data to estimate the costs of violent crime. *The Economic Journal*, 128:1545–1569, 2018.
- Jan Kabátek and David C Ribar. Daughters and divorce. *The Economic Journal*, 131(637): 2144–2170, 2021.
- Caspar Kaiser. Using memories to assess the intrapersonal comparability of wellbeing reports. *Journal of Economic Behavior & Organization*, 193:410–442, 2022.
- Caspar Kaiser and Andrew J Oswald. The scientific value of numerical measures of human feelings. *Proceedings of the National Academy of Sciences*, 119(42):e2210412119, 2022.

- Caspar Kaiser and Alberto Prati. From likert changes to changing likert: rethinking satisfaction measurement over time. *Unpublished Working Paper*, 2025.
- Caspar Kaiser and Maarten Vendrik. How much can we learn from happiness data? University of Oxford, Mimeo, 2023.
- Judd B Kessler, Corinne Low, and Colin D Sullivan. Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review*, 109(11):3713–3744, 2019.
- Iris Kesternich, Bettina Siflinger, James P Smith, and Joachim K Winter. The effects of world war ii on economic and health outcomes across europe. *Review of Economics and Statistics*, 96(1):103–118, 2014.
- Andreas Knabe, Steffen Rätzl, Ronnie Schöb, and Joachim Weimann. Dissatisfied with life but having a good day: Time-use and well-being of the unemployed. *The Economic Journal*, 120:867–889, 2010.
- Christian Krekel, Ganga Shreedhar, Helen Lee, Claire Marshall, Alison Boler, Allison Smith, and Paul Dolan. Happy to help: welfare effects of a nationwide volunteering programme. *Review of Economics and Statistics*, pages 1–64, 2024.
- Ilyana Kuziemko, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva. How elastic are preferences for redistribution? evidence from randomized survey experiments. *American Economic Review*, 105(4):1478–1508, 2015.
- Richard Layard, Andrew E Clark, Francesca Cornaglia, Nattavudh Powdthavee, and James Vernoit. What predicts a successful life? A life-course model of well-being. *The Economic Journal*, 124:F720–F738, 2014.
- Kenneth Lee, Edward Miguel, and Catherine Wolfram. Experimental evidence on the economics of rural electrification. *Journal of Political Economy*, 128(4):1523–1565, 2020.
- Steven D Levitt. Heads or tails: The impact of a coin toss on major life decisions and subsequent happiness. *The Review of Economic Studies*, 88:378–405, 2021.
- Wei Li. The ”miseries” of sex imbalance: Evidence using subjective well-being data. *Journal of Development Economics*, 151:102634, 2021.
- Armando N Meier. Emotions and risk attitudes. *American Economic Journal: Applied Economics*, 14(3):527–558, 2022.

- Reto Odermatt and Alois Stutzer. (mis-) predicted subjective well-being following life events. *Journal of the European Economic Association*, 17(1):245–283, 2019.
- Andrew J Oswald. Happiness and economic performance. *The economic journal*, 107(445): 1815–1831, 1997.
- Andrew J Oswald. On the curvature of the reporting function from objective reality to subjective feelings. *Economics Letters*, 100:369–372, 2008.
- Andrew J Oswald and Stephen Wu. Well-being across America. *Review of Economics and Statistics*, 93:1118–1134, 2011.
- Andrew J Oswald, Eugenio Proto, and Daniel Sgroi. Happiness and productivity. *Journal of labor economics*, 33(4):789–822, 2015.
- Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi: 10.1136/bmj.n71.
- Ricardo Perez-Truglia. The effects of income transparency on well-being: Evidence from a natural experiment. *American Economic Review*, 110:1019–1054, 2020.
- Alberto Prati and Claudia Senik. Is it possible to raise national happiness? CEP Discussion Paper 2068, Centre for Economic Performance, London School of Economics and Political Science, 2025. URL <https://cep.lse.ac.uk/pubs/download/dp2068.pdf>.
- Jan Priebe, Ute Rink, and Henry Stemmler. Disability and risk preferences: Experimental and survey evidence from vietnam. *The Economic Journal*, 134(664):3390–3427, 2024.
- Emma Riley. Resisting social pressure in the household using mobile money: Experimental evidence on microenterprise investment in uganda. *American Economic Review*, 114(5): 1415–1447, 2024.
- Luis Sarmiento, Nicolas Wagner, and Aleksandar Zaklan. The air quality and well-being effects of low emission zones. *Journal of Public Economics*, 227:105014, 2023.
- Bruce Schneider, Scott Parker, Dan Ostrosky, David Stein, and Gary Kanow. A scale for the psychological magnitude of number. *Perception & Psychophysics*, 16(1):43–46, 1974. doi: 10.3758/BF03203247.

- Carsten Schröder and Shlomo Yitzhaki. Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, 92:337–358, 2017.
- Wenbiao Sha. The political impacts of land expropriation in china. *Journal of Development Economics*, 160:102985, 2023.
- Prakarsh Singh and William A Masters. Performance bonuses in the public sector: Winner-take-all prizes versus proportional payments to reduce child malnutrition in india. *Journal of Development Economics*, 146:102295, 2020.
- Jörg L Spenkuch, Edoardo Teso, and Guo Xu. Ideology and performance in public organizations. *Econometrica*, 91(4):1171–1203, 2023.
- Victor Stango and Jonathan Zinman. We are all behavioural, more, or less: A taxonomy of consumer decision-making. *The Review of Economic Studies*, 90(3):1470–1498, 2023.
- Raphael Studer. Does it matter how happiness is measured? evidence from a randomized controlled experiment. *Journal of Economic and Social Measurement*, 37(4):317–336, 2012. doi: 10.3233/JEM-120364.
- Ana Tur-Prats. Family types and intimate partner violence: A historical perspective. *Review of Economics and Statistics*, 101(5):878–891, 2019.
- UK HMRC Treasury. Wellbeing guidance for appraisal: Supplementary green book guidance. Report, HM Treasury, 2021. URL <https://www.gov.uk/government/publications/green-book-supplementary-guidance-wellbeing>.
- Bernard Van Praag. The welfare function of income in Belgium: An empirical investigation. *European Economic Review*, 2(3):337–369, 1971. doi: 10.1016/0014-2921(71)90045-6.
- Bernard MS Van Praag. Ordinal and cardinal utility: an integration of the two dimensions of the welfare concept. *Journal of econometrics*, 50(1-2):69–89, 1991.
- Bernard MS Van Praag and Nico L Van der Sar. Household cost functions and equivalence scales. *Journal of human Resources*, pages 193–210, 1988.
- Bernard MS Van Praag, Paul Frijters, et al. The measurement of welfare and well-being: The leyden approach. *Well-being: The foundations of hedonic psychology*, pages 413–433, 1999.
- Maarten C Vendrik. Adaptation, anticipation and social interaction in happiness: An integrated error-correction approach. *Journal of Public Economics*, 105:131–149, 2013.

Michael Vlassopoulos, Abu Siddique, Tabassum Rahman, Debayan Pakrashi, Asad Islam, and Firoz Ahmed. Improving women’s mental health during a pandemic. *American Economic Journal: Applied Economics*, 16(2):422–455, 2024.

Appendix

A Proofs

A.1 Proof of Proposition 1

A.1.1 OLS case

A version of this proof originally appeared in Kaiser and Vendrik (2023). We here reproduce a shorter version in our notation, which will be useful for later proofs.

As in section 2.4, let l_k be the real value we assign to the k^{th} response category of the untransformed variable r_i , and let the labels assigned to each category of the transformed variable \tilde{r}_i be given by \tilde{l}_k . Hence, for any transformation f , we have $f(r_i = l_k) = \tilde{l}_k$. Now note that:

$$\begin{aligned}\tilde{r}_i &= \sum_{k=1}^K \tilde{l}_k \mathbf{1}(r_i = k) \\ &= \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{1}(r_i \leq k) + \tilde{l}_K \\ &= \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) d_{k,i} + \tilde{l}_K\end{aligned}$$

Stacking over individuals, we can thus write $\tilde{\mathbf{r}} = \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{d}_k + \tilde{l}_K \mathbf{I}$. Now notice that:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \tilde{\mathbf{r}} \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \left(\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{d}_k + \tilde{l}_K \mathbf{I} \right) \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \left(\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) ((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}')^{-1} \hat{\boldsymbol{\beta}}_k^{(d)} + \tilde{l}_K \mathbf{I} \right) \\ &= \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\boldsymbol{\beta}}_k^{(d)} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \tilde{l}_K \mathbf{I}\end{aligned}$$

Recall that the first element of $\hat{\boldsymbol{\beta}}$ records a constant. The second term in the last line is therefore a vector with all but the first element equal to zero. Hence, for coefficient $\hat{\beta}_m$

associated with covariate X_{im} , we can write $\hat{\beta}_m = \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\beta}_{km}^{(d)}$. Since $\tilde{l}_k - \tilde{l}_{k+1} < 0$ for all positive monotonic transformations, if $\text{sgn}(\hat{\beta}_{km}^{(d)})$ is constant across k , every positive monotonic transformation of r_i yields the same sign for $\hat{\beta}_m$. However, if $\text{sgn}(\hat{\beta}_{km}^{(d)}) \neq \text{sgn}(\hat{\beta}_{k'm}^{(d)})$ for at least one k and k' , then there will always be a choice of labels such that either $\tilde{l}_k - \tilde{l}_{k+1}$ or $\tilde{l}_{k'} - \tilde{l}_{k'+1}$ is sufficiently large to switch the sign of $\hat{\beta}_m$ (since either can be made arbitrarily large without affecting the other).

A.1.2 Fixed-effects case

Suppose we have panel data for respondents i and time period t . We collect the within-person means across $t = 1, 2, \dots, T_i$ of all covariates in $\bar{\mathbf{X}}$. The within-person means of $\tilde{\mathbf{r}}$ and \mathbf{d}_k are collected in $\bar{\tilde{\mathbf{r}}}$ and $\bar{\mathbf{d}}_k$, respectively. The demeaned values of \mathbf{X} , $\tilde{\mathbf{r}}$, and \mathbf{d}_k are then given by $\dot{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{X}}$, $\dot{\tilde{\mathbf{r}}} = \tilde{\mathbf{r}} - \bar{\tilde{\mathbf{r}}}$, and $\dot{\mathbf{d}}_k = \mathbf{d}_k - \bar{\mathbf{d}}_k$, respectively. The fixed effects estimator can then be written as $\hat{\beta}_{FE} = (\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} \dot{\mathbf{X}}' \dot{\tilde{\mathbf{r}}}$ and the result of Proposition 1 follows by the same argument.

A.1.3 2-SLS Case

To also cover the IV case, it is sufficient to show that all but the first element of $\hat{\beta}_{IV}$ are equal to $\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\beta}_{IV,k}^{(d)}$, where $\hat{\beta}_{IV}$ and $\hat{\beta}_{IV,k}^{(d)}$ are, respectively, IV estimates of regressions of $\tilde{\mathbf{r}}$ and \mathbf{d}_k on \mathbf{X} with excluded instruments \mathbf{Z} . In the just-identified case, $\hat{\beta}_{IV} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \tilde{\mathbf{r}}$ and $\hat{\beta}_{IV,k}^{(d)} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{d}_k$. Thus, analogous to the OLS case, we have:

$$\begin{aligned} \hat{\beta}_{IV} &= (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \tilde{\mathbf{r}} \\ &= (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \left(\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{d}_k + \tilde{l}_K \mathbf{I} \right) \\ &= (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \left(\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) ((\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}')^{-1} \hat{\beta}_{IV,k}^{(d)} + \tilde{l}_K \mathbf{I} \right) \\ &= \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\beta}_{IV,k}^{(d)} + (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \tilde{l}_K \mathbf{I} \end{aligned}$$

As in the OLS case, the term $(\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \tilde{l}_K \mathbf{I}$ is just an IV estimate of a regression of the constant term $\tilde{l}_K \mathbf{I}$. All but the first element will therefore be zero. Hence, as required, all but the first element of $\hat{\beta}_{IV}$ are equal to $\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\beta}_{IV,k}^{(d)}$.

A.1.4 Continuous Case

In principle r_i could be measured on a continuous scale. An analogous result to Proposition 1 holds in this case.

Proposition A1 (Non-reversal condition with continuous outcomes). *Let r_i be a continuous variable with support $[r_{\min}, r_{\max}]$ and let $f : [r_{\min}, r_{\max}] \rightarrow \mathbb{R}$ be any continuously differentiable, strictly increasing function (i.e., $f'(t) > 0$ for all $t \in [r_{\min}, r_{\max}]$). Define the transformed variable $\tilde{r}_i = f(r_i)$. Then, the sign of the OLS coefficient $\hat{\beta}_m$ on covariate X_{im} in the regression*

$$\tilde{r}_i = \mathbf{X}_i \hat{\boldsymbol{\beta}} + \epsilon_i,$$

is invariant under all such transformations f if and only if the coefficient $\hat{\beta}_m^d(t)$ obtained from the regression of the dichotomized variable $\mathbb{1}\{r_i \leq t\}$ on \mathbf{X}_i is of the same sign for every $t \in [r_{\min}, r_{\max}]$.

Since f is continuously differentiable and strictly increasing, we can write

$$f(r_i) = f(r_{\max}) - \int_{r_i}^{r_{\max}} f'(t) dt.$$

Noting that for any $r_i \in [r_{\min}, r_{\max}]$ we have

$$\int_{r_i}^{r_{\max}} f'(t) dt = \int_{r_{\min}}^{r_{\max}} f'(t) \mathbb{1}\{r_i \leq t\} dt,$$

it follows that

$$f(r_i) = f(r_{\max}) - \int_{r_{\min}}^{r_{\max}} f'(t) \mathbb{1}\{r_i \leq t\} dt.$$

Stacking observations and regressing \tilde{r}_i on \mathbf{X}_i yields

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{r}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(f(r_{\max})\mathbf{1} - \int_{r_{\min}}^{r_{\max}} f'(t) \mathbf{1}\{r_i \leq t\} dt\right).$$

Since $f(r_{\max})$ is constant, it affects only the intercept. For the coefficient on X_{im} , we obtain, analogous to the discrete case:

$$\hat{\beta}_m = - \int_{r_{\min}}^{r_{\max}} f'(t) \hat{\beta}_m^d(t) dt,$$

where $\hat{\beta}_m^d(t)$ is the coefficient on X_{im} from the regression of $\mathbb{1}\{r_i \leq t\}$ on \mathbf{X}_i . Because $f'(t) > 0$ for all t , the overall coefficient $\hat{\beta}_m$ is essentially a weighted average (with a negative sign) of the $\hat{\beta}_m^d(t)$. Thus, if $\hat{\beta}_m^d(t)$ has the same sign for every $t \in [r_{\min}, r_{\max}]$, then the sign

of $\hat{\beta}_m$ is fixed regardless of the choice of f . Conversely, if there is any interval of values for t where $\hat{\beta}_m^d(t)$ takes a different sign, one may choose f so that the weights $f'(t)$ shift the overall sign of $\hat{\beta}_m$.

A.2 Proof of Proposition 2

From Equation 1 and Assumptions 1 and 2 we have:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{r}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{s} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\zeta}$$

Given Assumption 1, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{s}$ is a consistent estimator of β . Given Assumption 2, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\zeta}$ is a consistent estimator of γ . Assumption 2 implies that $\text{sgn}(\beta_m) = \text{sgn}(\beta_m - \gamma_m)$. Thus, since $\text{sgn}(\hat{\beta}_m)$ is a consistent estimator of $\text{sgn}(\beta_m - \gamma_m)$, $\text{sgn}(\hat{\beta}_m)$ is also a consistent estimator of $\text{sgn}(\beta_m)$. By satisfying the non-reversal condition, $\text{sgn}(\hat{\beta}_m)$ is invariant under all positive monotonic transformations. Thus, $\text{sgn}(\hat{\beta}_m)$ is a consistent estimator of $\text{sgn}(\beta_m)$ for all positive monotonic transformations of r_i .

A.3 Standard errors under monotonic transformations

We here provide additional details for computing the variance-covariance matrix of estimated coefficients under arbitrary monotonic transformations of the response scale.

A.3.1 Residual decomposition

For any monotonic transformation $\tilde{r}_i = f(r_i)$, we show that the residuals from a regression of $\tilde{\mathbf{r}}$ on \mathbf{X} can be expressed as a weighted combination of residuals from regressions of the dichotomized variables \mathbf{d}_k . The residuals from the transformed regression are:

$$\tilde{\mathbf{e}} = \tilde{\mathbf{r}} - \mathbf{X}\hat{\beta}^{(\tilde{r})}. \quad (8)$$

From the proof of Proposition 1 (Appendix A.1), we know that $\tilde{\mathbf{r}} = \sum_{k=1}^{K-1}(\tilde{l}_k - \tilde{l}_{k+1})\mathbf{d}_k + \tilde{l}_K\mathbf{I}$ and $\hat{\beta}^{(\tilde{r})} = \sum_{k=1}^{K-1}(\tilde{l}_k - \tilde{l}_{k+1})\hat{\beta}_k^{(d)} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{l}_K\mathbf{I}$. Substituting these expressions:

$$\tilde{\mathbf{e}} = \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{d}_k + \tilde{l}_K \mathbf{I} - \mathbf{X} \left(\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\boldsymbol{\beta}}_k^{(d)} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \tilde{l}_K \mathbf{I} \right) \quad (9)$$

$$= \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{d}_k + \tilde{l}_K \mathbf{I} - \sum_{k=1}^{K-1} \mathbf{X} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\boldsymbol{\beta}}_k^{(d)} - \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \tilde{l}_K \mathbf{I} \quad (10)$$

$$= \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) (\mathbf{d}_k - \mathbf{X} \hat{\boldsymbol{\beta}}_k^{(d)}) + \tilde{l}_K \mathbf{I} - \tilde{l}_K \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{I} \quad (11)$$

$$= \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{e}_{dk} + \tilde{l}_K (\mathbf{I} - \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{I} \quad (12)$$

$$= \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \mathbf{e}_{dk}. \quad (13)$$

The last equality follows because $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix onto the column space of \mathbf{X} . Since \mathbf{X} includes a constant, \mathbf{I} lies in its column space, making $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I} = \mathbf{I}$.

A.3.2 Variance-covariance matrix expressions

Using the residual decomposition above, we can express the variance-covariance matrix estimator $\hat{\boldsymbol{\Omega}}$ for different error structures.

Homoskedastic standard errors. Under the assumption of homoskedastic errors, the variance estimator is:

$$\hat{\Omega}_{vanilla} = \hat{\sigma}^2 = \frac{1}{N-M} \sum_{i=1}^N \tilde{e}_i^2 = \frac{1}{N-M} \sum_{i=1}^N \left[\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) e_{dk,i} \right]^2, \quad (14)$$

where N is the number of observations and M is the number of regressors, and $e_{dk,i}$ is the residual for observation i from the regression of d_{ki} on \mathbf{X} .

Heteroskedasticity-robust standard errors. The Huber-White heteroskedasticity-robust variance estimator is:

$$\hat{\Omega}_{robust} = \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \tilde{e}_i^2 = \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i' \left[\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) e_{dk,i} \right]^2. \quad (15)$$

Clustered standard errors. For G clusters, the clustered variance estimator is:

$$\hat{\Omega}_{clustered} = \sum_{g=1}^G \left(\sum_{i \in g} \mathbf{x}_i \tilde{e}_i \right) \left(\sum_{i \in g} \mathbf{x}_i \tilde{e}_i \right)'. \quad (16)$$

Substituting the residual decomposition:

$$\hat{\Omega}_{\text{clustered}} = \sum_{g=1}^G \left(\sum_{i \in g} \mathbf{x}_i \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) e_{dk,i} \right) \left(\sum_{i \in g} \mathbf{x}_i \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) e_{dk,i} \right)'. \quad (17)$$

A.4 Proof of Proposition 3

We begin with the identity established in the proof of Proposition 1:

$$\hat{\beta}_m^{(\tilde{r})} = \sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\beta}_{km}^{(d)}$$

For any two variables m and n , the ratio of their coefficients is:

$$\frac{\hat{\beta}_m^{(\tilde{r})}}{\hat{\beta}_n^{(\tilde{r})}} = \frac{\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\beta}_{km}^{(d)}}{\sum_{k=1}^{K-1} (\tilde{l}_k - \tilde{l}_{k+1}) \hat{\beta}_{kn}^{(d)}}$$

We consider two cases, depending on whether the coefficient in the denominator is reversible.

Case 1: $\hat{\beta}_n^{(\tilde{r})}$ is not reversible

If $\hat{\beta}_n^{(\tilde{r})}$ is not reversible across all positive monotonic transformations, then by Proposition 1, all $\hat{\beta}_{kn}^{(d)}$ share the same sign. First assume that all $\hat{\beta}_{kn}^{(d)} > 0$. Let $\tilde{w}_k = -(\tilde{l}_k - \tilde{l}_{k+1})$. Given that $\tilde{l}_k - \tilde{l}_{k+1} < 0$ for all positive monotonic transformations, we have $\tilde{w}_k > 0$. We can rewrite the ratio as:

$$\begin{aligned} \frac{\hat{\beta}_m^{(\tilde{r})}}{\hat{\beta}_n^{(\tilde{r})}} &= \frac{\sum_{k=1}^{K-1} \tilde{w}_k \hat{\beta}_{km}^{(d)}}{\sum_{k=1}^{K-1} \tilde{w}_k \hat{\beta}_{kn}^{(d)}} = \frac{\sum_{k=1}^{K-1} \tilde{w}_k \hat{\beta}_{kn}^{(d)} \frac{\hat{\beta}_{km}^{(d)}}{\hat{\beta}_{kn}^{(d)}}}{\sum_{k=1}^{K-1} \tilde{w}_k \hat{\beta}_{kn}^{(d)}} = \frac{1}{\sum_{k=1}^{K-1} \tilde{w}_k \hat{\beta}_{kn}^{(d)}} \sum_{k=1}^{K-1} \tilde{w}_k \hat{\beta}_{kn}^{(d)} \frac{\hat{\beta}_{km}^{(d)}}{\hat{\beta}_{kn}^{(d)}} \\ &= \sum_{k=1}^{K-1} \frac{\tilde{w}_k \hat{\beta}_{kn}^{(d)}}{\sum_{j=1}^{K-1} \tilde{w}_j \hat{\beta}_{jn}^{(d)}} \frac{\hat{\beta}_{km}^{(d)}}{\hat{\beta}_{kn}^{(d)}} = \sum_{k=1}^{K-1} \alpha_k \frac{\hat{\beta}_{km}^{(d)}}{\hat{\beta}_{kn}^{(d)}} \end{aligned}$$

Since $\tilde{w}_k > 0$ and $\hat{\beta}_{kn}^{(d)} > 0$ for all k (by assumption), we have $\alpha_k > 0$ for all k . Additionally, $\sum_{k=1}^{K-1} \alpha_k = 1$. Therefore, the ratio $\hat{\beta}_m^{(\tilde{r})}/\hat{\beta}_n^{(\tilde{r})}$ is a convex combination of the ratios $\hat{\beta}_{km}^{(d)}/\hat{\beta}_{kn}^{(d)}$, where the weights are given by $\alpha_k \equiv \frac{\tilde{w}_k \hat{\beta}_{kn}^{(d)}}{\sum_{j=1}^{K-1} \tilde{w}_j \hat{\beta}_{jn}^{(d)}}$.

Thus, the ratio $\hat{\beta}_m^{(\tilde{r})}/\hat{\beta}_n^{(\tilde{r})}$ must lie between the minimum and maximum values of $\hat{\beta}_{km}^{(d)}/\hat{\beta}_{kn}^{(d)}$:

$$\min_k \frac{\hat{\beta}_{km}^{(d)}}{\hat{\beta}_{kn}^{(d)}} < \frac{\hat{\beta}_m^{(\tilde{r})}}{\hat{\beta}_n^{(\tilde{r})}} < \max_k \frac{\hat{\beta}_{km}^{(d)}}{\hat{\beta}_{kn}^{(d)}}$$

By choosing appropriate values for \tilde{w}_k (which corresponds to choosing an appropriate positive monotonic transformation), we can make the ratio $\hat{\beta}_m^{(\tilde{r})}/\hat{\beta}_n^{(\tilde{r})}$ arbitrarily close to either bound. For example, to approach the maximum value $\max_k \frac{\hat{\beta}_{km}^{(d)}}{\hat{\beta}_{kn}^{(d)}}$, we could choose a transformation where \tilde{w}_k is very large for the k that maximizes $\frac{\hat{\beta}_{km}^{(d)}}{\hat{\beta}_{kn}^{(d)}}$ and very small for all other values of k . Finally, when the signs of the $\hat{\beta}_{kn}^{(d)}$ are all negative, the same argument applies, except that the inequalities are reversed due to the negative sign in the denominator. However, the bounds remain the same.

Case 2: $\hat{\beta}_n^{(\tilde{r})}$ is reversible

Now consider the case where $\hat{\beta}_n^{(\tilde{r})}$ can be reversed by some positive monotonic transformation. In that case, the ratio $\frac{\hat{\beta}_m^{(\tilde{r})}}{\hat{\beta}_n^{(\tilde{r})}}$ is not bounded. To see this, note that since $\hat{\beta}_n^{(\tilde{r})}$ is reversible, we can find a transformation such that $\hat{\beta}_n^{(\tilde{r})} = \varepsilon$ for some arbitrarily small $\varepsilon > 0$. Depending on the sign of $\hat{\beta}_m^{(\tilde{r})}$ for that transformation, this will cause the ratio $\frac{\hat{\beta}_m^{(\tilde{r})}}{\hat{\beta}_n^{(\tilde{r})}}$ to be arbitrarily large negative (for $\hat{\beta}_m^{(\tilde{r})} < 0$) or positive (for $\hat{\beta}_m^{(\tilde{r})} \geq 0$). By the same argument, we can always find another transformation such that $\hat{\beta}_n^{(\tilde{r})} = \epsilon$ for some arbitrarily small $\epsilon < 0$, and obtain an arbitrarily large positive (for $\hat{\beta}_m^{(\tilde{r})} < 0$) or large negative (for $\hat{\beta}_m^{(\tilde{r})} \geq 0$) ratio.²⁶

A.5 Derivation of $\max\text{Var}(\Delta\tilde{\mathbf{I}})$

We here show that $\max\text{Var}(\Delta\tilde{\mathbf{I}}) = \left(\frac{1}{K-1} - \frac{1}{(K-1)^2}\right) (l_K - l_1)^2$. Let differences between adjacent labels be given by $d_k \equiv \tilde{l}_{k+1} - \tilde{l}_k$ for $k = 1, \dots, K-1$. The variance of these differences is given by:

$$\text{Var}(\Delta\tilde{\mathbf{I}}) = \frac{1}{K-1} \sum_{k=1}^{K-1} (d_k - \bar{d})^2 \quad (18)$$

Where $\bar{d} = \frac{1}{K-1} \sum_{k=1}^{K-1} d_k = \frac{l_K - l_1}{K-1}$ is the mean difference. Now note that the variance is maximized when these differences are as spread out as possible. Given the constraint that all differences must be positive (our **Monotonicity** constraint) and sum to $L = l_K - l_1$ (our **Normalization** constraint), the maximum variance occurs when one difference approaches L and all other $K-2$ differences approach 0. The maximum variance is then:

$$\max\text{Var}(\Delta\tilde{\mathbf{I}}) = \frac{1}{K-1} [(L - \bar{d})^2 + (K-2)(0 - \bar{d})^2] \quad (19)$$

²⁶We exclude the degenerate case here where $\hat{\beta}_m^{(\tilde{r})}$ switches sign for exactly the same transformation as $\hat{\beta}_n^{(\tilde{r})}$.

Substituting $\bar{d} = \frac{L}{K-1}$:

$$\max \text{Var}(\Delta \tilde{\mathbf{l}}) = \frac{1}{K-1} \left[\left(L - \frac{L}{K-1} \right)^2 + (K-2) \left(\frac{L}{K-1} \right)^2 \right] \quad (20)$$

$$= \frac{1}{K-1} \left[L^2 \left(\frac{K-2}{K-1} \right)^2 + (K-2) \frac{L^2}{(K-1)^2} \right] \quad (21)$$

$$= \frac{L^2}{(K-1)^3} [(K-2)^2 + (K-2)] \quad (22)$$

$$= \frac{L^2(K-2)(K-1)}{(K-1)^3} \quad (23)$$

$$= \frac{K-2}{(K-1)^2} L^2 \quad (24)$$

$$= \left(\frac{1}{K-1} - \frac{1}{(K-1)^2} \right) (l_K - l_1)^2 \quad (25)$$

B Making C comparable across scales with varying numbers of response options

We here provide a justification for setting $\alpha = 2 \log_{10}(K-1)$ in the cost function $C_\alpha(\tilde{\mathbf{l}})$ when comparing transformations across scales with varying numbers of categories. We also show why our standard cost function (i.e., setting $\alpha = 2$) becomes problematic as the number of labels increases.

Consider a continuous function $f : [0, 1] \rightarrow [0, 1]$ with which we plan to recode our dependent variable r . This restriction to the unit interval is without loss of generality, as any monotonic transformation can be normalized to this domain and range. Depending on the number of response options K for r , we can think of this function as being *sampled* at K equidistant points (resulting in $K-1$ differences between adjacent points). The pattern of differences between response options in turn approximates the derivative of the function, scaled by the sampling interval.

When we sample a continuous function at K equidistant points, each difference can be expressed as:

$$d_i \Delta \tilde{l}_k = f(x_{i+1}) - f(x_i) \approx f'(x_i) \cdot \Delta x = f'(x_i) \cdot \frac{1}{K-1} \quad (26)$$

In the context of our response scale transformation, these differences d_i correspond precisely to the differences between adjacent labels $\tilde{l}_{k+1} - \tilde{l}_k$, where the sampling points x_i correspond to the normalized positions of the original labels l_k in the interval $[0, 1]$.

To see how the variance of differences scales with the number of points, we calculate:

$$\text{Var}(d) = \frac{1}{K-1} \sum_{i=1}^{K-1} (d_i - \bar{d})^2 \quad (27)$$

where \bar{d} is the mean difference:

$$\bar{d} = \frac{1}{K-1} \sum_{i=1}^{K-1} d_i = \frac{f(1) - f(0)}{K-1} = \frac{1}{K-1} \quad (28)$$

Substituting our expressions for d_i and \bar{d} :

$$\text{Var}(d) = \frac{1}{K-1} \sum_{i=1}^{K-1} \left(f'(x_i) \cdot \frac{1}{K-1} - \frac{1}{K-1} \right)^2 \quad (29)$$

$$= \frac{1}{K-1} \sum_{i=1}^{K-1} \frac{1}{(K-1)^2} (f'(x_i) - 1)^2 \quad (30)$$

$$= \frac{1}{(K-1)^2} \sum_{i=1}^{K-1} \frac{1}{(K-1)} (f'(x_i) - 1)^2 \quad (31)$$

As K increases, this sum approaches an integral:

$$\frac{1}{(K-1)^2} \sum_{i=1}^{K-1} \frac{1}{(K-1)} (f'(x_i) - 1)^2 \approx \frac{1}{(K-1)^2} \int_0^1 (f'(x) - 1)^2 dx \quad (32)$$

If we denote the variance of the derivative function over $[0, 1]$ as $\sigma_{f'}^2$, which is a fixed value for any given function f , then:

$$\text{Var}(d) \approx \frac{\sigma_{f'}^2}{(K-1)^2} \quad (33)$$

Hence, the variance of differences scales by a factor of $1/(K-1)^2$ for a fixed pattern of non-linearity as the number of sampling points increases.

Now we may notice that since $\max \text{Var}(d) = \left(\frac{1}{K-1} - \frac{1}{(K-1)^2} \right) \approx \frac{1}{K-1}$ for large K , we have:

$$\frac{\text{Var}(d)}{\max \text{Var}(d)} \approx \frac{\sigma_{f'}^2 / (K-1)^2}{1/(K-1)} = \sigma_{f'}^2 \frac{1}{(K-1)} \quad (34)$$

This ratio, therefore, scales approximately by a factor $1/(K-1)$ for any fixed continuous function as the sampling resolution (i.e., the number of response options) increases. We would

like to reduce this dependency on K in our cost function. Although completely eliminating this dependency would require knowing the variance of the derivative of the transformation function ($\sigma_{f'}^2$) in advance, we can mitigate it through our choice of exponent α in the cost function. Specifically, we choose an exponent that makes $\left(\frac{1}{K-1}\right)^{1/\alpha}$ constant:

$$\alpha = 2 \log_{10}(K - 1) \quad (35)$$

With this adjustment, for any value of K we obtain $\left(\frac{1}{K-1}\right)^{1/(2 \log_{10}(K-1))} = 10^{-1/2} \approx 0.316$. Notably, this adjustment works perfectly when the variance of the derivative of the transformation $\sigma_{f'}^2$ equals 1, while for other values of $\sigma_{f'}^2$, the dependency on K is substantially reduced but not eliminated.²⁷ Thus, with this adjustment, the cost function will yield more comparable values across scales with different numbers of response categories for the same type of transformation. Moreover, for the commonly used 11-point scales, this approach conveniently gives us $\alpha = 2 \log_{10}(10) = 2$, which is the setting we use in the main text.

C Further evidence on γ

C.1 Worst-case estimates for γ when $C > 0$

Despite finding in section 3.3 that $\gamma_m \approx 0$ if scale use were linear (i.e., for $C = 0$), it remains unclear how γ_m would behave for non-linear scale use (i.e. $C > 0$). We perform a worst-case analysis on the potential influence and magnitude of γ_m in the case where scale use is non-linear. We do so in two steps:

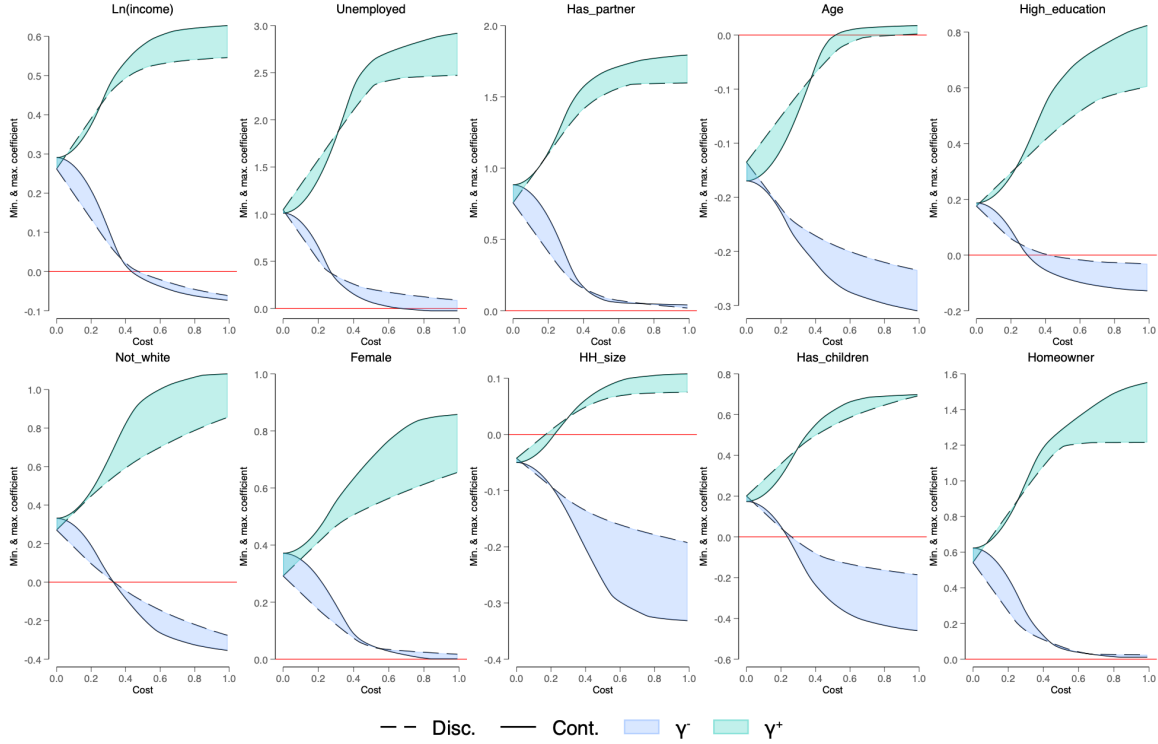
1. Using our continuous measure, and for increasing values of $C \in [0, 1]$ and every covariate m , we search for a transformation that yields a maximally positive and a maximally negative coefficient $\hat{\beta}_m^{(\tilde{r}^{(cont)})}$. This may, of course, involve a reversal of coefficient signs compared to the original coefficient assuming linear scale use $\hat{\beta}_m^{(r^{(cont)})}$.
2. We then check what maximal/minimal coefficient we would have obtained with a transformation of the same maximum cost if we only had our discrete 11-point variable. The

²⁷To see this, we note the full expression:

$$C_\alpha \approx \left(\frac{\sigma_{f'}^2}{K-1} \right)^{1/(2 \log_{10}(K-1))} = (\sigma_{f'}^2)^{1/(2 \log_{10}(K-1))} \cdot 10^{-1/2}$$

When $\sigma_{f'}^2 = 1$, the first term equals $1^{1/(2 \log_{10}(K-1))} = 1$. As in the unadjusted case for fixed α , for values of $\sigma_{f'}^2 > 1$, our cost will decrease as K increases. In contrast, for $\sigma_{f'}^2 < 1$, it will increase as K increases. However, this remaining dependency on K is much weaker than in the case of fixed α .

Figure A1: Worst-case evidence on γ_m when $C > 0$ (Prolific data)



Note: This figure displays worst-case scenarios for γ_m under non-linear scale use across several socioeconomic characteristics. The shaded regions represent the range of possible coefficient values achievable through transformations at each cost C , with teal regions (γ_m^+) showing maximum possible coefficients and blue regions (γ_m^-) showing minimum possible coefficients. Solid lines represent coefficients from continuous measurements. Dashed lines show coefficients from discrete measurements.

difference between coefficients $\hat{\beta}_m^{(\tilde{r}^{(cont)})}$ and $\hat{\beta}_m^{(\tilde{r}^{(disc)})}$ gives us a worst-case estimate of γ_m under non-linear scale use.

Unfortunately, as discussed in Appendix B, it is not, in general, possible to make the cost perfectly comparable across scales with vastly different numbers of response options. In order to at least ensure some comparability, we cannot let α be fixed (c.f. Section 2.4). Instead, we let $\alpha = 2\log_{10}(K - 1)$, as also derived in Appendix B. Thus, in what follows, when we write “ C ” we mean $C_{\alpha=2\log_{10}(K-1)}$.

The results of this analysis, for several socio-economics variables, and across many values for C , are shown in Figure A1. The shaded regions represent the range of possible coefficient values obtainable through transformations at a given cost C , with the upper region (teal) corresponding to γ_m^+ (i.e. where we maximise coefficients) and the lower region (blue) corresponding to γ_m^- (i.e. where we minimise coefficients). The dashed black lines show the

coefficient from the discrete measurement, while the solid lines show the coefficient from the continuous measurement.

We do not generally see that a reversal is possible at a lower cost for the continuous measure.²⁸ However, as C approaches 1, it is almost universally the case that the range of possible coefficient values is somewhat larger for the continuous measure than for the discrete measure. As a consequence, the continuous measure leads to an additional reversal for the case of unemployment (at $C = 0.67$), while no reversal for unemployment is possible in the discrete case. These wider ranges for the continuous scale imply larger potential values of γ_m under extreme non-linear scale use. However, we have little empirical evidence in favour of such strongly non-linear scale use (c.f. Section 3)

C.2 Effect of number of response categories

The previous evidence suggests that the possible range of coefficient values will be wider when using more response options. Intuitively, this is because the scope for within-category heterogeneity will be larger when there are fewer response categories. To further understand the magnitude of this phenomenon, we now extend our analysis of the previous section in two ways. First, we replicate the analysis on the same three additional datasets as already discussed in section 3.3. Second, we now also consider 3-point and 7-point scales, alongside our original 11-point discrete scale. Given that we do not *observe* data on these types of measures, we create these 3-point and 7-point scales by discretizing our continuous measurement at equidistant points.²⁹

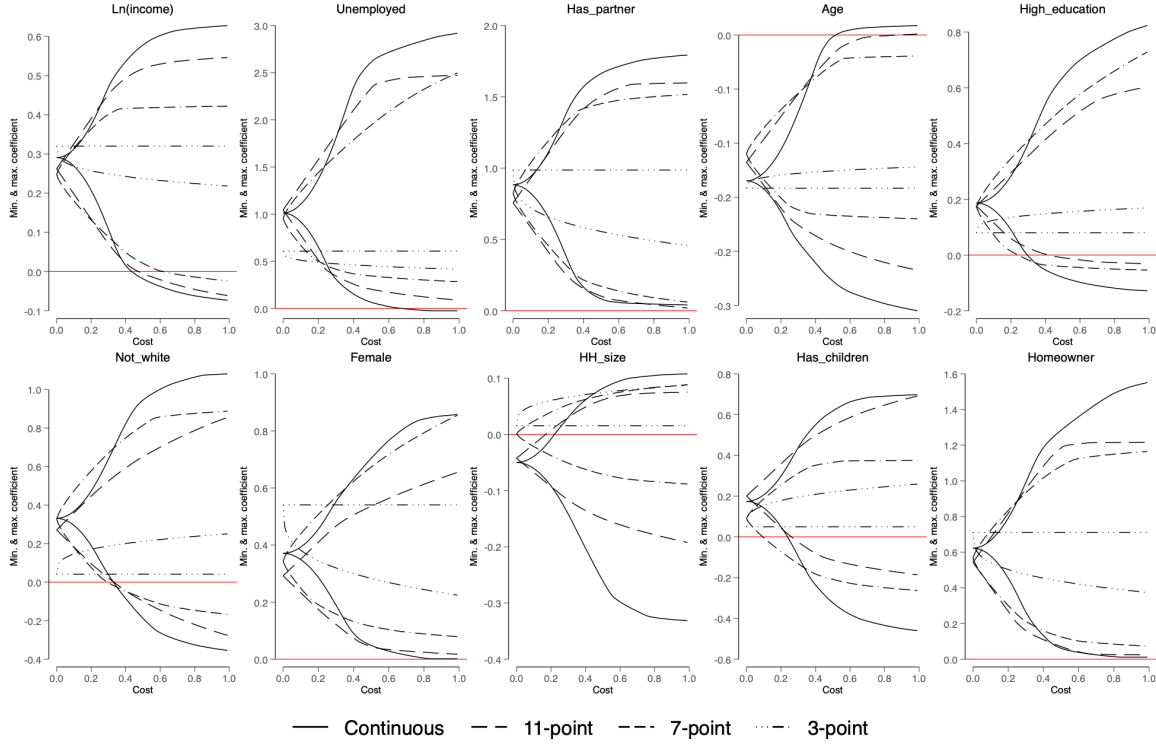
Figure A2 and Figures A8-A10 show the results. In line with the evidence of the previous section, we broadly observe that increasing the number of response categories also increases the possible spread of coefficient values at very high costs. As expected, we observe the smallest coefficient spreads for three response categories, and the largest spreads for our continuous measures. In turn, this again makes it generally more likely to reverse a coefficient when more response categories are available – especially when allowing for large values for C .

As a further piece of evidence, and to show this more systematically, we analyze for all of the variables and datasets discussed thus far, as well as pooling across datasets, how the mean cost of reversal (and the share of feasible reversals), varies with the number of categories. Since we do not observe all these n -point response scales, we construct them by discretizing

²⁸The S-shaped pattern observed in Figure A1 reflects how our cost adjustment affects scales with different numbers of categories. For small costs, our α adjustment decreases the cost parameter for the continuous measurement relative to what a fixed $\alpha = 2$ would yield, while for larger costs, it increases the relative cost.

²⁹This yields $\{0, 5, 10\}$ for the three-point scale and $\{0, 1.66, 3.33, 5, 6.66, 8.33, 10\}$ for the seven-point scale.

Figure A2: Coefficient spreads when $C > 0$ for different discretizations of r (Prolific data)



Note: This figure shows coefficient ranges for different variables across various scale discretizations (continuous, 11-point, 7-point, and 3-point scales) as the cost parameter C increases from 0 to 1. At high costs and for each variable, continuous measurements typically show the largest possible range of values, followed by 11-point, 7-point, and finally 3-point scales.

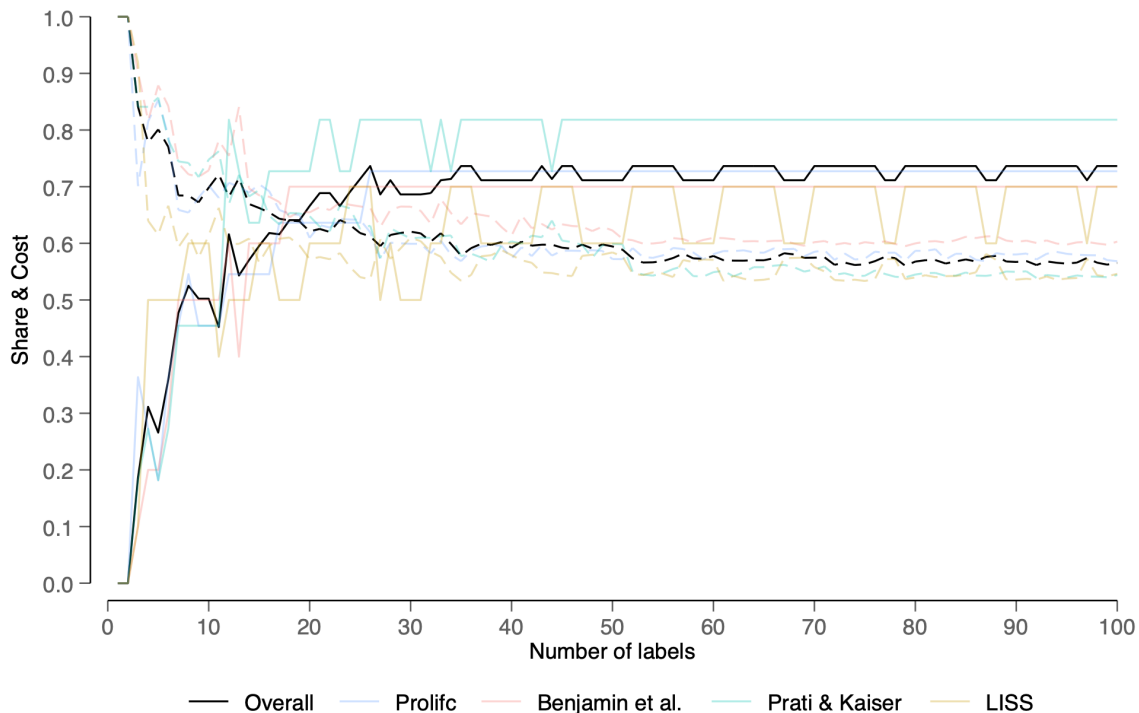
our original continuous 0-10 measure by rounding using $r_i^{(nlabs)} = \text{round}(r_i^{(cont)}, 10/(nlabs - 1))$, where the second argument of $\text{round}(\cdot, \cdot)$ gives the units to which we round. Figure A3 shows our results.

We generally observe the share of reversible coefficients (as indicated by the solid line) to increase when the number of response options is low, stabilizing at about 30 response options.³⁰ Likewise, the mean cost of reversals tends to decline with more response options. Overall, we observe that the share of reversible coefficients is about 20%-points higher for continuous scales than for discrete 11-point scales.

How do these findings impact the results of the main text? It seems clear that the share of reversible coefficient would be larger if satisfaction was measured on continuous scales in the literature. We can get some sense of this by comparing the share of reversible results in our replication effort when distinguishing between coefficients based on 10 or more categories, or

³⁰When only 2 response options are available, reversals are *never* possible. In that case the conditional of Proposition 1 is always trivially met.

Figure A3: Costs of reversals and shares of reversible coefficients as a function of the available number of response categories



Note: This figure illustrates how the number of available response categories affects both the share of reversible coefficients (solid lines) and the mean cost of reversals (dashed lines) across different datasets (as well as pooling across them). The share of reversible coefficients generally increases with the number of response options until approximately 30 categories, after which it stabilizes. Conversely, the mean cost of reversals tends to decline as more response options become available. With only 2 response options, reversals are impossible as the conditional of Proposition 1 is trivially met in this case. This is reflected by the zero values at the left edge of the graph.

coefficients based on fewer categories changes (there is only paper (34 estimates) in Wellbase with more than 11 categories). Figure A4 shows the results of this exercise. As expected, the share of reversible results is much larger in the case of results with 10 or more categories.

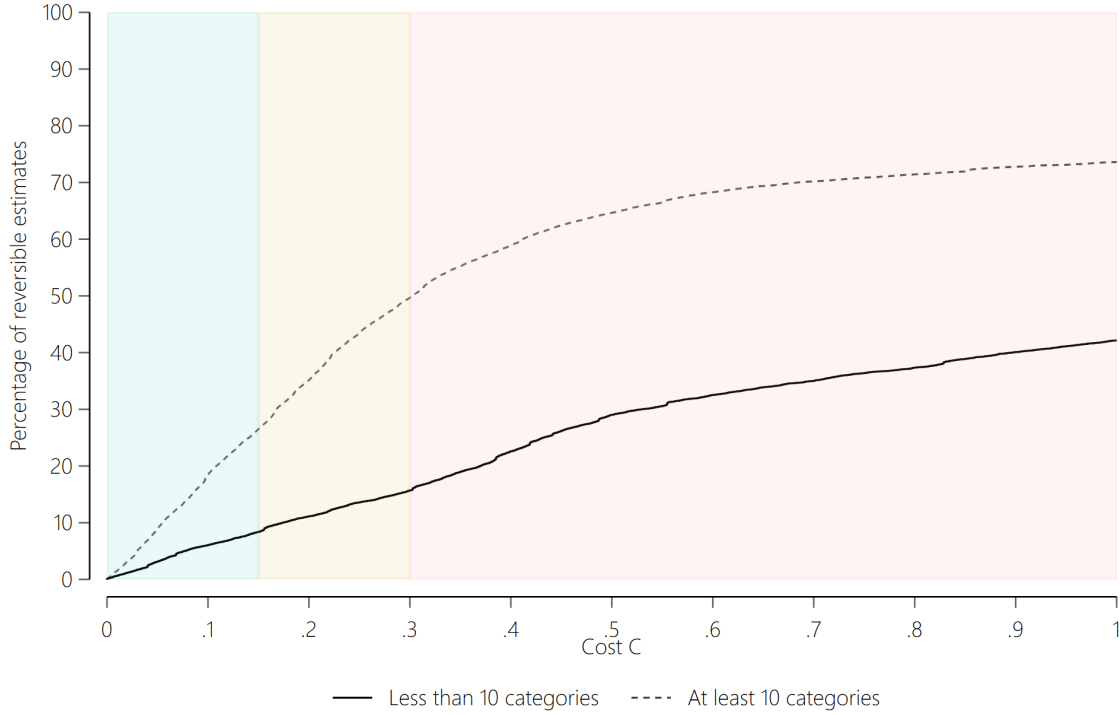
C.3 Implications

Three conclusions emerge from this analysis.

First, if satisfaction were measured continuously, the share of reversible results in the wellbeing literature would likely be somewhat higher than what we currently observe with discrete scales.

Second, however, such reversals would rely on extremely well-targeted transformations

Figure A4: Sign Reversal per categories



Notes: The figure shows the percentage of reversible estimates in *WellBase* by the number of response categories in the satisfaction scale. The solid line represents reversibility shares for scales with fewer than 10 response categories, while the dashed line represents scales with at least 10 categories.

that exploit a worst-case scenarios for within-category heterogeneity in more discrete measures. When we simply look at the magnitude of γ_m in the case of $C = 0$ (i.e. linear scale use), we find that γ_m is typically close to zero. This suggests that the assumption of favourable within-category heterogeneity is reasonable for low values of C .

Third, researchers should be especially cautious when working with wellbeing data based on few response categories (e.g., 3-point or 4-point scales). For such data, it is not feasible to conservatively assess the robustness of results again positive monotonic transformations of the response scale.

Figure A5: Evidence on γ_m when $C = 0$ (data from Benjamin et al.)

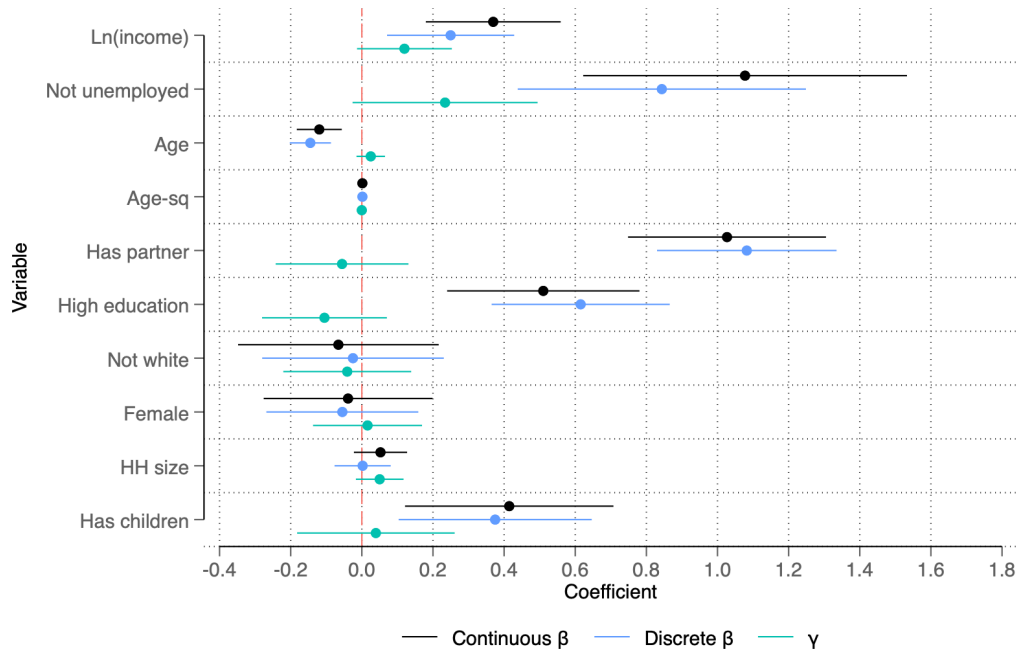


Figure A6: Evidence on γ_m when $C = 0$ (data from Prati & Kaiser)

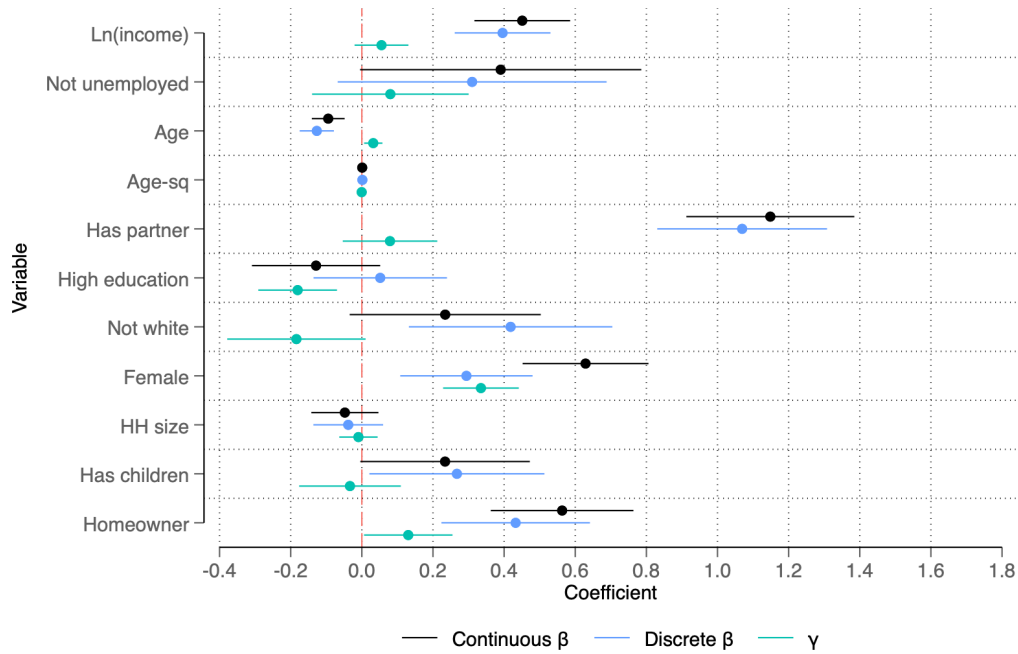


Figure A7: Evidence on γ_m when $C = 0$ (data from LISS)

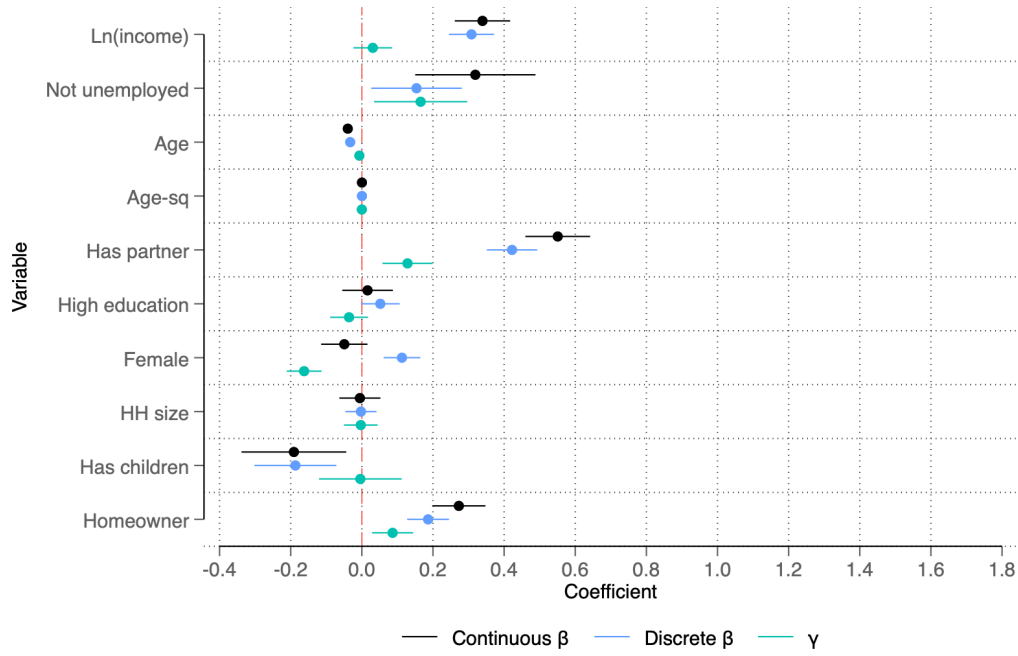
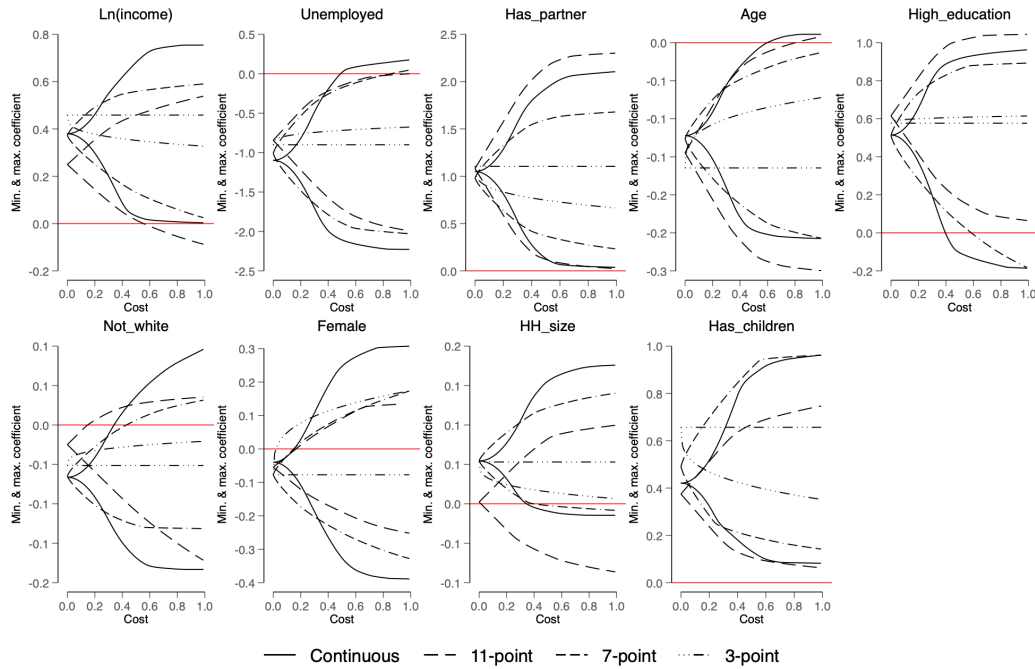
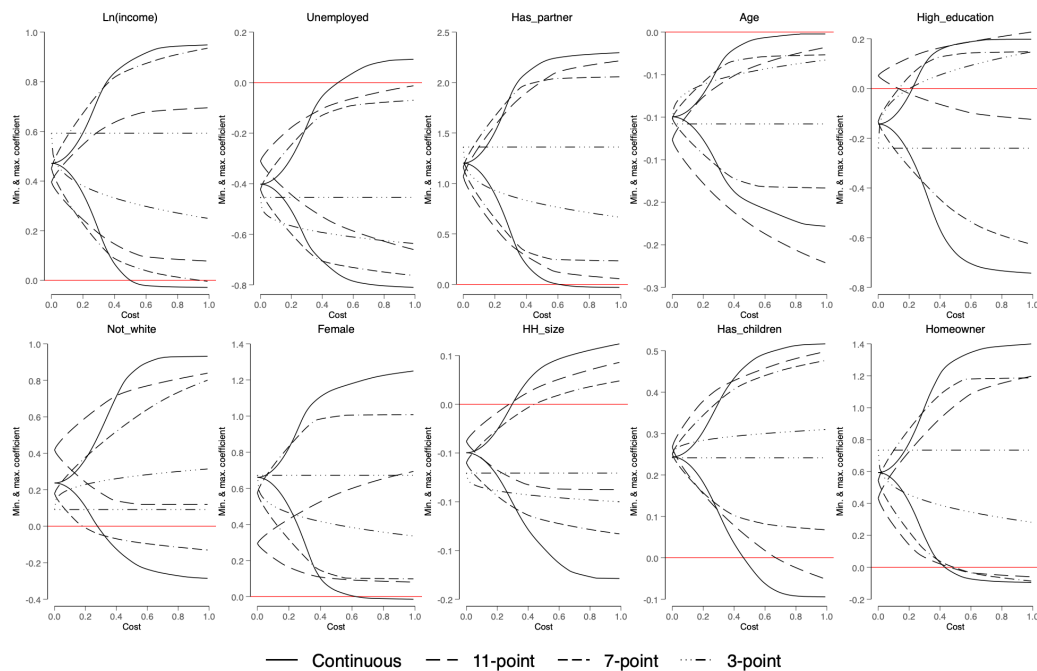


Figure A8: Coeff. spreads for different discretizations of r (Benjamin et al. data)



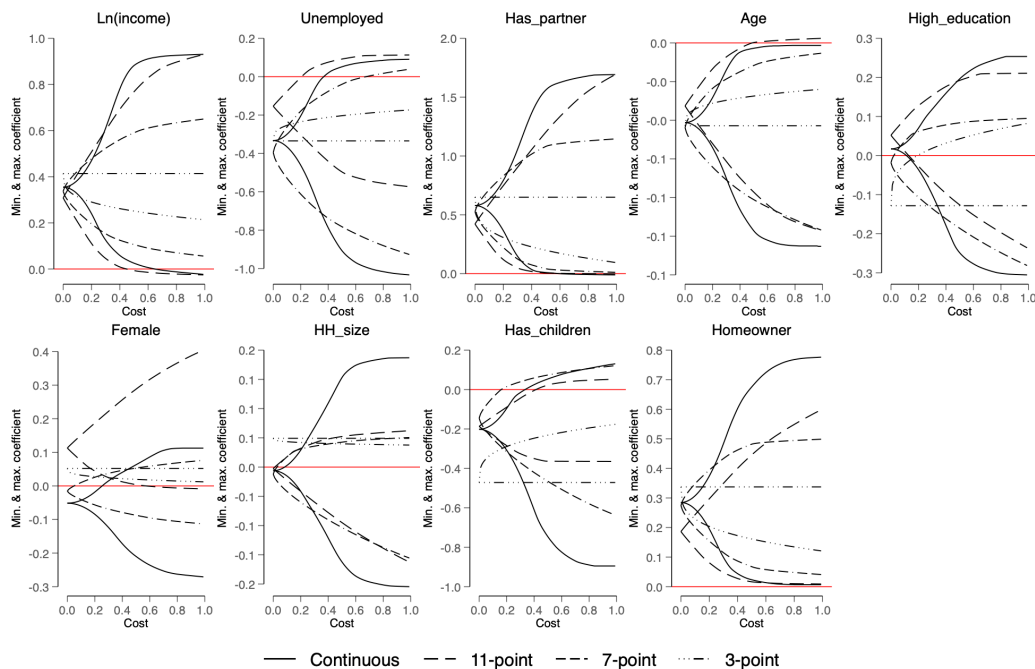
Note: C.f. note on Figure A2.

Figure A9: Coeff. spreads for different discretizations of r (Prati & Kaiser data)



Note: C.f. note on Figure A2.

Figure A10: Coeff. spreads for different discretizations of r (LISS data)



Note: C.f. note on Figure A2.

D Survey screenshots

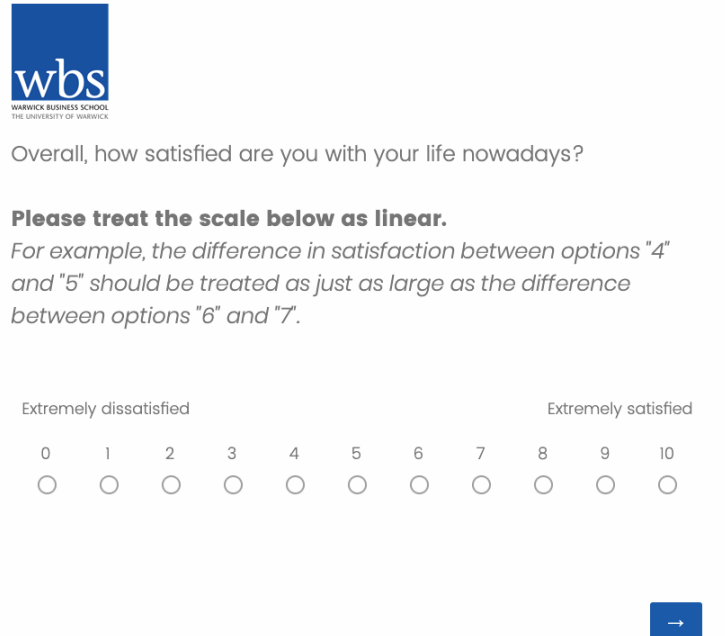
Figure A11: Screenshot of standard life satisfaction question



The screenshot shows a survey interface for Warwick Business School (WBS). At the top left is the WBS logo. The question is "Overall, how satisfied are you with your life nowadays?". Below the question is a horizontal scale from 0 to 10. The scale is labeled "Extremely dissatisfied" at the left end and "Extremely satisfied" at the right end. Each number from 0 to 10 has a corresponding radio button below it. At the bottom right of the scale is a blue button with a white right-pointing arrow.

Note: Screenshot taken from Qualtrics.

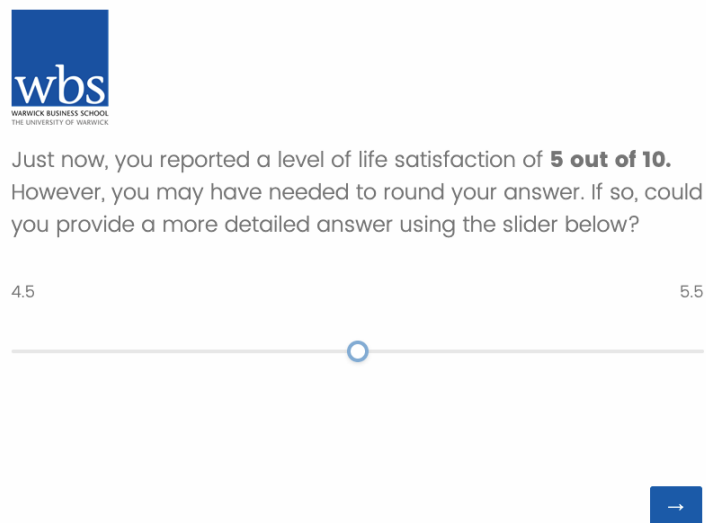
Figure A12: Screenshot of life satisfaction question with linear prompt



The screenshot shows a survey interface for Warwick Business School (WBS). At the top left is the WBS logo. The question is "Overall, how satisfied are you with your life nowadays?". Below the question is a bold prompt: "Please treat the scale below as linear." followed by an italicized explanation: "For example, the difference in satisfaction between options '4' and '5' should be treated as just as large as the difference between options '6' and '7'." Below this is a horizontal scale from 0 to 10. The scale is labeled "Extremely dissatisfied" at the left end and "Extremely satisfied" at the right end. Each number from 0 to 10 has a corresponding radio button below it. At the bottom right of the scale is a blue button with a white right-pointing arrow.

Note: Screenshot taken from Qualtrics.

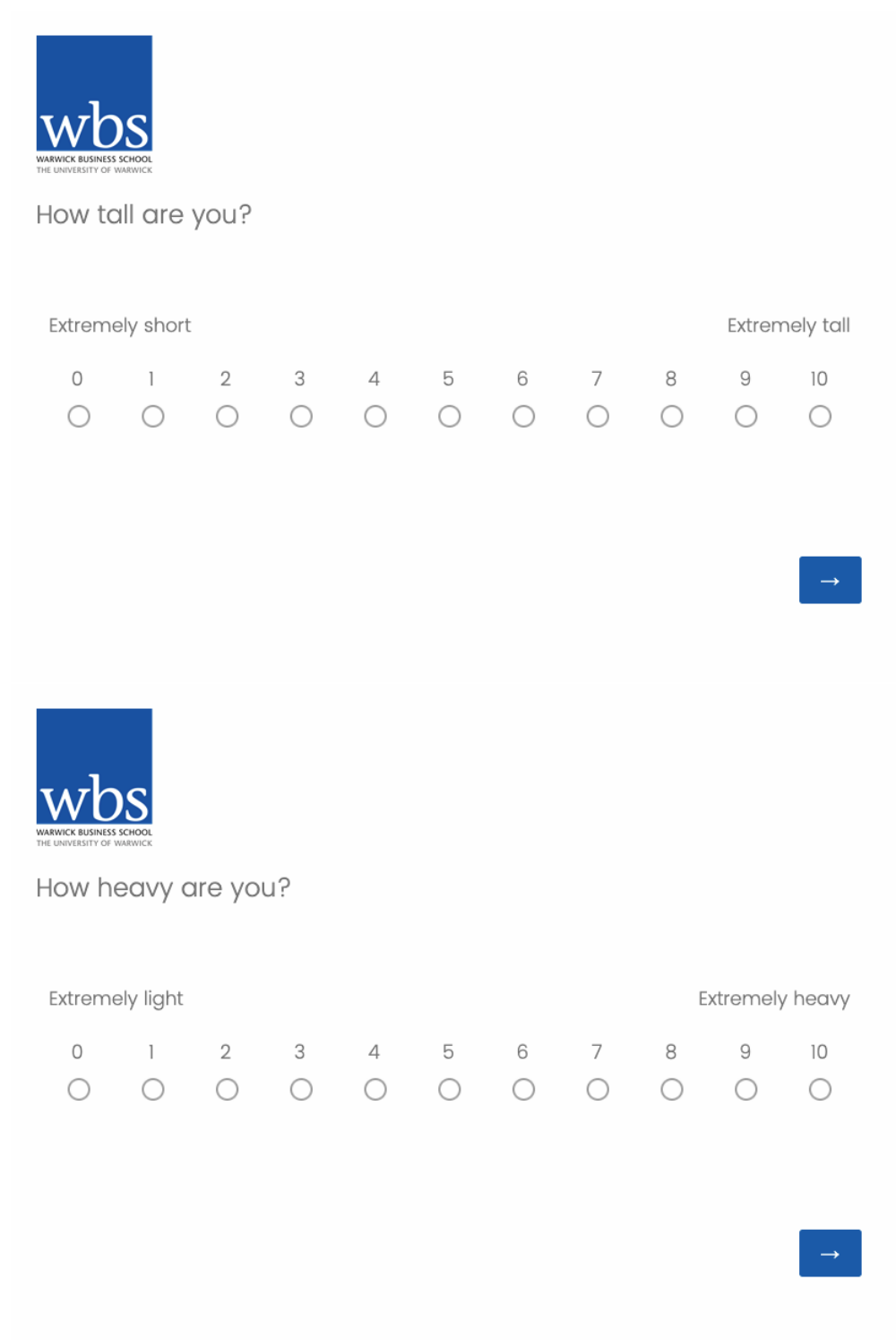
Figure A13: Screenshot of follow-up question to obtain a continuous measure of satisfaction



The screenshot shows a survey interface for Warwick Business School (WBS). At the top left is the WBS logo. The text reads: "Just now, you reported a level of life satisfaction of **5 out of 10**. However, you may have needed to round your answer. If so, could you provide a more detailed answer using the slider below?". Below the text is a horizontal slider bar with numerical labels "4.5" on the left and "5.5" on the right. A blue circular slider handle is positioned exactly halfway between 4.5 and 5.5. At the bottom right of the slider area is a blue square button with a white right-pointing arrow.

Note: Screenshot taken from Qualtrics. The slider allows for 100 different values within a given category. Different versions of this question are provided for every possible response option in the original life satisfaction question.

Figure A14: Screenshot of question on subjective height and weight



The image shows two screenshots of a survey interface. The top screenshot is for the question "How tall are you?". It features the Warwick Business School (WBS) logo at the top left. Below the question, there is a horizontal scale from 0 to 10. The scale is labeled "Extremely short" at the left end and "Extremely tall" at the right end. Each number from 0 to 10 has a corresponding radio button below it. A blue button with a right-pointing arrow is located at the bottom right of the scale.

WBS
WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

How tall are you?

Extremely short

Extremely tall

0 1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

→

The bottom screenshot is for the question "How heavy are you?". It also features the WBS logo at the top left. Below the question, there is a horizontal scale from 0 to 10. The scale is labeled "Extremely light" at the left end and "Extremely heavy" at the right end. Each number from 0 to 10 has a corresponding radio button below it. A blue button with a right-pointing arrow is located at the bottom right of the scale.

WBS
WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

How heavy are you?

Extremely light

Extremely heavy

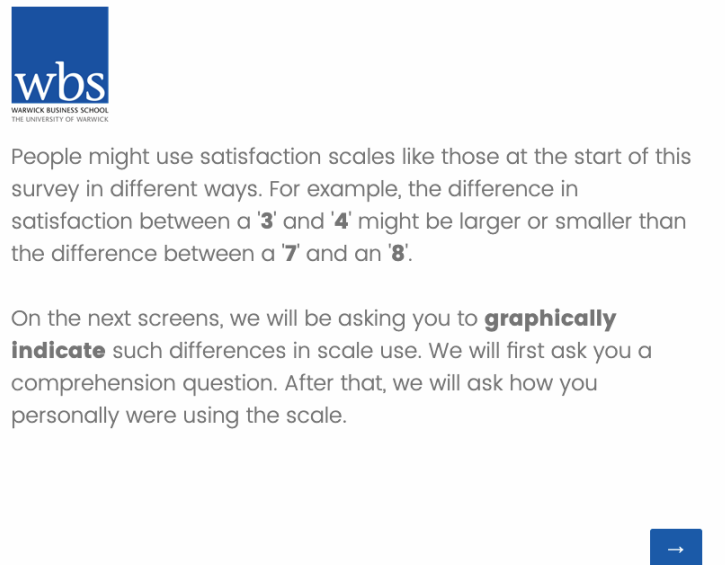
0 1 2 3 4 5 6 7 8 9 10

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

→

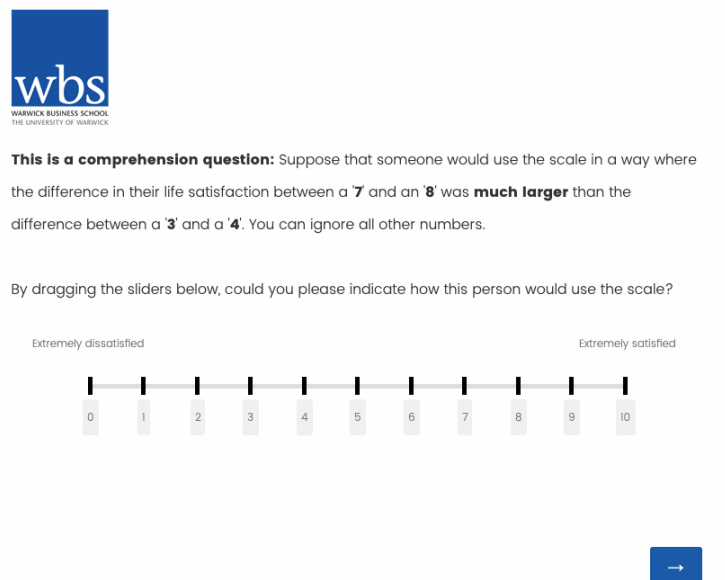
Note: Screenshot taken from Qualtrics. Questions are presented on separate pages.

Figure A15: Screenshot of introduction material to sliders questions



Note: Screenshot taken from Qualtrics.

Figure A16: Screenshot of comprehension question for sliders



Note: Screenshot taken from Qualtrics.

Figure A17: Screenshot of main sliders question

Warwick Business School
THE UNIVERSITY OF WARWICK

By dragging the sliders below, could you show us how **you personally** were using the scale to express your life satisfaction at the start of the survey?

Extremely dissatisfied Extremely satisfied

0 1 2 3 4 5 6 7 8 9 10

Large distances at bottom Large distances at top Equal distances Large distances in the middle Large distances at the extremes

(press the buttons to move all sliders)

→

Note: Screenshot taken from Qualtrics.

Figure A18: Screenshot of verification question when respondents indicate linear scale use

Warwick Business School
THE UNIVERSITY OF WARWICK

You did not move any of the sliders in the previous question.

Does this mean that, for you, the difference in satisfaction between response categories is always the same?
(i.e. the difference in satisfaction between, say, a '3' and a '4', is just as large as the difference between a '7' and an '8').

☐ No

☐ Yes

☐ Neither, please elaborate:

→

Note: Screenshot taken from Qualtrics.

E WellBase Details

Table A1: Description of Papers included in WellBase

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
Banks et al. (2010)	I am satisfied with my life	4	.	.	None
Clark and Senik (2010)	Taking all things together, how happy would you say you are?	11	.	✓	None
	All things considered, how satisfied are you with your life as a whole nowadays?	11			
	How satisfied are you with how your life has turned out so far?	11			
Knabe et al. (2010)	All things considered, how satisfied are you with your life as a whole these days?	11	✓	✓	None
Oswald and Wu (2011)	In general, how satisfied are you with your life?	6	✓	✓	Income logged-transformed and change of LFS reference category for Section 4.2.4
Bertrand (2013)	Taken all together, how would say things are these days - would you say that you are very happy, pretty happy or not too happy?	3	.	.	None
Vendrik (2013)	All things considered, how satisfied are you with your life as a whole these days?	11	✓	✓	None
Ashraf et al. (2014)	How satisfied are you with your life as a whole these days?	5	.	.	None
Frijters et al. (2014)	Here is a scale from 0 to 10, where "0" dissatisfied and "10" means that you are completely satisfied. Please enter the number which corresponds with how satisfied or dissatisfied you are with the way life has turned out so far.	11	✓	.	None
Kesternich et al. (2014)	On a scale from 0 to 10 where 0 means completely dissatisfied and 10 means completely satisfied, how satisfied are you with your life?	11	.	.	None

Continued on next page

Table A1 – *Continued from previous page*

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
Layard et al. (2014)	Here is a scale from 0 to 10. On it, “0” means that you are completely dissatisfied and “10” means that you are completely satisfied. Please tick the box with the number above it which shows how dissatisfied or satisfied you are about the way your life has turned out so far.	11	✓	✓	Income unstandardized for Section 4.2.4
Bloom et al. (2015)	How satisfied are you with your life as a whole these days?	7	.	.	None
Campante and Yanagizawa-Drott (2015)	Taking all things together, would you say you are: not at all happy, not very happy, quite happy, very happy?	4	.	.	Converted binary r to original continuous r
	How satisfied are you with your life as a whole these days?	10			
Dinkelman and Schulhofer-Wohl (2015)	Taking everything into account, how satisfied is the household with the way it lives these days?	5	.	.	Original measure of r in log; delogged for Well-Base
Oswald et al. (2015)	How would you rate your happiness at the moment?	6	.	.	Ordered probit replaced by OLS
Aghion et al. (2016)	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11	.	✓	None
	In general, how satisfied are you with your life?	4			
Clark et al. (2016)	How satisfied are you with your life, all things considered?	11	✓	.	None
Danzer and Danzer (2016)	To what extent are you satisfied with your life in general at the present time?	5	✓	✓	None
Gerritsen (2016)	How dissatisfied or satisfied are you with your life overall?	7	.	✓	None
Glaeser et al. (2016)	In general, how satisfied are you with your life?	4	.	.	Regressions based on propriety data missing

Continued on next page

Table A1 – *Continued from previous page*

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
Haushofer and Shapiro (2016)	Taking all things together, would you say you are ‘very happy’ (1), ‘quite happy’ (2), ‘not very happy’ (3), or ‘not at all happy’ (4)?”	4	.	.	None
	All things considered, how satisfied are you with your life as a whole these days?	11			
Cheng et al. (2017)	How satisfied are you with your life, all things considered?	11	.	.	None
	How dissatisfied or satisfied are you with your life overall?	7			
	All things considered, how satisfied are you with your life in general?	10			
	All things considered, how satisfied are you with your life?	11			
Blattman and Dercon (2018)	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11	.	.	None
Blumenstock et al. (2018)	All things considered, how satisfied are you with life as a whole?	11	.	.	None
De Neve et al. (2018)	On the whole, are you very satisfied, fairly satisfied, not very satisfied, or not at all satisfied with the life you lead?	4	.	.	None
	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11			
	In general, how satisfied are you with your life?	4			

Continued on next page

Table A1 – *Continued from previous page*

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
Johnston et al. (2018)	All things considered, how satisfied are you with your life?	11	✓	✓	Income logged-transformed and change of LFS reference category for Section 4.2.4
Dolan et al. (2019)	Overall, how satisfied are you with your life now-days?	11	✓	.	None
Fisher and Zhu (2019)	All things considered, how satisfied are you with your life?	11	.	.	None
Guriev and Treisman (2019)	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11	.	.	None
Heffetz and Reeves (2019)	In general, how satisfied are you with your life?	4	.	.	None
Odermatt and Stutzer (2019)	How satisfied are you with your life, all things considered?	11	✓	✓	None
Tur-Prats (2019)	How satisfied are you with your life as a whole these days?	10	.	.	Converted binary r to original continuous r
Allcott et al. (2020)	During the past 4 weeks, I was satisfied with my life	7	.	.	None
Blakeslee et al. (2020)	All things considered, how satisfied are you with your life as a whole these days?	10	.	.	None
Haushofer et al. (2020)	Taking all things together, would you say you are ‘very happy’ (1), ‘quite happy’ (2), ‘not very happy’ (3), or ‘not at all happy’ (4)?”	4	.	.	None
	All things considered, how satisfied are you with your life as a whole these days?	11			
Lee et al. (2020)	All things considered, how satisfied are you with your life as a whole these days?	10	.	✓	None

Continued on next page

Table A1 – *Continued from previous page*

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
Perez-Truglia (2020)	Will you mostly describe yourself as: Very happy; Quite happy; Not particularly happy; Not at all happy How satisfied are you with your life?	4	✓	.	Probit-adjusted OLS replaced by OLS
	How satisfied are you with your life, all things considered?	11			
Singh and Masters (2020)	How satisfied are you with your life, all things considered?	6	.	.	None
Aksoy and Tumen (2021)	All things considered, I am satisfied with my life now	5	.	.	None
Bessone et al. (2021)	How happy are you today?	5	.	.	None
	All things considered, how satisfied are you with your life as a whole?	10			
Bryan et al. (2021)	How would you describe your satisfaction with life?	4	.	.	
	Taking all things together, would you say you are	10			
Chen and Fang (2021)	Please think about your life-as-a-whole. How satisfied are you with it?	5	.	.	None
Dalton et al. (2021)	How satisfied are you with your life at this point?	10	.	.	None
Flèche (2021)	In general, how satisfied are you with your life?	11	✓	✓	Regressions with Municipality FE not reproduced
Huang et al. (2021)	Are you happy?	11	.	.	None
Kabátek and Ribar (2021)	How satisfied are you with the life you lead at the moment?	11	.	.	Ordered logit replaced by OLS
Levitt (2021)	All things considered, how happy are you as a whole right now?	10	.	.	None
Li (2021)	How happy are you?	5	.	.	None
	How satisfied are you with your life as a whole?	5			
Ajzenman et al. (2022)	All things considered, I am satisfied with my life now	5	.	.	None
Binder and Makridis (2022)	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11	✓	.	None

Continued on next page

Table A1 – *Continued from previous page*

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
Dahl et al. (2022)	Overall, how satisfied are you with your life?	11	.	.	None
Meier (2022)	How satisfied are you with your life, all things considered?	11	.	.	None
Adhvaryu et al. (2023)	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11	.	.	None
Bah et al. (2023)	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11	.	.	None
Carattini and Roesti (2023)	All things considered, how satisfied are you with your life as a whole nowadays? Ranges from 0 (extremely dissatisfied) to 10 (extremely satisfied)	11	✓	✓	SHP, ESS and SOM samples analyzed separately in Section 4.2.4
	Taking all things together, how happy would you say you are? - ranges from 0 (extremely unhappy) to 10 (extremely happy) In general, how satisfied are you with your life?	11			
	How satisfied as a whole, 1 (not at all) to 4 (very satisfied)	4			
Caria et al. (2023)	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11	.	.	None

Continued on next page

Table A1 – *Continued from previous page*

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
Coville et al. (2023)	All things considered, how satisfied are you with your life as a whole these days?	10	.	.	None
Edmonds et al. (2023)	Taking all things together, would you say you are: Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	4 11	 ✓	 .	 None
Gazeaud et al. (2023)	Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	11	.	.	None
Sarmiento et al. (2023)	How satisfied are you with your life, all things considered?	11	.	.	None
Sha (2023)	How satisfied are you with your life as a whole?	5	.	.	None
Stango and Zinman (2023)	How satisfied are you with your life as a whole these days?	100	.	.	None
Angelucci and Bennett (2024)	I am satisfied with my life	10	.	.	None
Ciancio et al. (2024)	How satisfied are you with your life, all things considered?	6	.	.	None
Clark and Zhu (2024)	All things considered, how satisfied are you with your life?	11	.	.	None
Giacobino et al. (2024)	Happiness question - wording not reported Life satisfaction question - wording not reported	4 10	 .	 .	 None

Continued on next page

Table A1 – *Continued from previous page*

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
Grimm et al. (2024)	Imagine for a moment that you are living the best life you can imagine living. Now, imagine a situation where your life is as bad as it could possibly be. Let's consider a scale from 1 to 6. Suppose we say that the top of the scale (6) represents the best possible life for you, and the bottom (1) represents the worst possible life for you. Which step of the scale best represents your current personal situation?	6	.	.	None
Krekel et al. (2024)	Overall, how satisfied are you with your life now-days?	11	.	.	None
Priebe et al. (2024)	Life Satisfaction question - not reported	5	.	.	None
Riley (2024)	Happiness question - not reported	5	.	.	None
	Life satisfaction question - not reported	10			
Vlassopoulos et al. (2024)	Taking all things together, how happy are you these days?	11	.	.	None.
	How satisfied are you with your life as a whole these days?	11			
Bjorvatn et al. (2025)	How happy are you with your life?	11	.	.	None
	In your opinion, where are you on the ladder of life at the moment?	11			
Courtemanche et al. (2025)	In general, how satisfied are you with your life?	11	✓	.	None

Note: This table lists all the papers included in *WellBase*.

Table A2: Risk of sign reversal and main conclusions of *WellBase*

Author	Test(s) of the paper	Source	Sign	Sig.	Reversal	Cost
Clark and Senik (2010)	Income	Table 7 - Column 4	+	1%	No	.
	Important to compare income	Table 7 - Column 4	-	1%	No	.
	Comparison direction: work colleagues	Table 7 - Column 4	-	5%	Yes	0.715
	Comparison direction: family members	Table 7 - Column 4	-	1%	Yes	0.164
	Comparison direction: others	Table 7 - Column 4	-	1%	Yes	0.138
	Comparison direction: don't compare	Table 7 - Column 4	-	1%	Yes	0.252
Knabe et al. (2010)	Unemployment	Table 5 - Column 3	-	1%	No	.
Bertrand (2013)	Having a job	Table 1 - Panel A	+	1%	No	.
	Being married	Table 1 - Panel A	+	1%	No	.
	Having a job and being married	Table 1 - Panel A	-	5%	No	.
	Having a job	Table 1 - Panel B	+	1%	No	.
	Having kids	Table 1 - Panel B	+	1%	No	.
	Having a job and having kids	Table 1 - Panel B	-	5%	No	.
Vendrik (2013)	Current own income	Table 1 - Column 5	+	1%	No	.
	Past own income (one year)	Table 1 - Column 5	-	NS	Yes	0.165
	Past own income (two years)	Table 1 - Column 5	-	NS	Yes	0.280
	Past own income (three years)	Table 1 - Column 5	+	10%	Yes	0.564
	Future own income (one year)	Table 1 - Column 5	+	1%	No	.
	Current reference income	Table 1 - Column 5	-	NS	Yes	0.321
	Past reference income (one year)	Table 1 - Column 5	-	10%	Yes	0.226
	Future reference income (one year)	Table 1 - Column 5	+	NS	Yes	0.053
Frijters et al. (2014)	Wage	Table 4 - Column 5	+	1%	Yes	0.226
	Employment	Table 4 - Column 5	+	1%	Yes	0.399
	Unemployment	Table 4 - Column 5	+	NS	Yes	0.084
	Married	Table 4 - Column 5	+	1%	Yes	0.733
	Poor Health	Table 4 - Column 5	-	1%	No	.
	Education	Table 4 - Column 5	+	NS	Yes	0.278
	Lagged satisfaction (age 46)	Table 4 - Column 5	+	1%	No	.
	Lagged satisfaction (age 42)	Table 4 - Column 5	+	1%	No	.
	Lagged satisfaction (age 33)	Table 4 - Column 5	+	1%	No	.
	Income	Table 1 - Column 3	+	1%	Yes	0.354
Layard et al. (2014)	Education	Table 1 - Column 3	+	1%	Yes	0.049
	Having a job	Table 1 - Column 3	+	1%	No	.
	Good conduct	Table 1 - Column 3	+	1%	No	.
	Having a partner	Table 1 - Column 3	+	1%	Yes	0.856

(Continued from previous page)

Author	Test(s) of the paper	Source	Sign	Sig.	Reversal	Cost
Campante and Yanagizawa-Drott (2015)	Self-perceived health	Table 1 - Column 3	+	1%	Yes	0.786
	Emotional health	Table 1 - Column 3	+	1%	No	.
	Female	Table 1 - Column 3	+	1%	Yes	0.15
	Ramadan hours	Table 2 - Column 12	+	1%	No	.
Oswald et al. (2015)	US States Fixed effects	Table 2 - Column 4	Mix	NS to 1%	44%	0.010 to 0.980
Aghion et al. (2016)	Job turnover rate	Table 2 - Column 3 - Panel B	+	5%	Yes	0.342
	Unemployment rate	Table 2 - Column 3 - Panel B	-	1%	No	.
	Job creation rate	Table 3 - Column 2 - Panel B	+	1%	Yes	0.575
	Job destruction rate	Table 3 - Column 2 - Panel B	-	1%	Yes	0.459
Clark et al. (2016)	Incidence of poverty	Table 2 - Column 1	-	1%	Yes	0.593
	Intensity of poverty	Table 2 - Column 1	-	1%	No	.
	0 to 1 years of poverty	Table 3 - Column 1	-	1%	No	.
	1 to 2 years of poverty	Table 3 - Column 1	-	1%	No	.
	2 to 3 years of poverty	Table 3 - Column 1	-	1%	No	.
	3 to 4 years of poverty	Table 3 - Column 1	-	1%	Yes	0.494
	4 to 5 years of poverty	Table 3 - Column 1	-	1%	Yes	0.324
	5 years of poverty or more	Table 3 - Column 1	-	1%	Yes	0.504
Danzer and Danzer (2016)	Radiation	Table 2 - Column 3	-	1%	Yes	0.962
Gerritsen (2016)	Income	Table 1 - Column 1	+	1%	No	.
	Hours of work	Table 1 - Column 1	+	5%	Yes	0.181
	Hours of work squared	Table 1 - Column 1	-	5%	Yes	0.115
Glaeser et al. (2016)	Population size	Table 1 - Column 2	-	5%	No	.
Cheng et al. (2017)	Age	Figure 2 - Panel A	-	1%	No	.
	Age squared	Figure 2 - Panel A	+	1%	No	.
	Age	Figure 2 - Panel B	-	1%	No	.
	Age squared	Figure 2 - Panel B	+	1%	No	.
	Age	Figure 2 - Panel C	-	1%	No	.
	Age squared	Figure 2 - Panel C	+	1%	No	.
	Age	Figure 2 - Panel D	-	1%	No	.
	Age squared	Figure 2 - Panel D	+	1%	No	.
De Neve et al. (2018)	Economic growth - World Sample	Table 1 - Column 1	+	1%	No	.
	Negative growth - World Sample	Table 1 - Column 2	-	1%	No	.
	Positive growth - World Sample	Table 1 - Column 2	+	NS	Yes	0.482
	Economic growth - European Sample	Table 1 - Column 3	+	1%	No	.

(Continued from previous page)

Author	Test(s) of the paper	Source	Sign	Sig.	Reversal	Cost
Johnston et al. (2018)	Negative growth - European Sample	Table 1 - Column 4	-	1%	No	.
	Positive growth - European Sample	Table 1 - Column 4	+	5%	No	.
	Economic growth - US Sample	Table 1 - Column 5	+	1%	No	.
	Negative growth - US Sample	Table 1 - Column 6	-	1%	No	.
	Positive growth - US Sample	Table 1 - Column 6	+	1%	No	.
	Victim of physical violence - Women sample	Table 3 - Column 1	-	1%	No	.
	Victim of physical violence - Men sample	Table 3 - Column 2	-	1%	No	.
Dolan et al. (2019)	Olympic games in London	Table 2 - Column 6	+	1%	No	.
Odermatt and Stutzer (2019)	Widowhood (zero to one year)	Table 2 - Column 2	-	1%	No	.
	Widowhood (five to six year)	Table 2 - Column 2	-	1%	Yes	0.081
	Unemployment (zero to one year)	Table 2 - Column 4	-	1%	No	.
	Unemployment (five to six year)	Table 2 - Column 4	-	1%	Yes	0.293
	Disability (zero to one year)	Table 2 - Column 6	-	1%	Yes	0.509
	Disability (five to six year)	Table 2 - Column 6	-	1%	Yes	0.182
	Plant closure (zero to one year)	Table 2 - Column 8	-	1%	No	.
Perez-Truglia (2020)	Plant closure (five to six year)	Table 2 - Column 8	-	1%	Yes	0.198
	Income Rank*2001–201*High Internet	Table 3 - Column 4	+	1%	No	.
	Income Rank*2001–2013*High Internet	Table 3 - Column 6	+	NS	Yes	0.428
Flèche (2021)	Centralization reforms	Table 1 - Column 4	-	1%	No	.
Levitt (2021)	All major life decisions after two months	Table 5 - Column 2 - Row 1	+	1%	No	.
	All major life decisions after two months	Table 5 - Column 3 - Row 1	+	NS	Yes	0.087
	All major life decisions after six months	Table 5 - Column 5 - Row 1	+	1%	No	.
	All major life decisions after six months	Table 5 - Column 6 - Row 1	+	5%	No	.
Li (2021)	First son * Sex ratio	Table 3 - Column 1	-	5%	No	.
	First son * Sex ratio	Table 3 - Column 2	-	1%	No	.
Dahl et al. (2022)	Post-reform*Immigrant	Table 1 - Column 4 - Panel A	-	1%	No	.
	Post-reform*Immigrant	Table 1 - Column 4 - Panel B	+	NS	Yes	0.123
Carattini and Roesti (2023)	Trust	Table 1 - Column 1	+	1%	No	.
Sarmiento et al. (2023)	LEZ introduction	Table 8 - Column 1	+	1%	Yes	0.310
Krekel et al. (2024)	Volunteering in England's NHS	Table 3 - Column 2	+	1%	No	.
Courtemanche et al. (2025)	Chain restaurant calorie posting laws	Table 5 - Column 2	+	1%	Yes	0.645

Note: This table lists the risk of sign reversal in all the papers included in *WellBase* for which at least half of the regressions printed uses a measure of cognitive subjective wellbeing as dependent variable.

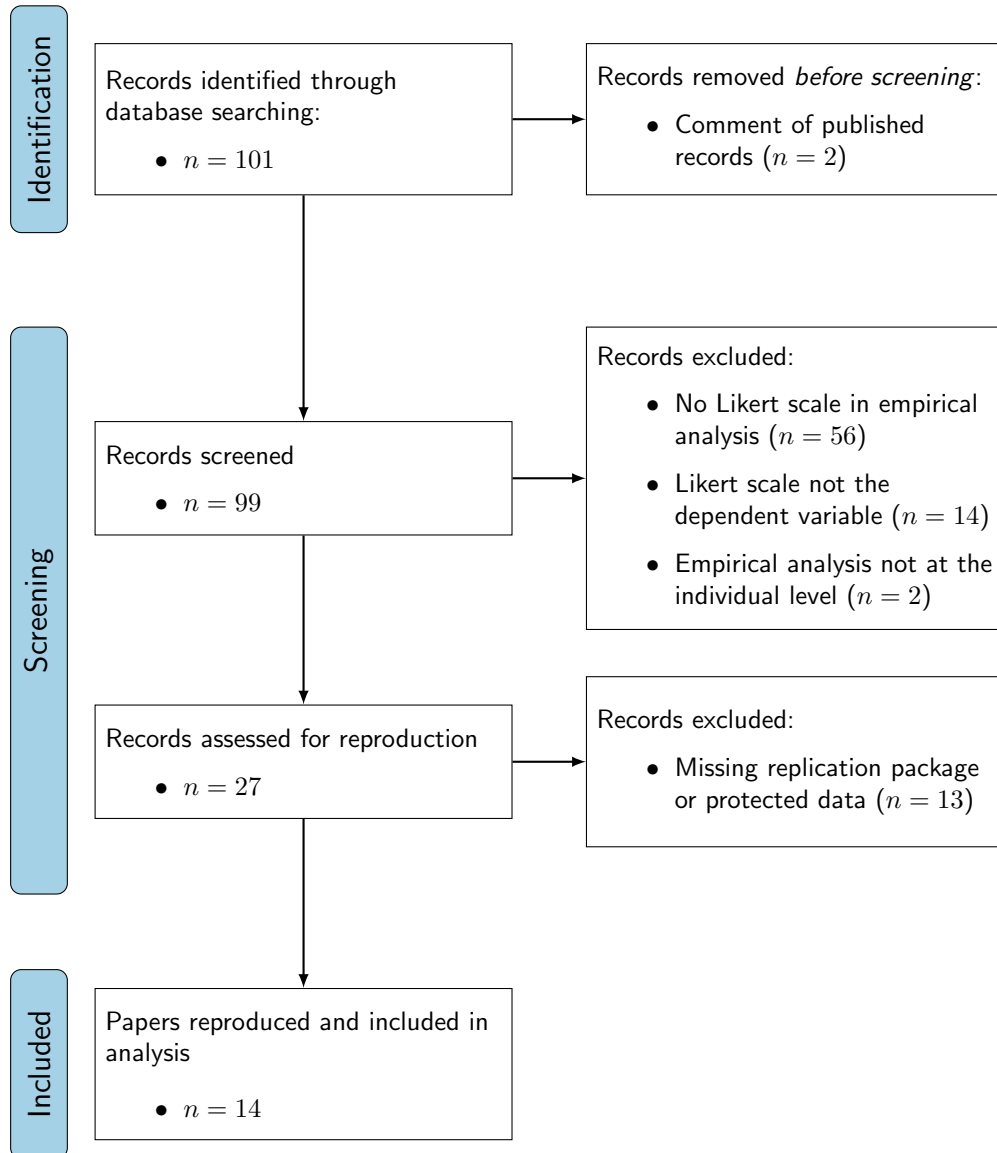
F Additional WellBase Tables and Figures

Table A3: Descriptive Statistics of WellBase at the regression level

	Mean	SD	Min	Max
About the well-being scales:				
<i>Number of response categories:</i>				
3-points scale	0.00		0	1
4-points scale	0.32		0	1
5-points scale	0.06		0	1
6-points scale	0.01		0	1
7-points scale	0.00		0	1
10-points scale	0.30		0	1
11-points scale	0.30		0	1
More than 11-points scale	0.00		0	1
<i>Type of question:</i>				
Life Satisfaction	0.72		0	1
Cantril Ladder	0.04		0	1
Happiness Question	0.24		0	1
About the estimation samples:				
Number of observations	178,759.94	417,261.63	59	2,471,360
Number of observations (logged)	9.70	2.45	4.08	14.72
About the econometric models:				
Number of controls	17.81	17.03	1	191
Individual FE	0.09		0	1
About the independent variables:				
Printed in manuscript	0.10		0	1
Printed in appendix	0.19		0	1
Not printed	0.71		0	1
Continuous variable	0.26		0	1
Time-varying variable	0.71		0	1
Two-stage least square	0.01		0	1
Individual-specific	0.81		0	1
Natural experiment, RCT and policy reform	0.13		0	1
Macroeconomic indicator	0.04		0	1
Absolute t-statistics	6.56	11.95	0.00	124.20
Absolute t-statistics (logged)	0.57	1.08	-7.78	4.66

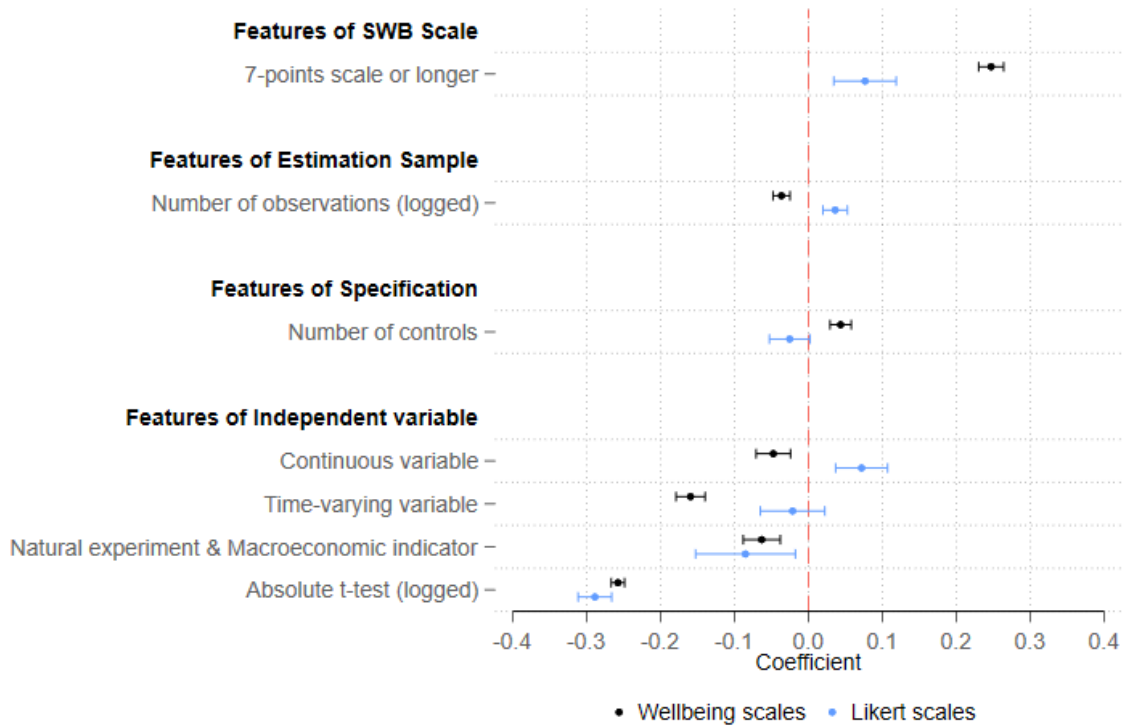
Note: These numbers refer to the sample of 1,601 regressions included in Well-Base.

Figure A19: PRISMA Chart - Likert scales



Note: This PRISMA chart describes the selection of papers included in the Likert scale analysis.

Figure A20: Predictors of the Probability of Sign-reversal - Wellbeing and Likert scales



Notes: The figure shows the coefficients from linear probability models estimating the probability of sign reversal for estimates from well-being and Likert scale regressions published in Top-5 Economics journals. Standard errors are clustered at the regression-paper level. Whiskers represent 95% confidence intervals.

Table A4: Descriptive Statistics of WellBase at the paper level

	Mean	SD	Min	Max
About the well-being scales:				
<i>Number of response categories:</i>				
3-points scale	0.01		0	1
4-points scale	0.13		0	1
5-points scale	0.16		0	1
6-points scale	0.04		0	1
7-points scale	0.05		0	1
10-points scale	0.20		0	1
11-points scale	0.40		0	1
More than 11-points scale	0.01		0	1
<i>Type of question:</i>				
Life Satisfaction	0.73		0	1
Cantril Ladder	0.11		0	1
Happiness Question	0.16		0	1
About the estimation samples:				
Number of observations	123,042.89	382,830.27	84	2,471,360
Number of observations (logged)	8.80	2.32	4.43	14.72
About the econometric models:				
Number of controls	20.30	25.24	1	153.44
Individual FE	0.11		0	1
About the independent variables:				
Printed in manuscript	0.22		0	1
Printed in appendix	0.19		0	1
Not printed	0.58		0	1
Continuous variable	0.27		0	1
Time-varying variable	0.72		0	1
Two-stage least square	0.01		0	1
Individual-specific	0.72		0	1
Natural experiment, RCT and policy reform	0.23		0	1
Macroeconomic indicator	0.03		0	1
Absolute t-statistics	4.69	7.16	0.64	33.10
Absolute t-statistics (logged)	0.47	0.89	-0.94	3.29

Note: These numbers refer to the sample of 72 papers included in Well-Base.

Table A5: Predictors of the Probability and Cost of Sign-reversal: Robustness Checks

	P(Sign-reversal)			Cost of sign-reversal	
	(1)	(2)	(3)	(4)	(5)
About the estimation sample:					
Number of observations (logged)	0.027*** (0.006)	0.019** (0.009)	0.048*** (0.006)	-0.029*** (0.003)	-0.026*** (0.004)
	(0.017)	(0.040)	(0.012)	(0.007)	(0.013)
About the econometric model:					
Number of controls	-0.000 (0.009)	-0.015 (0.014)	0.004 (0.006)	-0.004 (0.003)	-0.003 (0.006)
Individual FE	-0.001 (0.013)	0.120*** (0.035)	-0.000 (0.015)	-0.014** (0.007)	0.077* (0.041)
About the independent variable:					
Continuous variable	-0.032*** (0.007)	-0.030*** (0.007)	-0.034*** (0.007)	0.006 (0.005)	0.018*** (0.005)
Time-varying variable	-0.078*** (0.008)	-0.084*** (0.008)	-0.068*** (0.007)	0.056*** (0.004)	0.066*** (0.004)
Two-stage least squares	-0.033 (0.032)	-0.051 (0.031)	-0.035 (0.029)	0.021 (0.018)	0.036** (0.015)
Natural experiment	-0.059*** (0.014)	-0.082*** (0.014)	-0.037*** (0.010)	0.034*** (0.010)	0.050*** (0.010)
Macroeconomic indicator	-0.066*** (0.015)	-0.035** (0.014)	-0.062*** (0.012)	0.037*** (0.008)	0.015 (0.010)
Absolute t-statistics (logged)	-0.285*** (0.004)	-0.281*** (0.004)	-0.315*** (0.004)	0.193*** (0.002)	0.196*** (0.003)
Observations	28513	28513	28513	17240	17240
Journal FE	✓	.	.	✓	.
Paper FE	.	✓	.	.	✓

Notes: Columns (3) reports marginal effects from probit models while the other Columns report OLS coefficients. All regressions control for a dummy equal to one for wellbeing scales including at least seven categories, a categorical variable indicating whether the wellbeing measure is a life-satisfaction, Cantril Ladder or happiness question. Standard errors are clustered at the regression-paper level. Statistical significance is denoted as follows: * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Table A6: Descriptive Statistics per Type of Scale

	Wellbeing	Political Pref.	Trust	Risk	Other
7-points scale or longer	0.666 [0.472]	0.108*** [0.310]	0.657 [0.475]	1.000*** [0.000]	0.852 [0.355]
Number of observations (logged)	9.371 [2.178]	7.681*** [1.056]	7.631*** [1.114]	6.614*** [0.025]	7.581*** [1.301]
Number of controls	17.567 [6.517]	16.852*** [1.831]	12.545*** [17.044]	18.136 [7.187]	113.19*** [104.04]
Continuous variable	0.208 [0.406]	0.159*** [0.365]	0.573*** [0.495]	0.568*** [0.501]	0.182* [0.386]
Time-varying variable	0.573 [0.495]	0.538** [0.499]	0.579 [0.494]	0.727* [0.451]	0.130*** [0.337]
Natural experiment and Macroeconomic indicator	0.080 [0.272]	0.104** [0.305]	0.157*** [0.364]	0.000*** [0.000]	0.738*** [0.440]
Absolute t-test (logged)	0.237 [1.405]	0.354** [1.267]	0.600*** [1.407]	0.010 [1.178]	0.378*** [1.480]
Observations	10,186	1,501	492	44	866

Notes: These numbers refer to the samples of estimates based on wellbeing and other Likert scales published in Top-5 Economics journals. Standard deviations are in square brackets. Statistical significance, referring to differences between each column and the first column, is denoted as follows: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

G Additional tables on primary and secondary data

Table A7: Description of Primary Datasets

Short Name	Country	Time	Measure of r	Notes	Reference
Prolific	UK	2024	“Overall, how satisfied are you with your life nowadays?”	The discrete measure has 11 response options and mirrors the questions used in the UK APS. Continuous measure constructed by asking respondents about their location within a given discrete response option. Sample obtained via Prolific, with the nationally representative option.	Kaiser and Lepinteur (2025)
Benjamin et al.	USA	2022	Discrete measure is Cantril’s ladder of life (11 response options). Continuous measure asked: “How satisfied you are with your life?”	Continuous and discrete measure obtained with two questions in the same survey. Sample obtained via MTurk.	Benjamin et al. (2023b)
Prati & Kaiser	UK	2023-2024	“All things considered, how satisfied are you with your life nowadays?”	The discrete measure has 7 response options and mirrors the question used in the UKHLS. Continuous and discrete measure obtained with two questions in the same survey. Sample obtained via Prolific	Kaiser and Prati (2025)
LISS	NL	2011	“Taking all things together, how happy would you say you are?”	The discrete measure has 10 response options. In both measures, extremes are labelled “completely unhappy” and “completely happy”. Continuous and discrete measures obtained via two surveys administered one month apart. Sample based on long-standing https://www.lissdata.nl/ panel.	Studer (2012). Also used in Kaiser and Vendrik (2023)

Note: Description of datasets used in section 3 and Appendix C.

Table A8: Descriptive Statistics for Prolific data

	N	Mean	SD	Min	Max
Satisfaction measure					
Life satisfaction (discrete)	1238	6.28	2.07	0.00	10.00
LS (discrete unprompted)	621	6.38	1.97	0.00	10.00
LS (discrete linear prompt)	617	6.18	2.16	0.00	10.00
Life satisfaction (continuous)	1216	6.42	2.07	0.00	10.00
LS (continuous unprompted)	613	6.49	2.05	0.00	10.00
LS (continuous linear prompt)	603	6.35	2.08	0.20	10.00
Height & weight					
Height(cm)	1185	171.13	10.37	129.69	198.12
Weight(kg)	1186	81.60	24.83	40.82	192.32
Slider values					
Slider 1	606	1.07	0.84	0.00	8.60
Slider 2	606	1.95	1.07	0.00	8.60
Slider 3	606	2.85	1.20	0.40	8.60
Slider 4	606	3.86	1.16	0.70	8.70
Slider 5	606	4.94	1.10	1.20	8.90
Slider 6	606	6.03	1.18	1.30	9.30
Slider 7	606	7.05	1.24	1.30	10.00
Slider 8	606	7.98	1.05	4.30	10.00
Slider 9	606	8.94	0.69	6.60	10.00
Demographics					
Ln(Income)	1144	7.30	0.79	4.61	9.39
Unemployed	1243	0.96	0.20	0.00	1.00
Age	1211	46.82	15.82	18.00	87.00
Age Squared	1211	2442.55	1482.51	324.00	7569.00
Has partner	1243	0.65	0.48	0.00	1.00
Higher education	1243	0.54	0.50	0.00	1.00
Non-White	1243	0.18	0.38	0.00	1.00
Female	1199	0.51	0.50	0.00	1.00
Household Size	1214	2.64	1.25	1.00	8.00
Has Children	1243	0.26	0.44	0.00	1.00
Homeowner	1243	0.64	0.48	0.00	1.00

Note: Descriptive statistics for main Prolific data used in section 3 and Appendix C.

Table A9: Descriptive Statistics for Benjamin et al. data

	N	Mean	SD	Min	Max
Satisfaction measure					
Life satisfaction (discrete)	1494	6.69	2.28	0.00	10.00
Life satisfaction (continuous)	1494	6.61	2.54	0.00	10.00
Demographics					
Ln(Income)	1492	10.92	0.78	8.52	13.17
Unemployed	1471	0.10	0.30	0.00	1.00
Age	1493	45.95	12.79	21.62	83.62
Age Squared	1493	2274.42	1272.89	467.36	6992.06
Has partner	1494	0.50	0.50	0.00	1.00
Higher education	1484	0.64	0.48	0.00	1.00
Non-White	1494	0.27	0.44	0.00	1.00
Female	1494	0.55	0.50	0.00	1.00
Household Size	1494	2.79	1.62	1.00	12.00
Has Children	1493	0.35	0.48	0.00	1.00
Homeowner	1494	0.00	0.00	0.00	0.00

Note: Descriptive statistics for data from Benjamin et al. used in section 3 and Appendix C.

Table A10: Descriptive Statistics for Prati & Kaiser data

	N	Mean	SD	Min	Max
Satisfaction measure					
Life satisfaction (discrete)	1931	5.90	2.11	0.00	10.00
Life satisfaction (continuous)	1928	6.60	2.05	0.00	10.00
Demographics					
Ln(Income)	1926	7.33	0.81	5.52	9.10
Unemployed	1926	0.09	0.29	0.00	1.00
Age	1915	41.19	13.03	18.00	82.00
Age Squared	1915	1866.44	1183.20	324.00	6724.00
Has partner	1948	0.75	0.44	0.00	1.00
Higher education	1948	0.56	0.50	0.00	1.00
Non-White	1948	0.12	0.33	0.00	1.00
Female	1925	0.50	0.50	0.00	1.00
Household Size	1909	2.96	1.31	1.00	9.00
Has Children	1948	0.48	0.50	0.00	1.00
Homeowner	1948	0.65	0.48	0.00	1.00

Note: Descriptive statistics for data from Prati & Kaiser used in section 3 and Appendix C.

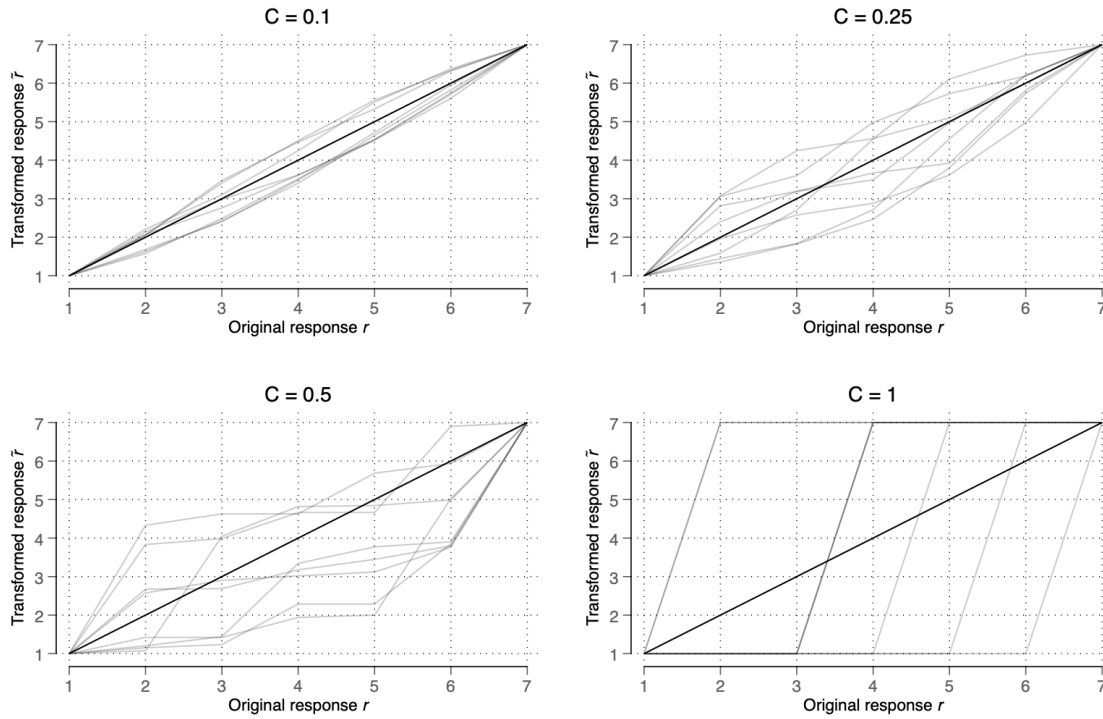
Table A11: Descriptive Statistics for LISS data

	N	Mean	SD	Min	Max
Happiness measure					
Happiness(discrete)	8548	7.17	1.19	0.00	9.00
Happiness(continuous)	8548	6.71	1.50	0.00	9.00
Demographics					
Ln(Income)	7801	7.83	0.52	4.61	12.10
Unemployed	8548	0.05	0.22	0.00	1.00
Age	8548	51.19	17.26	16.00	97.00
Age Squared	8548	2918.30	1717.87	256.00	9409.00
Has partner	8548	0.59	0.49	0.00	1.00
Higher education	8527	0.29	0.46	0.00	1.00
Non-White	8548	0.00	0.00	0.00	0.00
Female	8548	0.53	0.50	0.00	1.00
Household Size	8548	2.56	1.31	1.00	8.00
Has Children	8548	0.38	0.49	0.00	1.00
Homeowner	8548	0.65	0.48	0.00	1.00

Note: Descriptive statistics for data from LISS used in section 3 and Appendix C.

H Additional figures

Figure A21: Examples of different values for $C_{\alpha=2}$ (with 7 response options)



Note: Each line represents a possible transformation from r to \tilde{r} that satisfies a given cost C (with 7 response options).