Working
Paper
Series

doi.org/10.5287/ora-vyq7o5eqb

2503

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

Caspar Kaiser and Anthony Lepinteur

September 2025

cite this paper

Kaiser, C. and Lepinteur, A. (2025). *Measuring the Unmeasurable? Systematic Evidence on Scale Transformations in Subjective Survey Data.* University of Oxford Wellbeing Research Centre Working Paper 2503. doi.org/10.5287/ora-vyq7o5eqb

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

Caspar Kaiser*
Anthony Lepinteur†

September 1, 2025

Abstract

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 a systematic assessment of this linearity assumption, developing a general framework to quantify how easily empirical results can be overturned when it is relaxed. Using original experimental data, we show that respondents use survey scales in ways that deviate from linearity, but only mildly so. Focusing on wellbeing research, we then replicate 40,000+ 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 sizes, which are crucial for policy applications, are unreliable even under these modest non-linearities.

JEL Codes: I31, C18, C87
Key words: Likert scales, ordinal scales, wellbeing, life satisfaction, survey methods

*Warwick Business School, University of Warwick; Wellbeing Research Centre, University of Oxford.
†Department of Behavioural and Cognitive Science, University of Luxembourg; IZA Institute of Labour Economics, Bonn.
We thank Dan Benjamin, Anita Braga, Matthew Cashman, Andrew Clark, Elena Fumagalli, Leonard Goff, Ori Heffetz, Martijn Hendricks, Richard Heys, Christos Makridis, Giorgia Menta, Andrew Oswald, Kelsey O'Connor, 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, Groningen and Alcala, Warwick Business School, DIW, Freie Universität Berlin, General Conference of ISQOLS Rotterdam, Regional Conference of ISQOLS Johannesburg, General Conference IARIW 2024, Nanyang Technological University, Measuring Progress Workshop (STATEC), for their helpful comments and suggestions on earlier drafts of the paper. All errors remain our own. This project received ethical approval from the University of Warwick's Humanities and Social Sciences Research Ethics Committee (Ref.: 228/23-24).

1 Introduction

Ordered response scales, or 'Likert scales', are a standard 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 behavioural data.

Yet scepticism over the validity and use of such Likert scale measures remains. 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'). See, for example, Bertrand and Mullainathan (2001) or Benjamin et al. (2023a). The second concern asks whether responses are comparable across people and 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), Benjamin et al. (2023b) 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 this 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 issue applies to any construct measured with Likert scales, much of the methodological work focused on wellbeing. This is unsurprising: Economists have studied wellbeing, life satisfaction, and happiness for over fifty years (e.g. Easterlin (1974) or Van Praag (1971)). The modern study of wellbeing began in the 1990s, linking it to income, unemployment and macroeconomic conditions (Clark and Oswald 1994; Oswald 1997; Blanchflower and Oswald 2004). Today, wellbeing scales inform government policy, as seen in the UK Treasury's 2022 *Green Book* (UK HMRC Treasury 2021).

Within that literature, Ferrer-i Carbonell and Frijters (2004) were among the first to address the linearity concern. They showed that coefficients estimated from an ordered logit or probit models are similar to those 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 coefficient signs, Schröder and Yitzhaki (2017) provided conditions under which single-covariate regression results can be sign-reversed when allowing for a non-linear reporting function. They also showed that such sign reversals can indeed occur in practice; as did Bloem (2022) who broadened the analysis to a wider class of non-linear functions.

Bond and Lang (2019) generalised these ideas. They demonstrated that virtually all empirical findings based on Likert scales can be reversed via some 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.[1] 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 condition under which coefficients in OLS regressions with multiple covariates are reversible and applied this condition to a selected set of covariates.[2]

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. Even if 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 natural interpretation, with $C = 0$ indicating linear scale use, and $C = 1$ indicating (in a certain sense) 'maximally' non-linear scale use. We formally show that our specific cost function is a member of a broader class of measures that satisfy a series of natural desiderata. Using this cost function we can numerically determine the 'least non-linear' transformation capable of reversing regression results in terms of sign, significance, and relative magnitudes. From a partial identification perspective (e.g. Tamer 2010; Molinari 2020), this approach can be viewed as providing bounds to what would otherwise be impractically wide identified

[1] Liu and Netzer (2023) propose using survey response times to overcome the identification problem raised by Bond and Lang (2019). Their approach exploits *chronometric effects*: decisions tend to be faster when the latent state is further from the reporting threshold. They show, both theoretically and empirically, that response times thereby contain information about the distribution of the latent variable within categories, thereby relaxing the assumptions needed in standard ordered response models.

[2] These papers all focus on how a potentially non-linear reporting function may affect estimates of the conditional mean of underlying wellbeing. Rankings of the conditional median, in contrast, are invariant to such non-linear transformations (Chen et al. 2022; Bloem and Oswald 2022).

parameter sets when treating response scales as merely ordinal. The corresponding statistical machinery is general and applies to *any* bounded ordered scale. We provide corresponding Stata routines on Github.

Using new experimental data, we subsequently 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 73 papers, 1,610 regressions, with 3,430 coefficients of interest (and 28,522 coefficients overall). Using this dataset, we systematically assess the vulnerability of the published literature.

Our results show that respondents, on average, interpret and use wellbeing scales in a manner that does deviate from linearity, but only mildly so. Our upper bound estimate of this deviation 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 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%. The probability that a given estimate can be sign-reversed is systematically related to identifiable features of research design. Certain design choices, like leveraging natural experiments, are associated with substantially lower risks. 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. While coefficient signs for these determinants 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.

Our findings generalise beyond wellbeing scales. To show this, we reproduce 16 papers (23,104 coefficients) published in top-five economics journal. Each of these use Likert-scales to measure e.g. risk aversion, social trust, or political preferences. The prevalence and predictors of sign reversals for these measures closely mirror our wellbeing results.

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 presents our results based on it. Section 5 concludes. The ap-

pendices provide proofs, additional discussion, and further results. Replication codes are available here.

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.

Throughout, we primarily focus on the behaviour of OLS estimators under monotonic transformations of the response scale.[3] This is because our systematic replication exercise in Section 4 shows that the published economics literature overwhelmingly applies OLS to such scales. The central question, thus, is how robust this practice is.

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: $r_i \in \{1, 2, ..., k, ..., K\}$. We also observe a vector of covariates $\mathbf{X}_i$.

These responses measure an underlying but unobservable state s_i.[4] Suppose 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.

We are concerned with estimates from OLS regressions of $\tilde{r}_i$ on $\mathbf{X}_i$:

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

[3] It might instead be natural to analyse ordered response scales through threshold-crossing models (Klein and Sherman 2002). However, as shown by Bond and Lang (2019), very similar concerns apply to ordered probit/logit models.

[4] For 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.

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, only 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.

2.2 Sign reversals

If the sign of $\hat{\beta}_m^{(\tilde{r})}$ does not change under any positive monotonic transformation of the dependent variable r_i, then how we would code survey responses would not affect our estimates of the sign of their association with X_{im}. Define a new variable $d_{ki} \equiv \mathbb{1}(r_i \leq k)$ that dichotomises r_i at every response category. The following proposition then holds:

Proposition 1 (Non-reversal condition)**.** *The sign of $\hat{\beta}_m^{(\tilde{r})}$ 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 $k = 1, ..., K-1$.*

The proof in our notation appears in Appendix A.1. This condition can be read as establishing whether first-order stochastic dominance of r_i with respect to some variable X_i holds. As we show in Appendix C.1 this result extends to continuous outcomes, fixed effects, and two-stage least squares (2SLS) 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 dichotomisations, 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 merely a statement about the behaviour of OLS regression coefficients. To connect estimates $\hat{\beta}_m^{(\tilde{r})}$ from regressions of $\tilde{r}_i$ to underlying states s_i, we must make two 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* $\mathrm{E}(\varepsilon_i X_i) = 0$.

This is 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* $\mathrm{E}(\vartheta_i X_i) = 0$. *For coefficient γ_m corresponding to X_{im}, either* $\mathrm{sgn}(\beta_m) \neq \mathrm{sgn}(\gamma_m)$ *or* $\mathrm{sgn}(\beta_m) = \mathrm{sgn}(\gamma_m)$ *and* $|\beta_m| > |\gamma_m|$.

We can think of ζ_i as a measurement error associated with discretising 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 E 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 $\hat{\beta}_m^{(\tilde{r})}$ 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 discretising 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).

[5] To gain some intuition on this, 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. $\beta_m < 0$). 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.

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 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 size of a variable's association with the the outcome of interest.

However, such ratios are affected by non-linear transformations. The only exception occurs when the corresponding ratios $\hat{\beta}_{km}^{(d)}/\hat{\beta}_{kn}^{(d)}$ from regressions of dichotomised variables d_{ki} are constant across all k. Empirically, this is never the case. Yet, 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 $k = 1, \ldots, K-1$.*

See Appendix A.3 for the proof. Unfortunately, these bounds tend 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 behaviour of coefficient estimates 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 represents at best an *unusual* assumption about how people use survey scales. We thus 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.

There are several desiderata that any measure of departure from linear scale-use should possess. First, it should be minimised when gaps between categories are uniform (implying linear scale use) and maximised when all the scale's range is concentrated in a single jump between two categories. Second, greater dispersion in gap sizes should generally increase the value of the measure, i.e. increasingly unequal gaps should correspond to greater departures from linearity. Third, the measure should treat all positions on the scale equally.

A (say) compression between categories 2-3 should be just as 'costly' as the same compression between categories 8-9. In Appendix B we formalize these desiderata and derive a representation theorem characterising the class of functions satisfying them.

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 $f(l_k) = \tilde{l}_k$. Finally, let $\Delta\tilde{\mathbf{l}} \equiv [\tilde{l}_2 - \tilde{l}_1, \tilde{l}_3 - \tilde{l}_2, ..., \tilde{l}_K - \tilde{l}_{K-1}]$ capture these differences.

With this notation in place, we will now focus on a particular member of the class of functions we characterise in Appendix B. Specifically, we consider:

$$C_\alpha(\tilde{\mathbf{l}}) = \left(\frac{\text{Var}(\Delta\tilde{\mathbf{l}})}{\text{maxVar}(\Delta\tilde{\mathbf{l}})} \right)^{1/\alpha},$$

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

Any $\alpha > 0$ yields a valid 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).[6] Generally, smaller values for α make the cost function more lenient, allowing for stronger non-linearities at lower cost values. In Appendix B.4 we show that setting $\alpha = 2$ renders our cost function linearly homogeneous. Specifically, if we take some non-linear scale use characterised by $\Delta\tilde{\mathbf{l}}$ and 'mix' it with linear scale use $\mathbf{u} = [(l_K - l_1)/(K-1), ..., (l_K - l_1)/(K-1)]$ with weight λ to obtain $\Delta\tilde{\mathbf{l}}^{(\lambda)} = \lambda\Delta\tilde{\mathbf{l}} + (1-\lambda)\mathbf{u}$, we get $C(\Delta\tilde{\mathbf{l}}^{(\lambda)}) = \lambda \cdot C(\Delta\tilde{\mathbf{l}})$. This property yields a natural interpretation: a transformation with $C = 0.1$ is, in this sense, '10% non-linear'. We therefore use $\alpha = 2$ in the empirical sections and drop the α subscript.[7]

We 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 optimisation problem are shared across all applications:

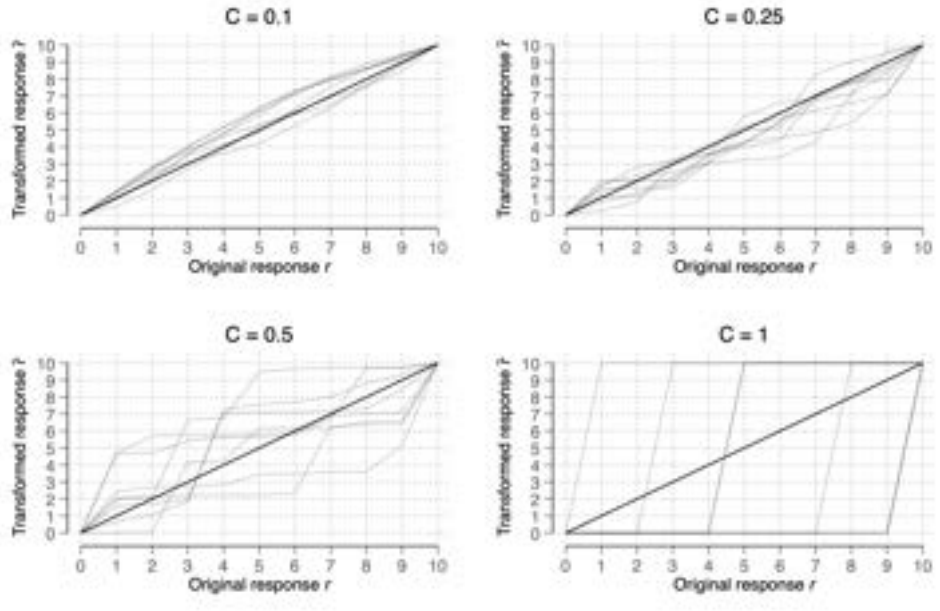
1. Normalisation: The 'length' of the scale must be preserved: $l_K - l_1 = \tilde{l}_K - \tilde{l}_1$.

2. Monotonicity: Transformed labels must be strictly increasing: $\tilde{l}_k - \tilde{l}_{k-1} > 0 \, \forall \, k \geq 2$.

[6] The normalised Theil Index is another member of the class of functions we characterise. Appendix Figure A9 shows results based on that index.

[7] However, 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 is possible to achieve visually strong non-linearities even for small values of C. In Appendix D we show that 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 vastly varying numbers of response options, we there propose setting $\alpha = 2\log_{10}(K-1)$.

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 $C_{\alpha=2} = 1$, 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.

Here, the **Normalisation** constraint ensures that transformations preserve the overall range of the scale.[8] This prevents arbitrary stretching or compression that would make 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: $\text{sgn}\big(\hat{\beta}_m^{(\tilde{r})}\big) \neq \text{sgn}\big(\hat{\beta}_m^{(r)}\big)$.

On the other hand, for coefficient ratios, we should constrain ourselves to achieving some target ratio within the bounds identified by Proposition 3:

[8]Some papers study potential stretching of the scale across respondents while maintaining the linearity assumption. See e.g. Benjamin et al. (2023b) or Fabian (2022). A fruitful avenue for future work is to combine these research streams.

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 the next section. For any application, we then find:

$$\tilde{\mathbf{l}}^* = \text{argmin}_{\tilde{\mathbf{l}}}\, C_\alpha(\tilde{\mathbf{l}}), \tag{2}$$

subject to the relevant constraints. In general, there may not be a unique solution to this optimisation problem. However, for any solution, $C_\alpha(\tilde{\mathbf{l}}^*)$ 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

To assess how significance levels change under monotonic transformations, we need the variance-covariance matrix of $\hat{\boldsymbol{\beta}}^{(\tilde{r})}$ from regressions of any transformed variable $\tilde{r}_i$. The variance-covariance matrix takes the standard form:

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

where $\hat{\boldsymbol{\Omega}}$ is an estimate of the covariance matrix of the residuals. The form of $\hat{\boldsymbol{\Omega}}$ 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 dichotomised regressions of d_{ki}. As shown in Appendix A.4 we have:

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

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

This decomposition allows us to compute $\hat{\boldsymbol{\Omega}}$ for any transformation using only results from the $K-1$ dichotomised regressions of d_{ki}. Appendix A.4 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.[9] We can do so by numerically maximising (for an upper

[9]Note here that is <u>not</u> the case that the p-value $p(\hat{\beta}_m^{(\tilde{r})})$ associated with some estimated coefficient $\hat{\beta}_m^{(\tilde{r})}$ is

bound) or numerically minimising (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 optimisation problem we previously specified:

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

We then numerically solve the optimisation problem of Equation 2 subject to constraints (**1**)-(**3c**). By choosing π appropriately (e.g. $\pi = 0.05$), this allows us to asses 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?

The previous section was 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.

There is little direct evidence on how respondents use survey response options. Existing work is mostly indirect. Psychophysics studies on how people interpret numbers suggest that, for bounded intervals analogous to survey scales, subjective and objective values are roughly linear (Banks and Coleman 1981; Banks and Hill 1974; Schneider et al. 1974). Earlier contributions in economics also point to near-linearity in scale use (Van Praag and Van der Sar 1988; Van Praag 1991; Van Praag et al. 1999). More recently, Kaiser and Oswald (2022) show that reported satisfaction in domains such as jobs, housing, and health predicts subsequent quitting actions in a near-linear fashion.

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.

3.1 Data and approach

We rely on 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 UK. See Appendix Table A4 for further details on data collection and Appendix Table A5 for descriptive statistics. We implement four different methods to estimate C. Given that our primary interest in our replication effort of Section 4 is on 'life satisfaction', our attempts at bounding C also tend to be specific to life satisfaction. As will become apparent, these

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 $p(\hat{\beta}_{km}^{(d)}) < 1 \; \forall \; k$.

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 randomised participants into two conditions. One half of participants is given a standard life satisfaction question: '*Overall, how satisfied are you with your life nowadays?*'.[10] 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. Our full survey, showing how these questions were presented to respondents, is available here.

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 adhere to our linearity prompt. Second, given randomisation, we assume that the distribution of underlying satisfaction is the same across both groups. For every value $r_{un}^{(disc)} \in \{0, 1, ..., k, ..., K\}$ of the unprompted discrete satisfaction data, we find the 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 denote the cumulative distribution functions of $r_{un}^{(disc)}$ and $r_{lin}^{(cont)}$. If scale use was 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 $r_{lin}^{*(cont)}$. 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 link). Towards the end of the survey – after all subjective questions are answered – we then ask respondents about their actual 'objective' 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. We can then read off in how far, expressed in

[10] This follows the phrasing used by the UK's Office of National Statistics in the Annual Population Survey.

terms of our cost C, respondents' average scale use deviated from linearity.

3.1.3 Interactive sliders

For our fourth method we interactively ask respondents how they interpreted the scale (this sample is restricted to respondents that were not 'prompted' in Section 3.1.1 – see this link for an interactive demo). We first explain to respondents that scale use might be non-linear. Then, as a comprehension check, we ask respondents to graphically indicate, using a set of interactive sliders, a pre-specified type of non-linear scale use (specifically a case in which the difference between a '3' and a '4' is larger than the difference between a '7' and an '8'). We only proceed with respondents who pass this check (82%). In the final step, we ask respondents to indicate their own scale use with the same set of interactive sliders. We provide respondents with several presets (incl. concave, convex, logistic, and inverse logistic scale use). When 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.

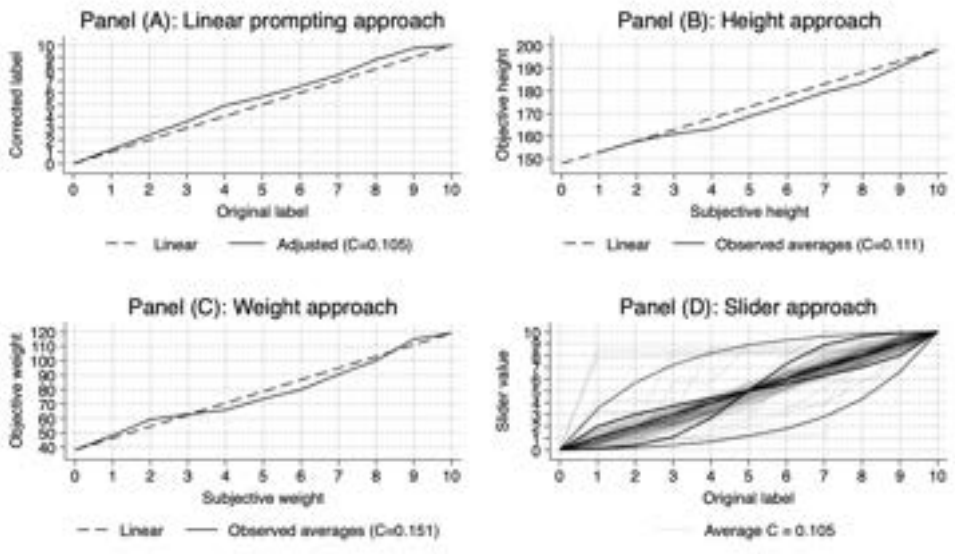
3.2 Results on scale use

Results are displayed in Figure 2. In each panel, the horizontal 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). 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 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 not using any of the pre-set options. Taking the average C across 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

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



Notes: Four different approaches to measuring non-linear scale use all point to similar conclusions. Panel (A) is based on a randomised experiment: the first half of our sample answered a standard satisfaction question, while the second half received explicit instructions to treat the scale linearly. The solid line shows how we would need to adjust response labels in the first group to match the distribution of the second 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≈1,200 UK residents recruited via Prolific.

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 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 call transformations with $0 \leq C < 0.15$ "*plausible*" or "*mild*". Transformations with $0.15 \leq C < 0.30$ will be called "*conservatively-plausible*". Transformations with $0.30 \leq C \leq 1.00$ will be 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 discretising 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 discretising 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.

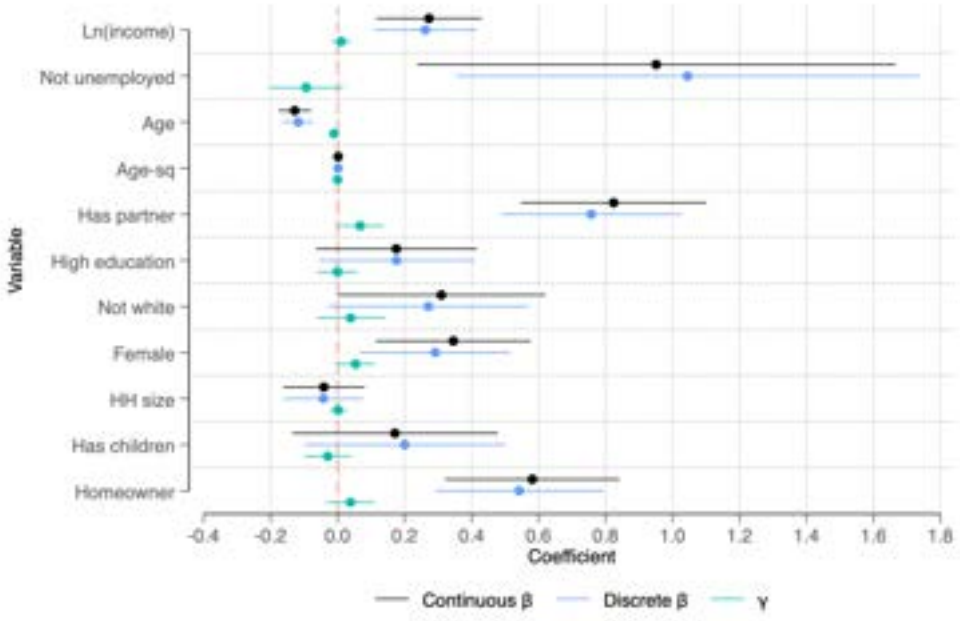
With this data, we evaluate what γ_m (which is key to Assumption 2) would be for each covariate m if scale use were linear. We rely on the following argument: The coefficient γ_m is intended to capture systematic within-category heterogeneity in underlying satisfaction.[11] When satisfaction is measured 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 scale use is linear, we may consider a (quasi-)continuous measurement of satisfaction r_i^{cont} to be a proxy for underlying satisfaction, i.e., $r_i^{cont} \approx s_i$. On this basis, we estimate two regressions: one using discrete r_i^{disc} and another using continuous r_i^{cont}. The difference in estimated coefficients, $\hat{\beta}_m^{cont} - \hat{\beta}_m^{disc}$, gives us an estimate of γ_m.[12]

Figure 3 presents our results. The figure shows estimates $\hat{\beta}_m^{cont}$, $\hat{\beta}_m^{disc}$, and $\hat{\gamma}_m$ for a set of standard socio-economic characteristics. In all cases, $\hat{\gamma}_m$ is close to zero and statistically insignificant (at the 5% level). For most variables, $\hat{\gamma}_m$ takes the same sign as $\hat{\beta}_m^{cont}$, causing the estimate of β_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. For none of the variables in this analysis

[11] In Assumption 2, $\zeta_i = \mathbf{X}_i\boldsymbol{\gamma} + \vartheta_i$ represents the measurement error associated with discretising 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}, and therefore do not systematically bias our coefficient estimates beyond the discretisation error we explicitly model.

[12] To see this more formally, note that since we assume $r_i^{cont} \approx s_i$ for $C = 0$, we have $r_i^{cont} = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$. For the discrete measure, we have $r_i^{disc} = \mathbf{X}_i(\boldsymbol{\beta} - \boldsymbol{\gamma}) + \varepsilon_i - \vartheta_i$. 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 discretisation. 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, unemployment strongly reduces satisfaction, higher household income increases it, and having a partner is beneficial. Women report slightly higher satisfaction than men.

does Assumption 2 look to be violated. This is evidence in favour of Assumption 2.

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 or happiness. Additional details and descriptive statistics are given in Appendix Tables A4 and A6. The main methodological difference in these compared to our own data is that answers to the continuous and the discrete question were given at different times in the survey. Thus, respondents were not forced to give their continuous answer as being located within a given discrete category. Results are shown in Appendix Figure A4. 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 violations of Assumption 2 twice: once for an 'education' dummy and once for a gender dummy. Overall, this is thus further evidence in support of Assumption 2.

Finally, although our evidence suggests $\gamma_m \approx 0$ if scale use were linear (i.e., for $C = 0$), it remains unclear how γ_m would behave for 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 $\boldsymbol{\gamma}$ 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 E, 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*

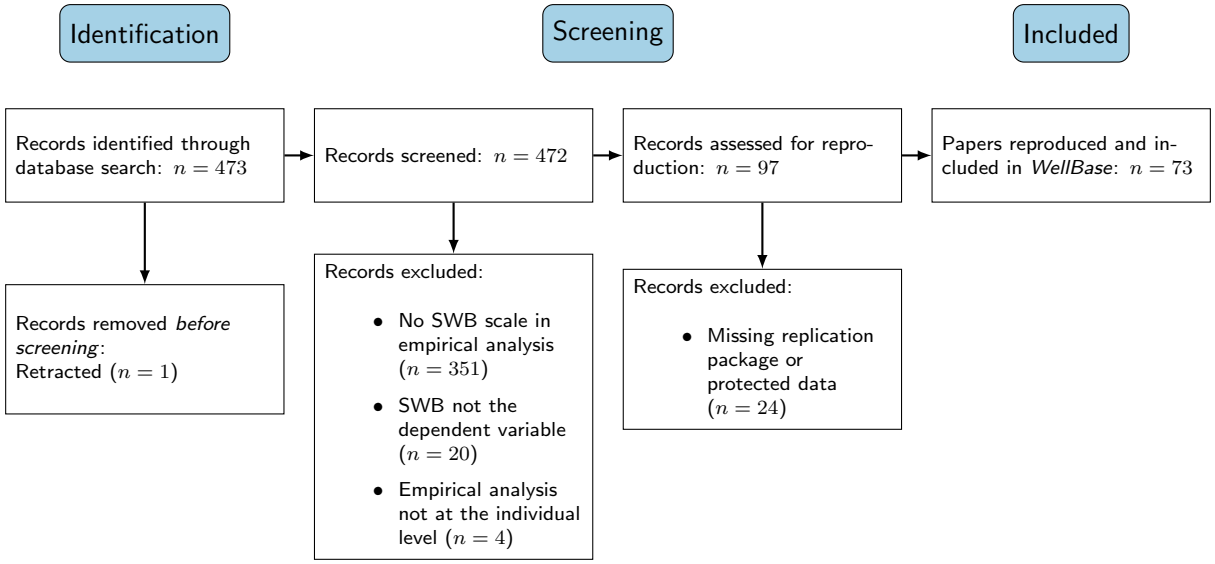
Subjective wellbeing is becoming increasingly central to policy. Among constructs measured using ordered response scales, it also is the main focus of methodological critiques. Drawing on our replication database, WellBase, this section provides the first systematic assessment of the robustness of the empirical economics of subjective wellbeing. *WellBase* includes 73 papers, 1,610 regressions and 28,522 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. There, we reproduce 23,104 coefficients across 16 papers.

4.1 Data

Our goal was to reproduce 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), which typically enforce data and code sharing, making reproduction more feasible. 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, was based on the following keywords: "Life Satisfaction", "Cantril Ladder", "Subjective well-being", and "Subjective wellbeing". The first two capture the most common cognitive wellbeing scales, while we added the latter two to capture any papers using less frequent cognitive wellbeing measures. See Figure 4 for a summary of our selection process.

Figure 4: Selection Process for *WellBase*.



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

In total, 97 articles were eligible for inclusion. Because of missing replication files or protected data, we reproduced 73 of these articles. Among these, we successfully reproduced all of the 1,610 relevant regressions in both the main manuscripts and any associated appendices (printed or online). Less than 2% of these regressions (spread across six 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 articles) used a binary dummy for high wellbeing as dependent variable. We re-estimated these regressions using the underlying full 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 5,322 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.

Table 1 provides an overview of the characteristics of the replicated estimates. 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

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

	Mean	SD	Min	Max
About the wellbeing scales:				
Number of response categories:				
3-point scale	0.00		0	1
4-point scale	0.27		0	1
5-point scale	0.14		0	1
6-point scale	0.00		0	1
7-point scale	0.00		0	1
10-point scale	0.22		0	1
11-point scale	0.36		0	1
More than 11-point scale	0.00		0	1
Type of question:				
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 (logged)	9.98	2.17	4.08	14.72
About the econometric models:				
Number of controls	34.07	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, or policy reform	0.04		0	1
Macroeconomic indicator	0.04		0	1
Absolute t-statistics (logged)	0.45	1.53	-9.08	6.16
Total number of estimates/regressions/papers:				28,522/1,610/73

Note: These numbers refer to the sample of 28,522 estimates included in *WellBase*.

control variables not shown in the printed articles.[13] About 4% relate to quasi-natural experiments (e.g., centralisation reforms in Switzerland, the London Olympics, or RCTs), while another 4% are macroeconomic factors (e.g., economic growth, inflation rates). Approximately 25% of coefficients relate to time-invariant characteristics (e.g., sex). Likewise, 25% of estimates are based on a continuous covariate (e.g., income, age).

Appendix Table A2 focuses on the 27 papers in *WellBase* for which at least half of the

[13]Most 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.

printed regressions use a wellbeing scale as dependent variable. In these studies, the main objective is to uncover the drivers of subjective wellbeing.[14] For each of these, Table A2 summarizes the hypotheses tested, and records the sign and significance of the main coefficients. A large number of these studies find that economic resources (e.g. household income or labour earnings) are associated with higher levels of reported wellbeing. Reported wellbeing also systematically declines following major adverse life events, including physical violence (Johnston et al. 2018), exposure to the Chernobyl disaster (Danzer and Danzer 2016), or falling into poverty (Clark et al. 2016). Several papers examine policy or environmental changes, such as centralisation 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 how sensitive the WellBase estimates are to the 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.

4.2.1 On sign reversals: Documenting the risk of reversal

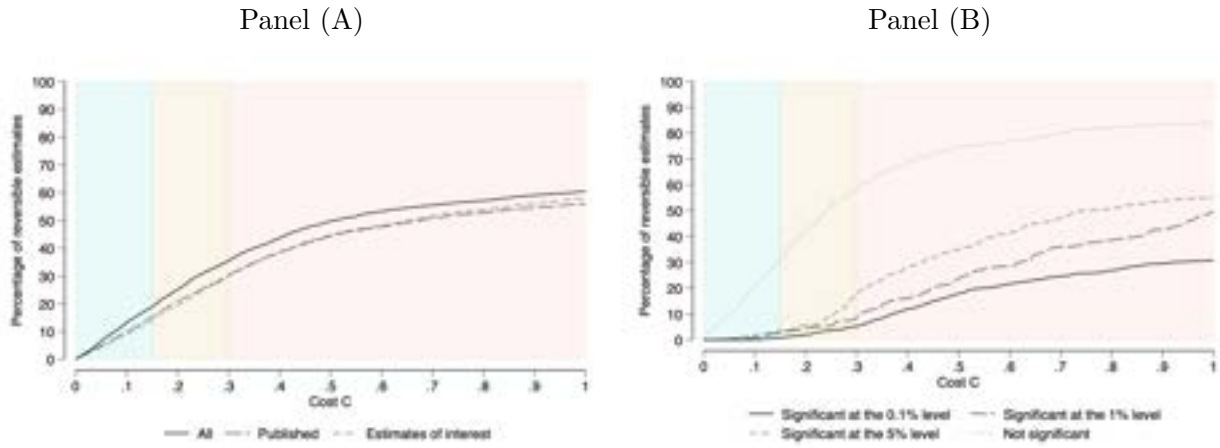
Figure 5 shows the share of point estimates whose sign can be reversed by applying some positive monotonic transformation of the response scale with a cost of at most C.

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.

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

[14] In the remaining papers in *WellBase*, wellbeing is not the *primary* outcome of interest.

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



Notes: Most coefficients signs in published wellbeing research are robust to plausible departures from linear scale use. The figure shows the cumulative shares of coefficients included in *WellBase* for which their sign can be reversed by some positive monotonic transformation of the response scale with at most cost C. The case of $C = 0$ corresponds to assuming that scale-use is linear. 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 replaced by a purely ordinal interpretation. Based on the scale-use evidence presented in Section 3, shaded regions indicate "plausible" (green), "conservatively plausible" (yellow), and "implausible" (red) degrees of non-linearity. Panel (A) shows that at most 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 harder to reverse the sign of coefficients that are originally significant at the 5% level or below.

$C = 0$, 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 a 5% significance threshold.

Appendix Table A2 restricts attention to studies whose main objective is to study the determinants of subjective wellbeing. Its last two columns indicate whether a reversal is possible and, if so, what 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 explore whether the risk of sign reversal can be predicted by observable features of the research design. We estimate a linear probability model of the form:

$$\text{Rev}_{mpr} = \beta_0 + \beta_1 \ln(\#\text{Observations})_{pr} + \boldsymbol{\beta}_2 \mathbf{Model}_{pr} + \boldsymbol{\beta}_3 \mathbf{Estimate}_{mpr} + \boldsymbol{\beta}_4 \mathbf{X}_{pr} + \varepsilon_{mpr}, \quad (3)$$

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.[15]

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, reflecting practices through which researchers attempt to limit omitted variable bias. 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 2SLS. 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 RCTs), 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 categories, and a categorical variable differentiating among life satisfaction questions, the Cantril Ladder, and happiness questions.

We estimate two versions of Equation (3): 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) continues to hold.

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

$$\text{Cost}_{mpr} = \beta_0 + \beta_1 \ln(\#\text{Observations})_{pr} + \boldsymbol{\beta}_2 \mathbf{Model}_{pr} + \boldsymbol{\beta}_3 \mathbf{Estimate}_{mpr} + \boldsymbol{\beta}_4 \mathbf{X}_{pr} + \varepsilon_{mpr}, \quad (4)$$

Equation (4) mirrors Equation (3) but uses the minimum C needed for a sign reversal as the dependent variable. Comparing Equations (3) and (4) enables us to assess whether the probability of reversal and the ease of achieving it share common determinants. In both analyses, we cluster standard errors at the regression–paper ($r \times p$) level. Continuous independent variables are standardised using their means and SDs reported in Table A1.

Table 2 reports predictors of reversal risk in Columns (1) and (2) and reversal costs in

[15] We also estimate a probit model to assess the robustness of our findings. Marginal effects are reported in Table A3. Conclusions are the same.

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***	0.029***	0.016***	-0.038***
	(0.007)	(0.006)	(0.003)	(0.003)
About the econometric model:				
Number of controls	0.019*	0.001	-0.009**	-0.006
	(0.008)	(0.006)	(0.003)	(0.003)
Individual FE	0.084***	0.004	-0.045***	-0.010
	(0.021)	(0.016)	(0.009)	(0.008)
About the independent variable:				
Continuous variable	-0.080***	-0.034***	-0.008	-0.002
	(0.010)	(0.008)	(0.006)	(0.005)
Time-varying variable	-0.142***	-0.074***	0.065***	0.038***
	(0.009)	(0.007)	(0.005)	(0.004)
Two-stage least square	-0.056	-0.012	0.049**	0.016
	(0.033)	(0.032)	(0.018)	(0.017)
Natural experiment	-0.167***	-0.090***	0.091***	0.034***
	(0.017)	(0.014)	(0.013)	(0.009)
Macroeconomic indicator	0.145***	-0.026	-0.009	0.024**
	(0.015)	(0.014)	(0.009)	(0.008)
Absolute t-statistics (logged)		-0.275***		0.187***
		(0.004)		(0.002)
Observations	28,522	28,522	17,243	17,243
R^2	0.163	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 wellbeing 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 wellbeing 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: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Columns (3) and (4). We highlight three 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 the strongest predictor of robustness: estimates with originally larger t-statistics are substantially less prone to reversal and more costly to reverse. This 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 exogenous sources of variation (e.g., natural experiments) are both less likely to reverse and require larger departures from linearity.[16]

Overall, 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. Signs of highly significant results are far more likely to persist across scale transformations.

4.2.3 On significance reversals: Documenting the risk of reversal

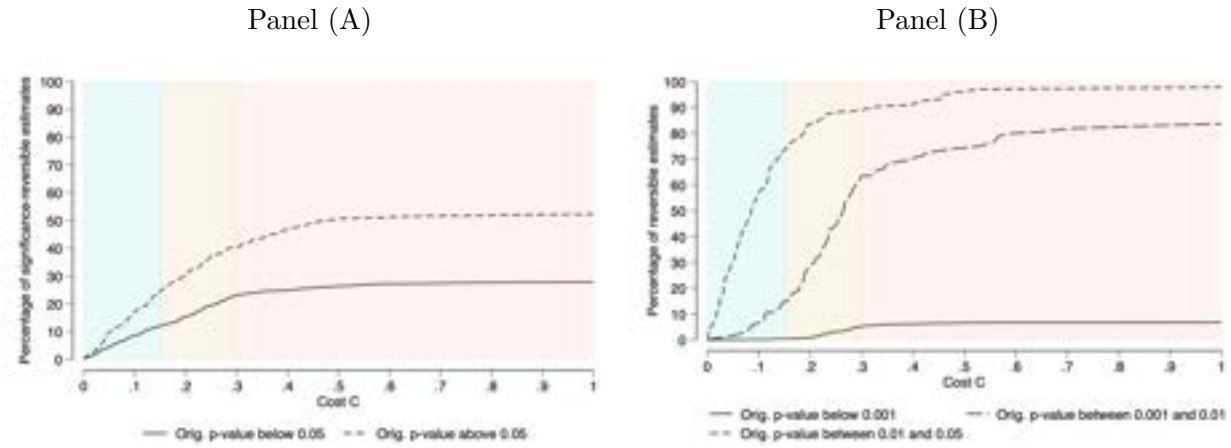
We now quantify the risk of *significance* reversals. To this end, we first divide all *estimates of interest* in *WellBase* into two groups: those initially significant at the 5% level, and those not significant at this level. For all estimates, we compute the maximum and minimum attainable p-values 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 on the possibility 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 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 originally insignificant coefficients 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 originally $0.01 < p \leq 0.05$ 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

[16]Some robustness checks are given Appendix Table A3. There we re-estimate Columns (2) and (4) of Table 2 while adding journal or paper fixed effects, employing a probit model instead of a linear probability model as well as a linear Cragg hurdle model where we jointly model the occurrence and cost of reversibility. Our conclusions are robust across these specifications.

Figure 6: *Significance*-reversal shares for different values of C in WellBase.



Notes: Cumulative shares of *coefficients of interest* included in WellBase 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 lose significance under 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.

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.

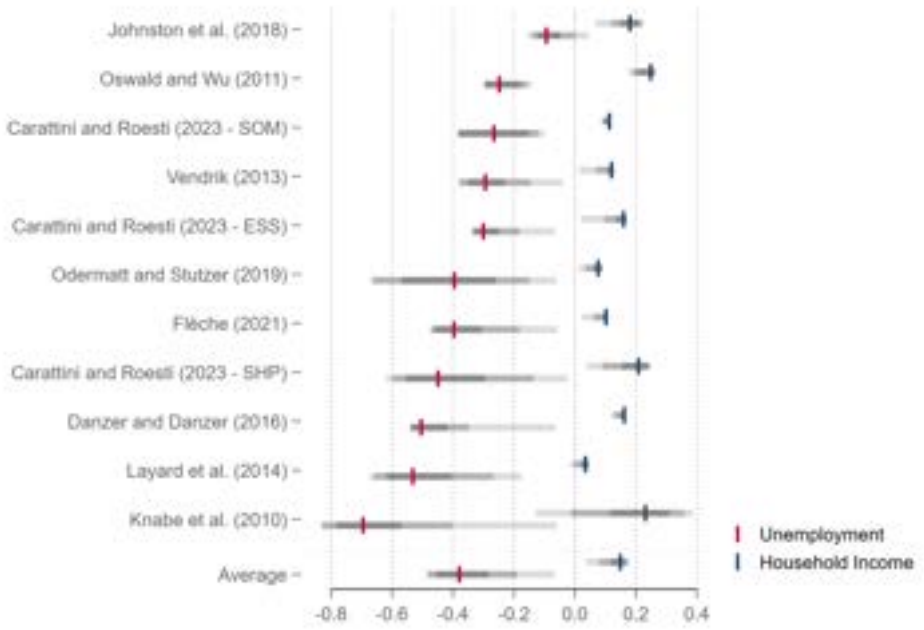
4.2.4 On relative magnitudes: The case of unemployment and income

Turning to relative magnitudes, we now focus on unemployment and the income–wellbeing relationship. The latter 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.[17] 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 0.16 SD increase in wellbeing.

[17]The only exception is Carattini and Roesti (2025), who used three distinct datasets, where we treat each dataset from their paper as a unique observation.

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



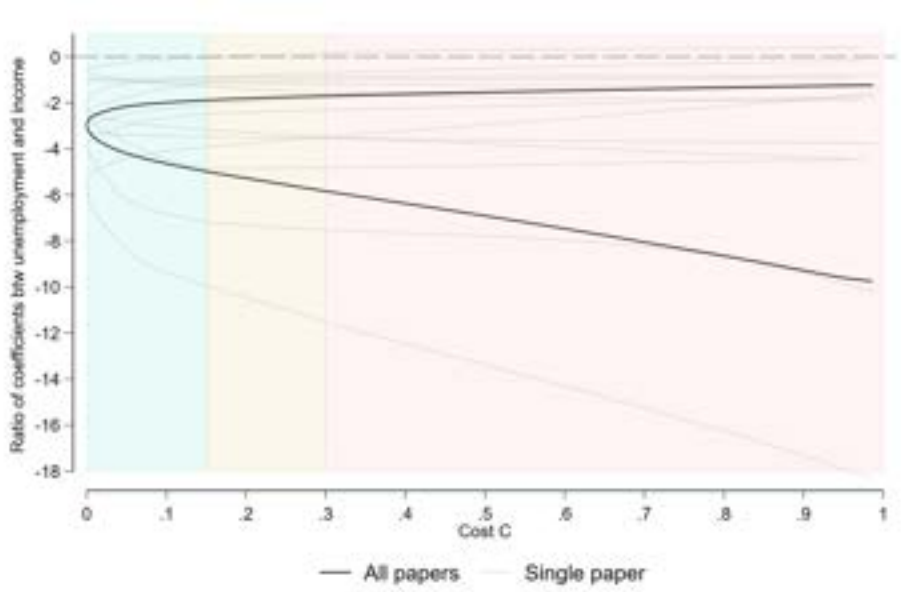
Notes: Standardised average point estimates for unemployment and the log of household income among papers included in *WellBase*. Papers are ranked by the average effect size of the unemployment coefficient. The overall average, weighted by the inverse of the standard error of the individual estimates (Borenstein et al. 2010), is displayed at the bottom. Wellbeing scales are standardised (mean of zero and standard deviation of one). Grey bars indicate the possible range of point estimates after applying positive monotonic transformations of the wellbeing scales. There are three shades of grey, with the darkest corresponding to $C < 0.15$, the middle to $0.15 \leq C < 0.30$, and the lightest to $0.30 \leq C$.

The grey bars in Figure 7 show how these estimates may vary as we depart from linear scale use. When taking the meta-analytic average across all studies, allowing 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 0.17 (0.18) SDs. Thus, the magnitudes of estimates vary widely, even under "plausible" transformations.

However, given that there is no natural absolute scale for wellbeing (linear or not), the absolute magnitudes of coefficients are not entirely 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(HH income) across different levels for C.[18] Each grey line corresponds to a different paper. The

[18]For 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 excluded four regressions on that basis: one in Knabe et al. (2010) and three in Layard et al. (2014) (c.f. Figure 8).

Figure 8: Ranges of unemployment-income ratios.



Notes: Ranges of marginal rate of substitution (MRS) between unemployment and log household income by paper (grey) and their average (black) across values of C. 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. This reveals a wide range: from just positive to –18.25.

black line shows the average ratio across all papers. We observe that this mean MRS can range from just positive to as low as –18.25. Under "likely" transformations, the ranges are only slightly narrower, ranging from zero to -10.

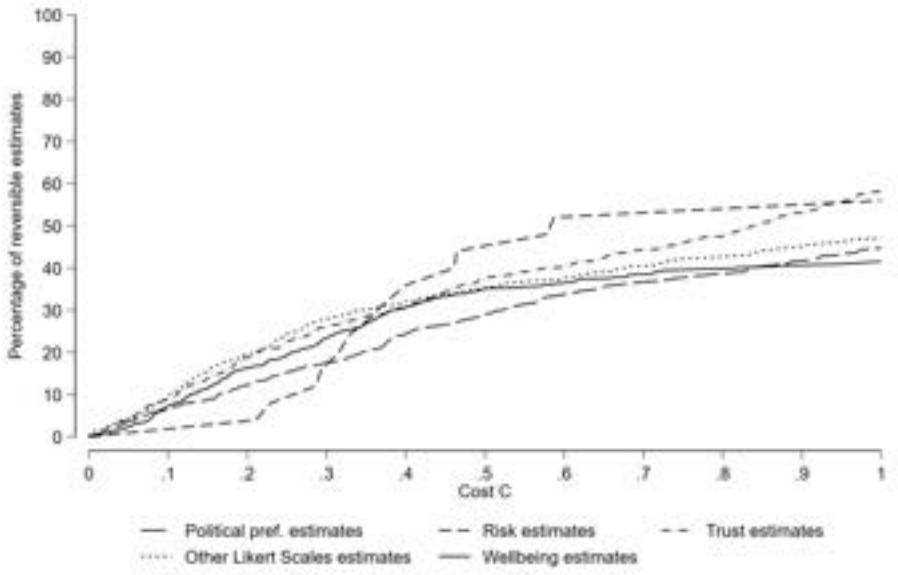
Thus, although the risk of sign and significance reversals appeared relatively small under "plausible" transformations of the wellbeing scale, the same cannot be said about the *magnitudes* of estimates. In the case of unemployment and income, two key drivers of wellbeing, both absolute and relative magnitudes turn out to be highly sensitive to even mild departures from linearity.

4.3 Likert Scales for Attitudes, Preferences and Perceptions

Our focus has so far been on wellbeing scales. But these are not the only constructs in economics measured using ordered 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

Figure 9: Comparing the risk of sign reversal between wellbeing scales and other Likert scales in Top-five 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. The figure shows that the risk of reversal is not unique to wellbeing. In many cases, results on constructs such as risk (56%), trust (58%), political preferences (44%), and 'other' constructs (47%) are sign reversible.

and May 2025 in the five leading economics journals[19] and retained those whose full text contained the term "Likert scale" or whose title included at least one of the following expressions: "attitudes", "risk aversion," "risk preferences," "trust," or "preferences for". This search strategy is unlikely to cover all Likert-scale based research published in top-five economics journals, but assembling a true census of all such published research is beyond the scope of this study.

As shown in Figure A10, we reproduced 16 articles for a total of 511 regressions and 23,104 estimates (8.61% of which are printed coefficients). Of the included papers, three contained contained Likert-scale measure of trust (Acemoglu et al. 2020; Algan and Cahuc 2010; Falk et al. 2018), four contained a Likert-scale measure of political preferences (Kuziemko et al. 2015; Alesina et al. 2018; Jha and Shayo 2019; Dechezleprêtre et al. 2025), two contained a Likert-scale measure of risk aversion (Dohmen et al. 2010; Jha and Shayo 2019), and nine contained a Likert-scale measure of other concepts including hiring interest, optimism, fear, political correctness attitudes, and work morale (Cohn et al. 2015; Jha and Shayo 2019; Kessler et al. 2019; Exley and Kessler 2022; Spenkuch et al. 2023; Braghieri 2024; Engelmann

[19]We count the following journals as part of the 'top five': *Quarterly Journal of Economics, American Economic Review, Journal of Political Economy, Review of Economic Studies, Econometrica.*

et al. 2024; Englmaier et al. 2024; Gagnon et al. 2025).

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 based on other Likert scales. The risk of sign reversal for wellbeing estimates in this subsample is around 41%. This is lower than that for risk (56%), trust (58%), political preferences (44%), and other the 'other' concepts (47%).

Finally, to explore whether the risks of sign reversal vary across concepts, we also replicated the analysis of Table 2. Appendix Figure A11 shows that the predictors of reversal risk are similar across types of measures: most notably, larger t-statistics robustly reduce reversal risk. Thus, 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 a rough upper bound on the cost of deviation from linearity of $C = 0.15$. 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 scale-use 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.

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%. Restricting attention to the subset of wellbeing studies in top-five journals, the risk is lower: around 40% under a purely ordinal interpretation. The corresponding values for our sample of non-wellbeing Likert scales lies between 44% and 58%. Among wellbeing-based coefficients with p-values below 0.05 — the ones typically emphasised 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. If the linearity assumption were fully abandoned, roughly 86% of the estimates originally significant at the 1% level would remain robust at the 5% level. However, for estimates with p-values between 0.05 and 0.01, the risk of significance reversal is much larger, 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 potential deviations from linear scale-use substantially affect the absolute size of these coefficients and easily alters 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 small original p-values. 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 0.05 and 0.01. Finally, estimated magnitudes and coefficient ratios are highly unstable. Here, too, do minimal non-linearities in scale use suffice to reverse researchers' substantive conclusions.

These results have 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, focus on a single type of wellbeing scale and could not pin-down the precise functional form by which scale use departs from linearity. Moreover, future work might fruitfully combine our analysis of how monotonic transformations affect regression results with current work on correcting for interpersonal differences in scale use (Prati and Senik 2025; Benjamin et al. 2023b). Further work may also draw on the partial identification literature to, for instance, provide confidence regions for identified coefficient ranges given some cost C (c.f. Imbens and Manski 2004; Tamer 2010). For current practice, we believe it useful for analysts to routinely probe the robustness of their headline results to monotonic transformations of their outcome variable. A contribution of this paper – and of our Stata routines – is to render such tests more tractable.

References

D. Acemoglu, A. Cheema, A. I. Khwaja, and J. A. Robinson. Trust in state and nonstate actors: Evidence from dispute resolution in Pakistan. *Journal of Political Economy*, 128: 3090–3147, 2020.

A. Alesina, S. Stantcheva, and E. Teso. Intergenerational mobility and preferences for redistribution. *American Economic Review*, 108:521–554, 2018.

Y. Algan and P. Cahuc. Inherited trust and growth. *American Economic Review*, 100: 2060–2092, 2010.

V. Angelini, D. Cavapozzi, L. Corazzini, and O. 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*, 76:643–666, 2014.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton, NJ, 2009.

W. P. Banks and M. J. Coleman. Two subjective scales of number. *Perception & Psychophysics*, 29:95–105, 1981.

W. P. Banks and D. K. Hill. The apparent magnitude of number scaled by random production. *Journal of Experimental Psychology*, 102:353, 1974.

D. J. Benjamin, K. Cooper, O. Heffetz, and M. Kimball. From happiness data to economic conclusions. *Annual Review of Economics*, 16:359–391, 2023a.

D. J. Benjamin, K. Cooper, O. Heffetz, M. S. Kimball, and J. Zhou. Adjusting for scale-use heterogeneity in self-reported well-being. Technical report, National Bureau of Economic

Research, 2023b.

M. Bertrand and S. Mullainathan. Do people mean what they say? Implications for subjective survey data. *American Economic Review*, 91:67–72, 2001.

D. G. Blanchflower and A. J. Oswald. Well-being over time in Britain and the USA. *Journal of Public Economics*, 88:1359–1386, 2004.

J. R. Bloem. How much does the cardinal treatment of ordinal variables matter? An empirical investigation. *Political Analysis*, 30:197–213, 2022. doi: 10.1017/pan.2020.55.

J. R. Bloem and A. J. Oswald. The analysis of human feelings: A practical suggestion for a robustness test. *Review of Income and Wealth*, 68:689–710, 2022.

T. N. Bond and K. Lang. The sad truth about happiness scales. *Journal of Political Economy*, 127:1629–1640, 2019.

M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1: 97–111, 2010.

L. Braghieri. Political correctness, social image, and information transmission. *American Economic Review*, 114:3877–3904, 2024.

S. Carattini and M. Roesti. Trust, happiness, and pro-social behavior. *Review of Economics and Statistics*, 107:967–981, 2025.

L.-Y. Chen, E. Oparina, N. Powdthavee, and S. Srisuma. Robust ranking of happiness outcomes: A median regression perspective. *Journal of Economic Behavior & Organization*, 200:672–686, 2022.

A. E. Clark and A. J. Oswald. Unhappiness and unemployment. *The Economic Journal*, 104:648–659, 1994.

A. E. Clark, C. D'Ambrosio, and S. Ghislandi. Adaptation to poverty in long-run panel data. *Review of Economics and Statistics*, 98:591–600, 2016.

A. Cohn, J. Engelmann, E. Fehr, and M. A. Maréchal. Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, 105: 860–885, 2015.

A. M. Danzer and N. Danzer. The long-run consequences of Chernobyl: Evidence on subjective well-being, mental health and welfare. *Journal of Public Economics*, 135:47–60, 2016.

A. Dechezleprêtre, A. Fabre, T. Kruse, B. Planterose, A. Sanchez Chico, and S. Stantcheva. Fighting climate change: International attitudes toward climate policies. *American Economic Review*, pages 1258–1300, 2025.

T. Dohmen, A. Falk, D. Huffman, and U. Sunde. Are risk aversion and impatience related

to cognitive ability? *American Economic Review*, 100:1238–1260, 2010.

P. Dolan, G. Kavetsos, C. Krekel, D. Mavridis, R. Metcalfe, C. Senik, S. Szymanski, and N. R. Ziebarth. Quantifying the intangible impact of the Olympics using subjective well-being data. *Journal of Public Economics*, 177:104043, 2019.

R. A. Easterlin. Does economic growth improve the human lot? Some empirical evidence. In P. A. David and M. W. Reder, editors, *Nations and households in economic growth*, pages 89–125. Academic Press, New York, 1974.

J. B. Engelmann, M. Lebreton, N. A. Salem-Garcia, P. Schwardmann, and J. J. van der Weele. Anticipatory anxiety and wishful thinking. *American Economic Review*, 114:926–960, 2024.

F. Englmaier, S. Grimm, D. Grothe, D. Schindler, and S. Schudy. The effect of incentives in nonroutine analytical team tasks. *Journal of Political Economy*, 132:2695–2747, 2024.

C. L. Exley and J. B. Kessler. The gender gap in self-promotion. *The Quarterly Journal of Economics*, 137:1345–1381, 2022.

M. Fabian. Scale norming undermines the use of life satisfaction scale data for welfare analysis. *Journal of Happiness Studies*, 23:1509–1541, 2022.

A. Falk, A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde. Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133:1645–1692, 2018.

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

S. Flèche. The welfare consequences of centralization: Evidence from a quasi-natural experiment in Switzerland. *Review of Economics and Statistics*, 103:621–635, 2021.

P. Frijters and C. Krekel. *A handbook for wellbeing policy-making: History, theory, measurement, implementation, and examples*. Oxford University Press, 2021.

N. Gagnon, K. Bosmans, and A. Riedl. The effect of gender discrimination on labor supply. *Journal of Political Economy*, 133:000–000, 2025.

G. W. Imbens and C. F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72:1845–1857, 2004.

S. Jha and M. Shayo. Valuing peace: The effects of financial market exposure on votes and political attitudes. *Econometrica*, 87:1561–1588, 2019.

D. W. Johnston, M. A. Shields, and A. Suziedelyte. Victimisation, well-being and compensation: Using panel data to estimate the costs of violent crime. *The Economic Journal*, 128:1545–1569, 2018.

C. Kaiser. Using memories to assess the intrapersonal comparability of wellbeing reports. *Journal of Economic Behavior & Organization*, 193:410–442, 2022.

C. Kaiser and A. J. Oswald. The scientific value of numerical measures of human feelings. *Proceedings of the National Academy of Sciences*, 119:e2210412119, 2022.

C. Kaiser and A. Prati. Why you should measure subjective changes. Warwick Business School, Mimeo, 2025.

C. Kaiser and M. Vendrik. How much can we learn from happiness data? University of Oxford, Mimeo, 2023.

J. B. Kessler, C. Low, and C. D. Sullivan. Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review*, 109:3713–3744, 2019.

R. W. Klein and R. P. Sherman. Shift restrictions and semiparametric estimation in ordered response models. *Econometrica*, 70(2):663–691, 2002.

A. Knabe, S. Rätzel, R. Schöb, and J. Weimann. Dissatisfied with life but having a good day: Time-use and well-being of the unemployed. *The Economic Journal*, 120:867–889, 2010.

I. Kuziemko, M. I. Norton, E. Saez, and S. Stantcheva. How elastic are preferences for redistribution? Evidence from randomized survey experiments. *American Economic Review*, 105:1478–1508, 2015.

R. Layard, A. E. Clark, F. Cornaglia, N. Powdthavee, and J. Vernoit. What predicts a successful life? A life-course model of well-being. *The Economic Journal*, 124:F720–F738, 2014.

S. Liu and N. Netzer. Happy times: Measuring happiness using response times. *American Economic Review*, 113:3289–3322, 2023. doi: 10.1257/aer.20221340.

A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*, volume 143 of *Mathematics in Science and Engineering*. Academic Press, New York, 1979.

F. Molinari. Microeconometrics with partial identification. *Handbook of econometrics*, 7: 355–486, 2020.

A. J. Oswald. Happiness and economic performance. *The Economic Journal*, 107:1815–1831, 1997.

A. J. Oswald. On the curvature of the reporting function from objective reality to subjective feelings. *Economics Letters*, 100:369–372, 2008.

M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021.

R. Perez-Truglia. The effects of income transparency on well-being: Evidence from a natural experiment. *American Economic Review*, 110:1019–1054, 2020.

A. Prati and C. Senik. Is it possible to raise national happiness? CEP Discussion Paper 2068, 2025.

E. Riley. Resisting social pressure in the household using mobile money: Experimental evidence on microenterprise investment in uganda. *American Economic Review*, 114: 1415–1447, 2024.

B. Schneider, S. Parker, D. Ostrosky, D. Stein, and G. Kanow. A scale for the psychological magnitude of number. *Perception & Psychophysics*, 16:43–46, 1974.

C. Schröder and S. Yitzhaki. Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, 92:337–358, 2017.

J. L. Spenkuch, E. Teso, and G. Xu. Ideology and performance in public organizations. *Econometrica*, 91:1171–1203, 2023.

R. Studer. Does it matter how happiness is measured? Evidence from a randomized controlled experiment. *Journal of Economic and Social Measurement*, 37:317–336, 2012.

E. Tamer. Partial identification in econometrics. *Annual Review of Econonmics*, 2:167–195, 2010.

UK HMRC Treasury. Wellbeing guidance for appraisal: Supplementary green book guidance. Report, HM Treasury, 2021.

B. Van Praag. The welfare function of income in Belgium: An empirical investigation. *European Economic Review*, 2:337–369, 1971.

B. M. Van Praag. Ordinal and cardinal utility: An integration of the two dimensions of the welfare concept. *Journal of Econometrics*, 50:69–89, 1991.

B. M. Van Praag and N. L. Van der Sar. Household cost functions and equivalence scales. *Journal of Human Resources*, pages 193–210, 1988.

B. M. Van Praag et al. The measurement of welfare and well-being: The Leyden approach. *Well-being: The foundations of hedonic psychology*, pages 413–433, 1999.

Appendix

A Proofs and derivations

A.1 Proof of Proposition 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.

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:

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

Stacking over individuals, we 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.2 Proof of Proposition 2

From Equation 1 and Assumptions 1 and 2 we have:

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

Given Assumption 1, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{s}$ is a consistent estimator of $\boldsymbol{\beta}$. Given Assumption 2, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\zeta}$ is a consistent estimator of $\boldsymbol{\gamma}$. 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 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: $\frac{\hat{\beta}_m^{(\tilde{r})}}{\hat{\beta}_n^{(\tilde{r})}} = \frac{\sum_{k=1}^{K-1}(\tilde{l}_k - \tilde{l}_{k+1})\hat{\beta}_{km}^{(d)}}{\sum_{k=1}^{K-1}(\tilde{l}_k - \tilde{l}_{k+1})\hat{\beta}_{kn}^{(d)}}$.

There are 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 $\tilde{w}_k > 0$. Then:

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

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.[20].

A.4 Derivation of variance-covariance matrices

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

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 dichotomised variables $\mathbf{d}_k$. The residuals from the transformed regression are:

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

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 $\hat{\boldsymbol{\beta}}^{(\tilde{r})} = \sum_{k=1}^{K-1}(\tilde{l}_k - \tilde{l}_{k+1})\hat{\boldsymbol{\beta}}_k^{(d)} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{l}_K\mathbf{I}$. Substituting these expressions:

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

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 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I} = \mathbf{I}$.

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

Using this decomposition, we can express the variance-covariance matrix estimator $\hat{\mathbf{\Omega}}$ for different error structures. For homoskedastic errors, the variance estimator is:

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

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 $\mathbf{X}$.

For the Huber-White heteroskedasticity-robust variance estimator we get:

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

Finally, for G clusters, the clustered variance estimator is:

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

A.5 Derivation of maxVar($\Delta\tilde{\mathbf{l}}$)

We here show that $\text{maxVar}(\Delta\tilde{\mathbf{l}}) = \frac{K-2}{(K-1)^2}(l_K - l_1)^2$. Let differences between adjacent labels be given by $d_k \equiv \tilde{l}_{k+1} - \tilde{l}_k$ for $k = 1, ..., K-1$. The variance of these differences is given by $\text{Var}(\Delta\tilde{\mathbf{l}}) = \frac{1}{K-1}\sum_{k=1}^{K-1}(d_k - \bar{d})^2$, 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 maximised 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 **Normalisation** constraint), the maximum variance occurs when one difference approaches L and all other $K-2$ differences approach 0. The maximum variance is then given by $\text{maxVar}(\Delta\tilde{\mathbf{l}}) = \frac{1}{K-1}\left[(L-\bar{d})^2 + (K-2)(0-\bar{d})^2\right]$. Substituting $\bar{d} = \frac{L}{K-1}$, we obtain:

$$\begin{aligned}
\text{maxVar}(\Delta\tilde{\mathbf{l}}) &= \frac{1}{K-1}\left[\left(L - \frac{L}{K-1}\right)^2 + (K-2)\left(\frac{L}{K-1}\right)^2\right] \\
&= \frac{1}{K-1}\left[L^2\left(\frac{K-2}{K-1}\right)^2 + (K-2)\frac{L^2}{(K-1)^2}\right] \\
&= \frac{L^2}{(K-1)^3}\left[(K-2)^2 + (K-2)\right] = \frac{L^2(K-2)(K-1)}{(K-1)^3} = \frac{K-2}{(K-1)^2}(l_K - l_1)^2
\end{aligned}$$

B Representation Theorem for Plausibility Measures

Here, we postulate (and motivate) a set of desiderata, or 'axioms', that any 'plausibility measure' of deviations from linear scale-use should satisfy. We derive a representation theorem to characterise the class of functions that satisfies these axioms. As we then show, our cost function of the main text is a member of this class. The specific cost function we use, setting $\alpha = 2$, further has the useful property of being linearly homogenous.

B.1 Theorem

Let $\Delta_k \equiv \tilde{l}_{k+1} - \tilde{l}_k$. To simplify notation, we only consider the case where $\sum_k^{K-1} \Delta_k = 1$, or, equivalently, where $\tilde{l}_K - \tilde{l}_1 = 1$. This restriction to the unit interval is without loss of generality, as any monotonic transformation can be normalised to this domain and range.

Proposition A1 (Representation Theorem for Plausibility Measures)**.** *Any plausibility measure* $\Pi : \mathcal{D} \to [0, 1]$ *satisfying Axioms 1-5 can be represented with the form:*

$$\Pi(\Delta) = h\left(\frac{\sum_{k=1}^{K-1} \psi(\Delta_k)}{\max_{\Delta' \in \mathcal{D}} \sum_{k=1}^{K-1} \psi(\Delta'_k)}\right)$$

where:

- $\psi : \mathbb{R}_+ \to \mathbb{R}$ *is continuous and strictly convex with unique minimum at* $\Delta_k = 1/(K-1)$
- $h : [0, 1] \to [0, 1]$ *is a continuous and increasing function with* $h(0) = 0$ *and* $h(1) = 1$
- $\mathcal{D}$ *is the set of all valid gap distributions:*
  $\mathcal{D} = \left\{\Delta \in \mathbb{R}^{K-1} : \Delta_k \geq 0 \text{ for all } k, \text{ and } \sum_{k=1}^{K-1} \Delta_k = 1\right\}$.

B.2 Axioms

We postulate 5 desiderata, or 'axioms', for any plausibility measure.

Axiom 1 (Normalisation)**.**

- $\min_{\Delta \in \mathcal{D}} \Pi(\Delta) = 0$ *if* $\Delta_k = \frac{1}{K-1}$ *for all* k *(uniform gaps)*
- $\max_{\Delta \in \mathcal{D}} \Pi(\Delta) = 1$ *if* $\Delta_k = 1$ *for some* k *and* $\Delta'_k = 0$ *for all other* $k' \neq k$ *(single-jump distributions)*

Motivation: It is convenient to know the range of our plausibility measure. The measure should be at its minimum for linear scale-use. Single-jump scale-use is intuitively the most extremely non-linear scale-use. The plausibility measure should be at its maximum there. The minimum at uniformity and the maximum at single-jumps is also implied by Axiom 3.

Axiom 2 (Symmetry)**.** Π *is invariant under permutations of the indices. That is, for any permutation* σ *of* $\{1, 2, ..., K-1\}$*:*

$$\Pi(\Delta_1, ..., \Delta_{K-1}) = \Pi(\Delta_{\sigma(1)}, ..., \Delta_{\sigma(K-1)})$$

Motivation: We have no reason to treat gaps at different positions differently. A (say) compression between categories 2-3 should be just as 'implausible' as the same compression between categories 8-9. Any plausibility measure should thus be invariant to permutations of the indices.

Axiom 3 (Strict Spread Sensitivity)**.** *If* Δ *is a strict mean-preserving spread of* Δ'*, then* $\Pi(\Delta) > \Pi(\Delta')$*.*

Motivation: Greater dispersion in gaps represents greater departure from linearity. If one distribution of gaps is a mean-preserving spread of another, it should have a higher plausibility cost. Uniform gaps are most plausible, while increasingly unequal gaps, some large, some small, are less plausible. This axiom captures this intuition. As we will see, this axiom alone implies that uniform gaps minimize the plausibility measure and single-jumps maximize it.

Axiom 4 (Continuity)**.** Π *is continuous.*[21]

Motivation: Without continuity, arbitrarily small adjustments to scale interpretation could produce discontinuous jumps in plausibility. Continuity is convenient for optimisation.

Axiom 5 (Monotonic Additive Separability)**.** Π *has the additive form:*

$$\Pi(\Delta) = g\left(\sum_{k=1}^{K-1} \varphi_k(\Delta_k)\right)$$

for some monotonic function g *and some functions* φ_k*.*

Motivation: Additivity ensures that one gap's deviation does not affect the plausibility of others. Monotonicity of g ensures consistency in how individual gap contributions aggregate.

[21] I.e. for any $\epsilon > 0$, there exists $\delta > 0$ s.t. when $\sqrt{\sum_{k=1}^{K-1}(\Delta_k - \Delta'_k)^2} < \delta$, we have $|\Pi(\Delta) - \Pi(\Delta')| < \epsilon$.

B.3 Proof

Step 1: Establish Identical Component Functions.

Axiom 5 states that $\Pi(\Delta) = g\left(\sum_{k=1}^{K-1} \varphi_k(\Delta_k)\right)$. By Axiom 2 (Symmetry), for any permutation σ, we have $g\left(\sum_{k=1}^{K-1} \varphi_k(\Delta_k)\right) = g\left(\sum_{k=1}^{K-1} \varphi_k(\Delta_{\sigma(k)})\right)$. Since g is monotonic (Axiom 5), we also need: $\sum_{k=1}^{K-1} \varphi_k(\Delta_k) = \sum_{k=1}^{K-1} \varphi_k(\Delta_{\sigma(k)})$. Now consider the specific case where $\Delta = (x, y, 0, ..., 0)$ with $x + y = 1$. Under the transposition σ that swaps positions 1 and 2: $\varphi_1(x) + \varphi_2(y) = \varphi_1(y) + \varphi_2(x) \iff \varphi_1(x) - \varphi_2(x) = \varphi_1(y) - \varphi_2(y)$. Since this holds for all x, y, the difference $\varphi_1(x) - \varphi_2(x)$ must be constant. By similar arguments for all pairs of indices, all φ_k differ only by constants. Since adding constants to all φ_k can be absorbed into g, we can without loss of generality set all $\varphi_k = \psi$ for some common function ψ. We thus get: $\Pi(\Delta) = g\left(\sum_{k=1}^{K-1} \psi(\Delta_k)\right)$.

Step 2: Determine Convexity of ψ and Direction of g

Define the inner sum as $S(\Delta) = \sum_{k=1}^{K-1} \psi(\Delta_k)$. Axiom 3 states that for any mean-preserving spread Δ of Δ', we have $\Pi(\Delta) > \Pi(\Delta')$. This implies that Π is strictly Schur-convex (Marshall and Olkin 1979). Since $\Pi(\Delta) = g(S(\Delta))$ where g is monotonic (Axiom 5), and Π is strictly Schur-convex, we know that $S(\Delta)$ is strictly Schur-convex (if g is increasing) or strictly Schur-concave (if g is decreasing). A sum of the form $S(\Delta) = \sum_{k=1}^{K-1} \psi(\Delta_k)$ is strictly Schur-convex (Schur-concave) iff ψ is strictly convex (concave). Therefore, we have two possible cases:

- **Case A**: ψ is strictly convex and g is strictly increasing
- **Case B**: ψ is strictly concave and g is strictly decreasing

These cases are duals: If (ψ, g) is a valid concave/decreasing pair, we can define $\psi' = -\psi$ (strictly convex) and $g'(x) = g(-x)$ (increasing), which represents the same plausibility measure: $\Pi(\Delta) = g(S(\Delta)) = g(\sum \psi(\Delta_k)) = g(-\sum[-\psi(\Delta_k)]) = g'(\sum \psi'(\Delta_k))$.

Convention: We choose Case A without loss of generality. Therefore, ψ is **strictly convex** and g is **increasing**.

Step 3: Note that g and ψ are continuous

By Axiom 4, Π is continuous. Since g is monotonic and therefore cannot smooth out discontinuities, both $S(\Delta) = \sum \psi(\Delta_k)$ and g must be **continuous**. Continuity of $S(\Delta)$ requires $\psi(\Delta_k)$ to be **continuous**.

Step 4: Derive the Unique Minimum of ψ

By Axiom 1, $\Pi(\Delta) = 0$ if all $\Delta_k = 1/(K-1)$ and $\Pi(\Delta) = 0$ is the minimum of $\Pi(\Delta)$. Let $\mathbf{u} = (1/(K-1), ..., 1/(K-1))$. Then: $\Pi(\mathbf{u}) = g\left((K-1) \cdot \psi(1/(K-1))\right) = 0$.

From Step 2, we know that g is monotonically increasing. Since $\Pi(\Delta)$ is minimised at $\Pi(\mathbf{u}) = 0$, it follows that the input to g must also be minimised at $\Delta = \mathbf{u}$. Therefore, $\sum_{k=1}^{K-1} \psi(\Delta_k)$ **achieves a minimum** when $\Delta_k = 1/(K-1)$ for all k. Because $\psi(\Delta_k)$ is strictly convex, this **minimum is unique**.

Convention: We can set $\psi(1/(K-1)) = 0$ without loss of generality. If instead $\psi(1/(K-1)) = c$ for some constant c, we could define $g'(x) = g(x - c(K-1))$ to achieve the same plausibility measure. Setting the minimum to zero simplifies the notation.

Step 5: Show that Π is maximised at single-jump distributions

Since ψ is strictly convex (Step 2) the sum $\sum_{k=1}^{K-1} \psi(\Delta_k)$ over $\mathcal{D} = \{\Delta : \sum \Delta_k = 1, \Delta_k \geq 0\}$ is maximised at the extreme points of $\mathcal{D}$. The single-jump distributions where $\Delta_j = 1$ for some j and $\Delta_k = 0$ for $k \neq j$ are the extreme points of $\mathcal{D}$. Finally, by symmetry, all single-jump distributions yield the same maximum value $M = \psi(1) + (K-2) \cdot \psi(0)$. Since g is monotonically increasing, Π is also maximised at single-jump distributions.

Step 6: Normalize

In step 4, we established that $\min_{\Delta \in \mathcal{D}} S(\Delta) = \sum_{k=1}^{K-1} \psi(\Delta_k) = 0$. In step 5, we showed that $\sum_{k=1}^{K-1} \psi(\Delta_k)$ obtains its maximum $M = \psi(1) + (K-2) \cdot \psi(0)$ at single-jump distributions. By Axiom 1 we must have: $\min_{\Delta \in \mathcal{D}} \Pi(\Delta) = g(0) = 0$ and $\max_{\Delta \in \mathcal{D}} \Pi(\Delta) = g(M) = 1$. Therefore, $g : [0, M] \to [0, 1]$, with g continuous (by step 3) and monotonically increasing (by step 2). Define $h : [0, 1] \to [0, 1]$ by $h(x) = g(M \cdot x)$. Then h is continuous and increasing, with $h(0) = g(0) = 0$ and $h(1) = g(M) = 1$. Now, for any Δ: $\Pi(\Delta) = g\left(\sum_{k=1}^{K-1} \psi(\Delta_k)\right) = g\left(M \cdot \frac{\sum_{k=1}^{K-1} \psi(\Delta_k)}{M}\right) = h\left(\frac{\sum_{k=1}^{K-1} \psi(\Delta_k)}{M}\right)$. Therefore: $\Pi(\Delta) = h\left(\frac{\sum_{k=1}^{K-1} \psi(\Delta_k)}{\max_{\Delta' \in \mathcal{D}} \sum_{k=1}^{K-1} \psi(\Delta'_k)}\right)$. This completes the proof.

B.4 The Variance-Based Cost and Linear Homogeneity

A particularly attractive choice satisfying the restrictions of the representation theorem of Proposition A1 is to set: $\psi(x) = (x - 1/(K-1))^2$ and $h(x) = \sqrt{x}$. This gives us:

$$\Pi(\Delta) = \sqrt{\frac{\sum_{k=1}^{K-1} (\Delta_k - 1/(K-1))^2}{\max_{\Delta'} \sum_{k=1}^{K-1} (\Delta'_k - 1/(K-1))^2}},$$

which corresponds to the cost function used in the main part of the paper (with $\alpha = 2$). To see this, note that when $\sum_{k=1}^{K-1} \Delta_k = 1$, we obtain $\sum_{k=1}^{K-1}(\Delta_k - 1/(K-1))^2 = (K-1)\cdot\text{Var}(\Delta)$. The maximum value, in turn, is given by $\max_{\Delta' \in \mathcal{D}} \sum_{k=1}^{K-1}(\Delta'_k - 1/(K-1))^2 = (K-2)/(K-1)$. From Appendix A.5, we have $\text{maxVar}(\Delta) = (K-2)/(K-1)^2$. We can therefore write $\Pi(\Delta) = \sqrt{\frac{\text{Var}(\Delta)}{\text{maxVar}(\Delta)}}$.

Usefully, this choice exhibits linear homogeneity. That is, for some new set of labels satisfying $\Delta^{(\lambda)} = \lambda\Delta + (1-\lambda)\mathbf{u}$, where $\mathbf{u} = (1/(K-1), ..., 1/(K-1))$, we obtain:

$$\Pi(\Delta^{(\lambda)}) = \lambda \cdot \Pi(\Delta)$$

To see this, consider the components of the interpolated distribution:

$$\begin{aligned} \Delta_k^{(\lambda)} = \lambda\Delta_k + (1-\lambda)\cdot\frac{1}{K-1} &\iff \Delta_k^{(\lambda)} - \frac{1}{K-1} = \lambda\left(\Delta_k - \frac{1}{K-1}\right) \\ &\iff \left(\Delta_k^{(\lambda)} - \frac{1}{K-1}\right)^2 = \lambda^2\left(\Delta_k - \frac{1}{K-1}\right)^2 \end{aligned}$$

Then, summing over all components and applying to the plausibility measure:

$$\Pi(\Delta^{(\lambda)}) = \sqrt{\frac{\lambda^2 \sum_{k=1}^{K-1}(\Delta_k - 1/(K-1))^2}{\max_{\Delta'} \sum_{k=1}^{K-1}(\Delta'_k - 1/(K-1))^2}} = \lambda \cdot \Pi(\Delta)$$

Hence, under this choice, if we make a scale e.g. '50% more uniform' (by mixing halfway with uniform), the plausibility cost is exactly halved.

C Supplemental Appendix

C.1 Proofs of Proposition 1 for FE, 2SLS, and continuous case

C.1.1 Fixed-effects case

Suppose we have panel data for respondents i and time period t. We collect the within-person means across $t = 1, 2, .., T_i$ 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 $\hat{\boldsymbol{\beta}}_{FE} = (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\dot{\tilde{\mathbf{r}}}$ and the result of Proposition 1 follows by the same argument.

C.1.2 2-SLS Case

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

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

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 $\hat{\boldsymbol{\beta}}_{IV}$ are equal to $\sum_{k=1}^{K-1}(\tilde{l}_k - \tilde{l}_{k+1})\hat{\boldsymbol{\beta}}^{(d)}_{IV,k}$.

C.1.3 Continuous Case

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

Proposition A2 (Non-reversal condition with continuous outcomes)**.** *Let r_i be a continuous variable with support $[r_{\min}, r_{\max}]$ and let $f : [r_{\min}, r_{\max}] \to \mathbb{R}$ 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 X_{im} in*

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

is invariant under all such transformations f iff the coefficient $\hat{\beta}_m^d(t)$ obtained from the regression of the dichotomised variable $\mathbb{1}\{r_i \leq t\}$ on $\mathbf{X}_i$ has 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

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

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, coefficient $\hat{\beta}_m$ is 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 $\hat{\beta}_m$.

D 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 $f : [0,1] \to [0,1]$ with which we plan to recode our depen-

dent variable r. As before, this restriction to the unit interval is without loss of generality, as (again) any monotonic transformation can be normalised 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:

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

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 normalised 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:

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

where $\bar{d}$ is the mean difference:

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

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

$$\begin{aligned} \mathrm{Var}(d) &= \frac{1}{K-1} \sum_{i=1}^{K-1} \left(f'(x_i) \cdot \frac{1}{K-1} - \frac{1}{K-1} \right)^2 \\ &= \frac{1}{K-1} \sum_{i=1}^{K-1} \frac{1}{(K-1)^2} \left(f'(x_i) - 1 \right)^2 = \frac{1}{(K-1)^2} \sum_{i=1}^{K-1} \frac{1}{(K-1)} \left(f'(x_i) - 1 \right)^2 \end{aligned}$$

As K increases, this sum approaches an integral:

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

Denote the variance of the derivative function over $[0, 1]$ as $\sigma^2_{f'}$. This is a fixed value for a given f. Then:

$$\mathrm{Var}(d) \approx \frac{\sigma^2_{f'}}{(K-1)^2}$$

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 $\text{maxVar}(d) = \left(\frac{1}{K-1} - \frac{1}{(K-1)^2}\right) \approx \frac{1}{K-1}$ for large K, we have:

$$\frac{\text{Var}(d)}{\text{maxVar}(d)} \approx \frac{\sigma^2_{f'}/(K-1)^2}{1/(K-1)} = \sigma^2_{f'}\frac{1}{(K-1)}$$

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 ($\sigma^2_{f'}$) 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:

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

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 $\sigma^2_{f'}$ equals 1, while for other values of $\sigma^2_{f'}$ the dependency on K is substantially reduced but not eliminated.[22] 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.

E Further evidence on γ

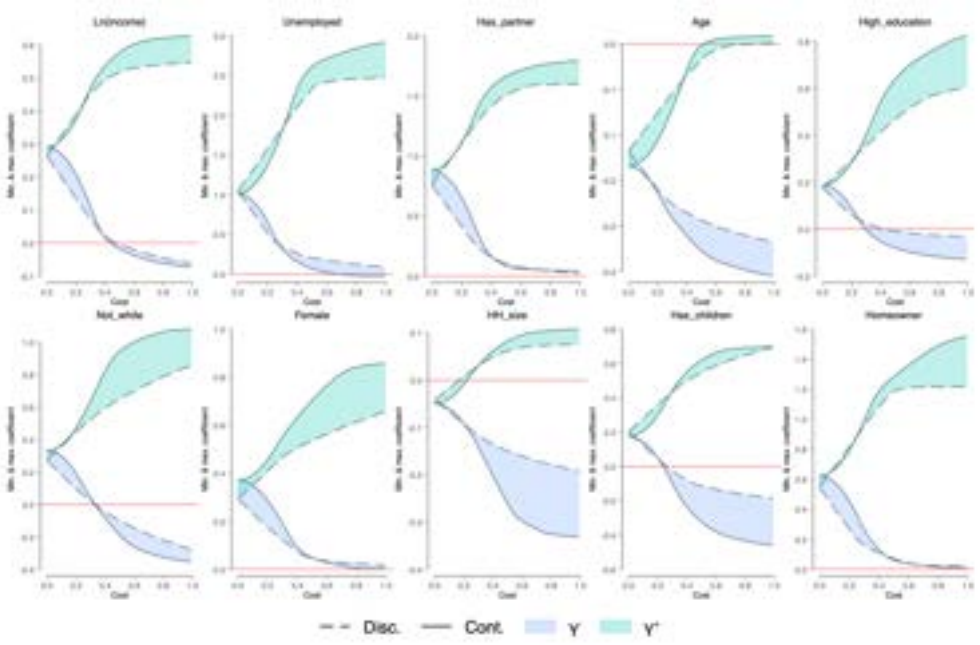
E.1 Worst-case estimates for γ when $C > 0$

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. for $C > 0$). We here perform a worst-case analysis on the potential influence and magnitude of γ_m in that case. We do so in two steps:

1. Using our continuous measure and every covariate m, we search for a transformation that yields a maximally positive and a maximally negative coefficient $\hat{\beta}_m^{(\tilde{r}^{(cont)})}$. Doing so

[22] To see this, we note the full expression: $C_\alpha \approx \left(\frac{\sigma^2_{f'}}{K-1}\right)^{1/(2\log_{10}(K-1))} = \left(\sigma^2_{f'}\right)^{1/(2\log_{10}(K-1))} \cdot 10^{-1/2}$
When $\sigma^2_{f'} = 1$, the first term equals $1^{1/(2\log_{10}(K-1))} = 1$. As in the unadjusted case for fixed α, for values of $\sigma^2_{f'} > 1$, our cost will decrease as K increases. In contrast, for $\sigma^2_{f'} < 1$, it will increase as K increases. However, this remaining dependency on 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 socio-economic 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.

may involve a reversal of coefficient signs compared to the original coefficient $\hat{\beta}_m^{(\tilde{r}^{(disc)})}$.

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 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 D, 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 D. 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.[23] 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)

E.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. We now extend our analysis 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 discretising our continuous measurement at equidistant points.[24]

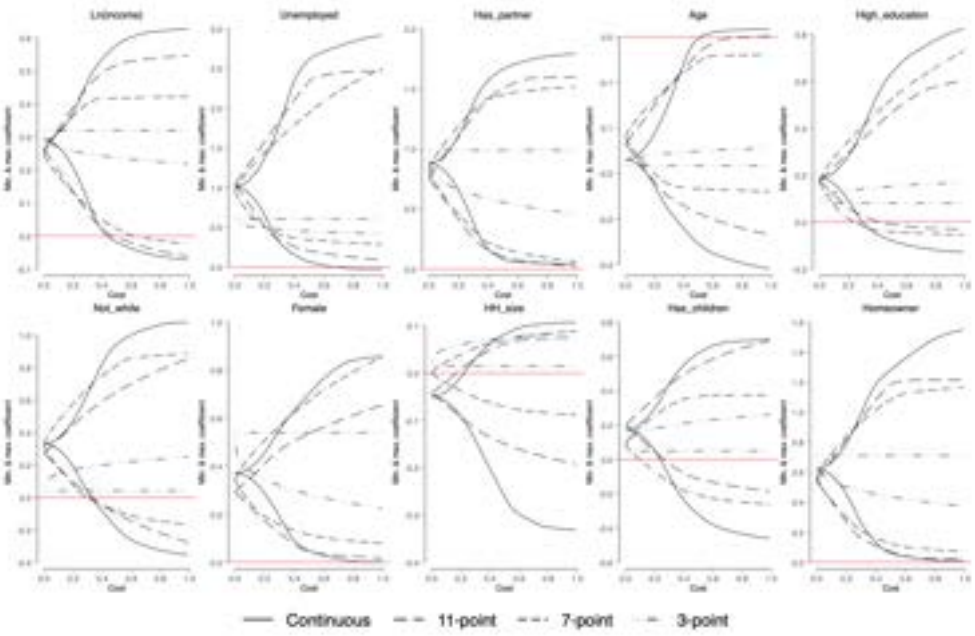
Figure A2 and Figures A5-A7 show the results. In line with the evidence of the previous section, we broadly observe that increasing the number of response categories also increases the possible spread of coefficient values at very high costs. As expected, we observe the smallest coefficient spreads for 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 discretising 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 round(.,.) 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, stabilising at about

[23] The 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.

[24] This 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 discretisations of r (Prolific data)



Note: This figure shows coefficient ranges for different variables across various scale discretisations (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.

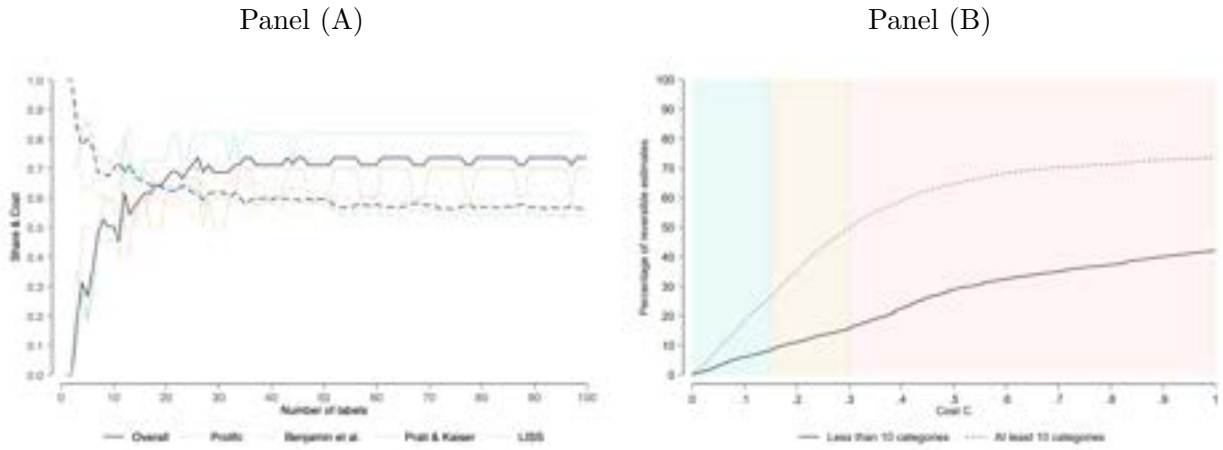
about 30 response options.[25] Likewise, the mean cost of reversals tends to decline with more response options. Overall, we observe that the share of reversible coefficients is about 20%-points higher for continuous scales than for discrete 11-point scales. It now seems clear that the share of reversible 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, with coefficients based on fewer categories. Panel (B) of Figure A3 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.

E.3 Implications

Three conclusions emerge. First, if satisfaction were measured continuously, the share of reversible results in the wellbeing literature would likely be somewhat higher than what we observe with discrete scales. Second, however, such reversals would rely on extremely well-targeted transformations 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

[25] With only 2 options, reversals are *never* possible and the conditional of Proposition 1 is trivially met.

Figure A3: Consequences of increasing the number of available response categories.

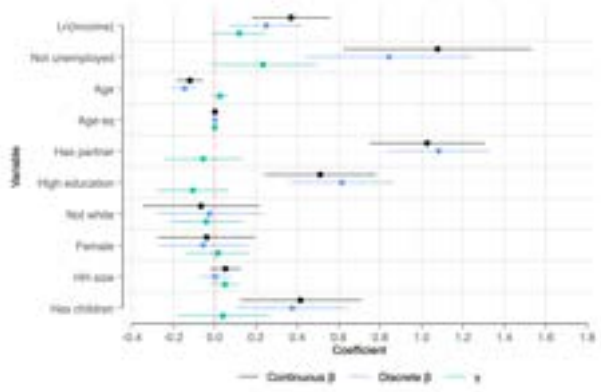


Note: Panel (A) illustrates how the number of available response categories affects both the share of reversible coefficients (solid lines) and the mean cost of reversals (dashed lines) across different datasets (as well as pooling across them). The share of reversible coefficients generally increases 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. Panel (B) shows cumulative percentages 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. See notes of Figure 5 for more details about C and the shaded regions.

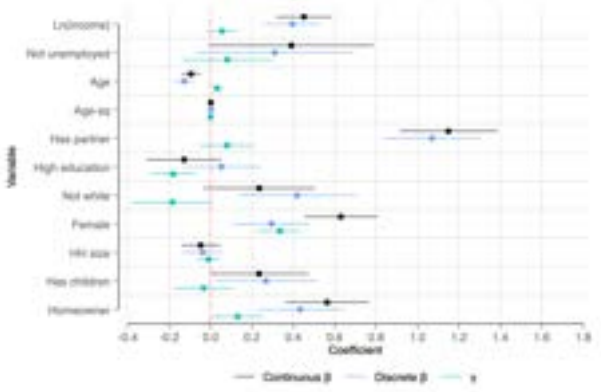
of $C = 0$ (i.e. linear scale use), we find that γ_m is typically close to zero. This suggests that the assumption of favourable within-category heterogeneity is reasonable for small C. Third, researchers should be especially cautious when working with wellbeing data based on few response categories (e.g., 3-point or 4-point scales). For these, it is not feasible to conservatively assess the robustness of results against transformations of the response scale.

Figure A4: Further evidence on γ_m when $C = 0$

Panel (A): Benjamin et al.



Panel (B): Prati & Kaiser



Panel (C): LISS

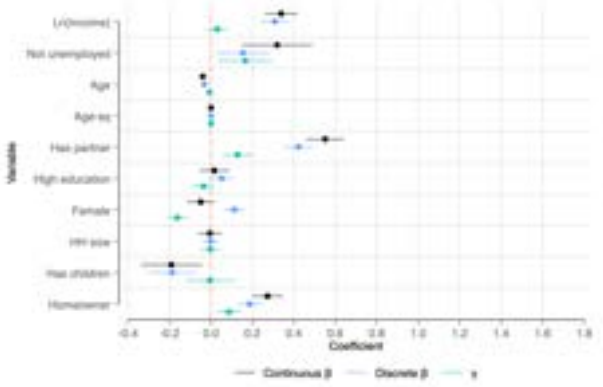


Figure A5: Coeff. spreads for different discretisations of r (Benjamin et al. data)

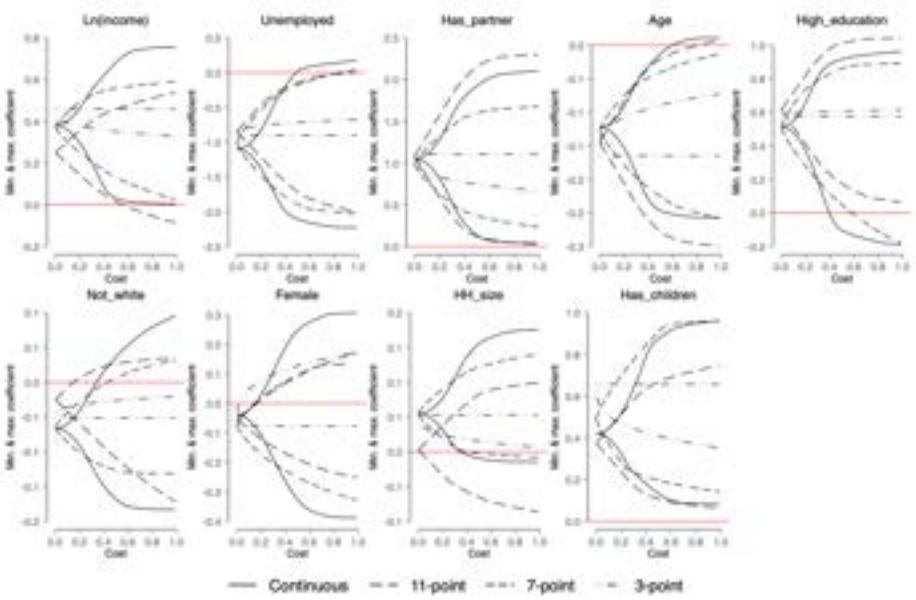


Figure A6: Coeff. spreads for different discretisations of r (Prati & Kaiser data)

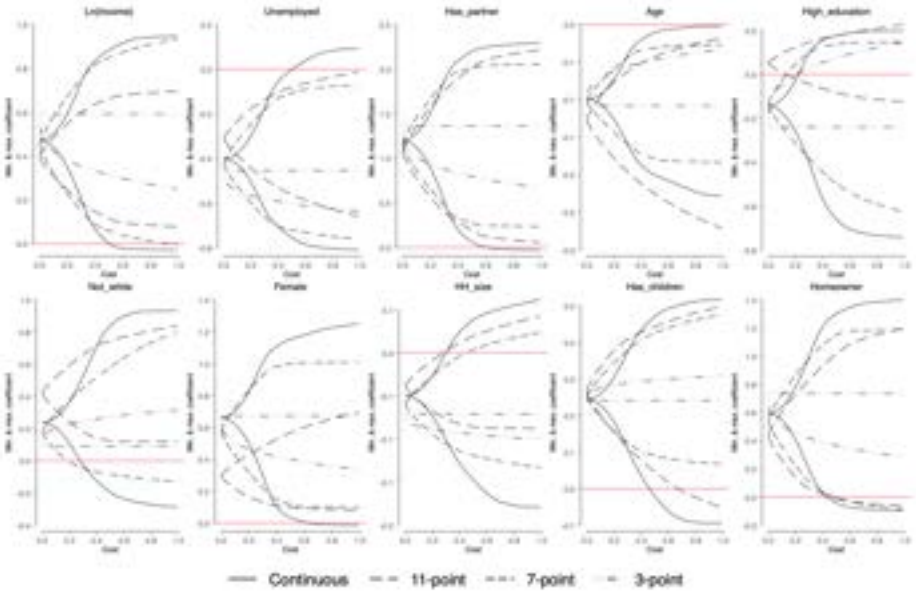
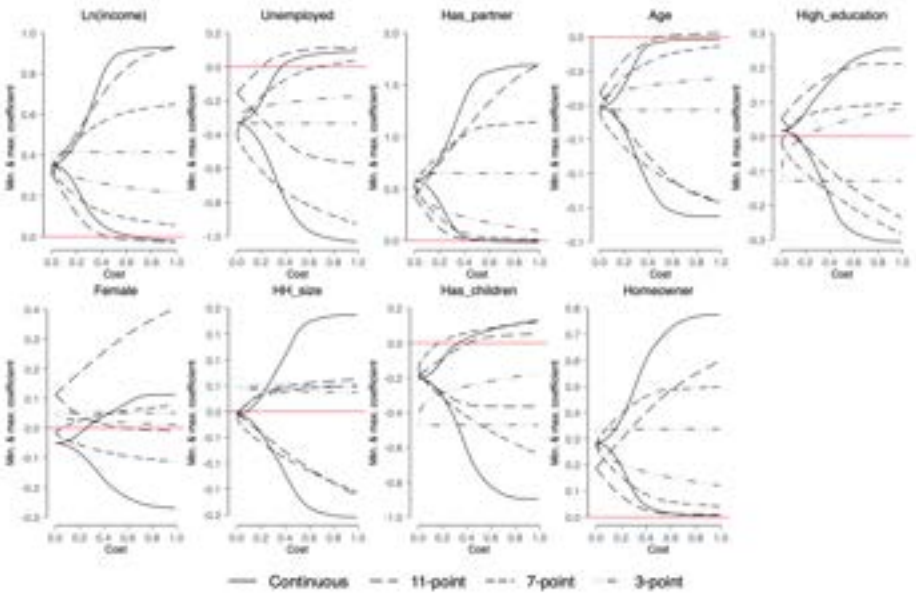


Figure A7: Coeff. spreads for different discretisations of r (LISS data)



F 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
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 WellBase
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	.	.	Some 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			
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 nowadays?	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
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

Continued on next page

Table A1 – *Continued from previous page*

Reference	Measure of r	Response options	Unemployment estimates	Income estimates	Comments
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 Probit-adjusted OLS replaced by OLS
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
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
Bha 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
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
Coville et al. (2023)	All things considered, how satisfied are you with your life as a whole these days?	10	.	.	None
	Taking all things together, would you say you are:	4			
Edmonds 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
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
Liu and Netzer (2023)	Taking all together, how would you say things are these days? Would you say that you are rather happy, neither happy nor unhappy or rather unhappy?	3	.	.	Ordered probit replaced by OLS
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 and Delavande (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	4	.	.	None
	Life satisfaction question - wording not reported	10			
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 nowadays?	11	.	.	Ordered logit replaced by OLS
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			
Carattini and Roesti (2025)	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 analysed 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			Ordered probit replaced by OLS
	How satisfied as a whole, 1 (not at all) to 4 (very satisfied)	4			
Courtemanche et al. (2025)	In general, how satisfied are you with your life?	11	✓	.	Ordered probit replaced by OLS

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	.
Oswald and Wu (2011)	US States Fixed effects	Table 2 - Column 4	Mix	NS to 1%	44%	0.010 to 0.980
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	.
Layard et al. (2014)	Income	Table 1 - Column 3	+	1%	Yes	0.354
	Education	Table 1 - Column 3	+	1%	Yes	0.049
	Having a job	Table 1 - Column 3	+	1%	No	.

(Continued from previous page)

Author	Test(s) of the paper	Source	Sign	Sig.	Reversal	Cost
	Good conduct	Table 1 - Column 3	+	1%	No	.
	Having a partner	Table 1 - Column 3	+	1%	Yes	0.856
	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
Campante and Yanagizawa-Drott (2015)	Ramadan hours	Table 2 - Column 12	+	1%	No	.
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	.
	Negative growth - European Sample	Table 1 - Column 4	-	1%	No	.

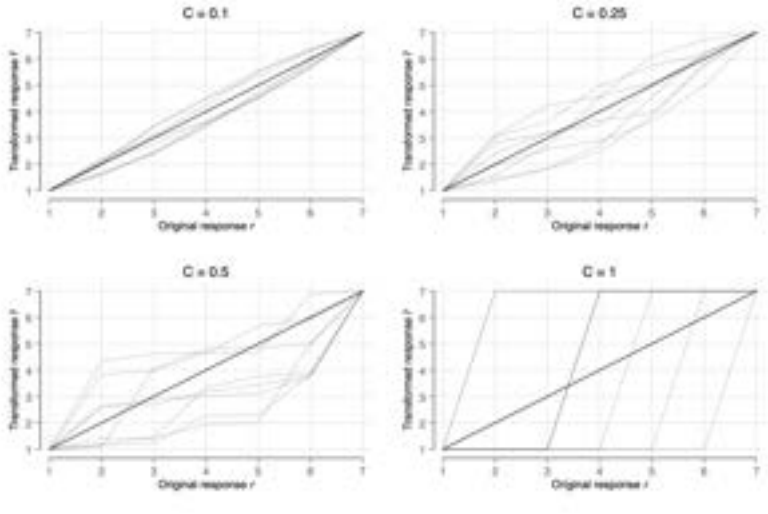
(Continued from previous page)

Author	Test(s) of the paper	Source	Sign	Sig.	Reversal	Cost
	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	.
Johnston et al. (2018)	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	.
	Plant closure (five to six year)	Table 2 - Column 8	-	1%	Yes	0.198
Perez-Truglia (2020)	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
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	.
Carattini and Roesti (2025)	Trust	Table 1 - Column 1	+	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.

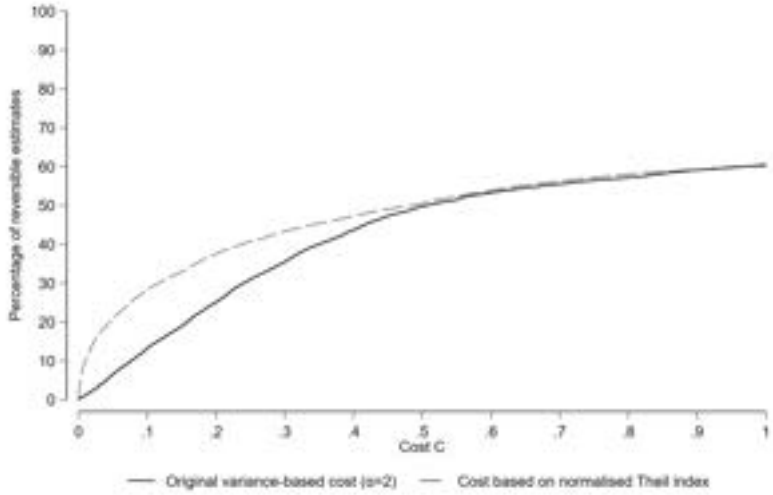
G Additional Tables and Figures

Figure A8: Examples of different values for $C_{\alpha=2}$ (with 7 response options)



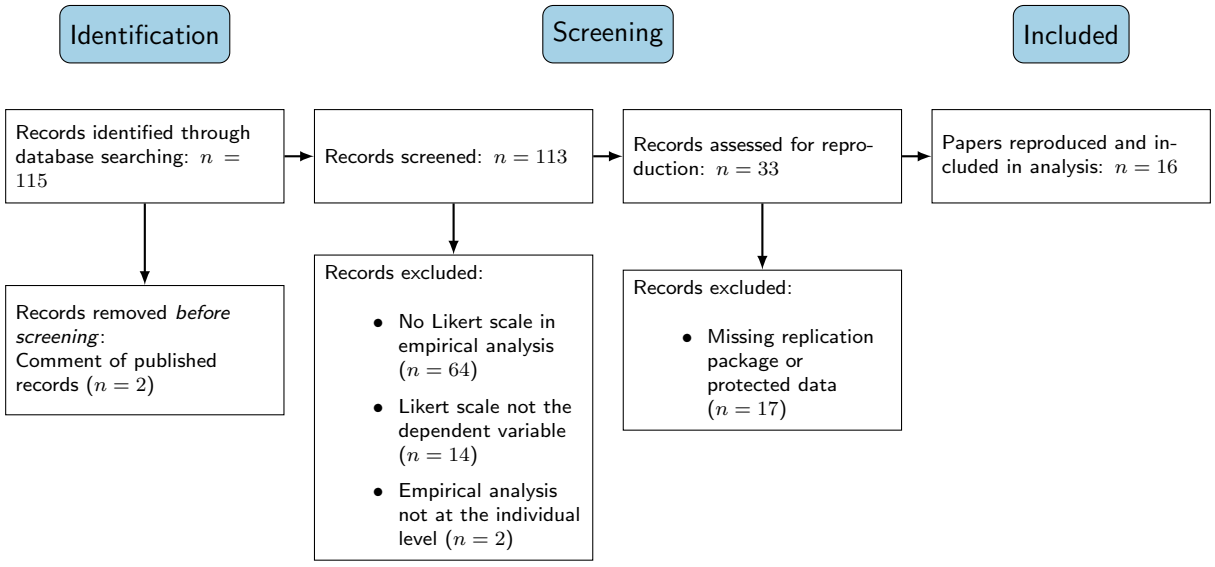
Note: Each line represents a possible transformation from r to $\tilde{r}$ that satisfies a given cost C.

Figure A9: Cumulative sign-reversal percentages when using Theil index as cost function



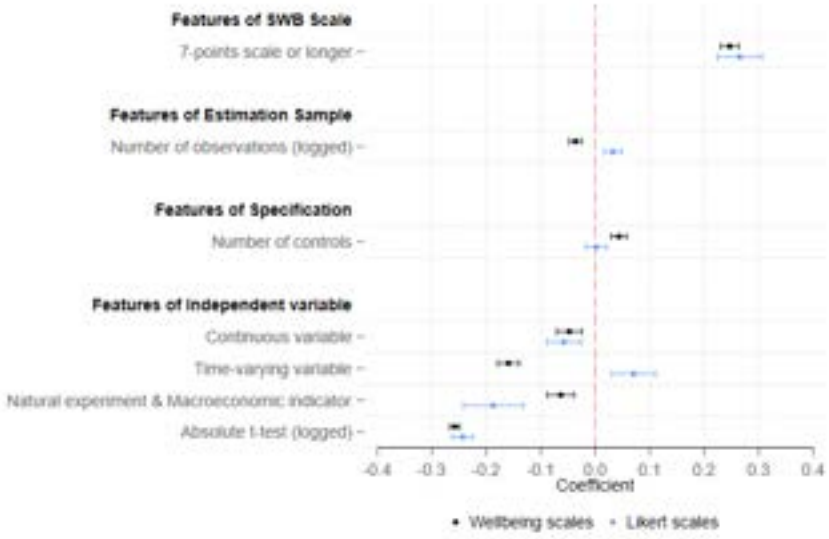
Note: Cumulative percentages of coefficients for which the sign can be reversed by at least one positive monotonic transformation of the response scale with at most cost C. When $C = 0$, this corresponds to the standard linearity assumption on scale use. When C may take on any value on the unit interval, 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.

Figure A10: PRISMA Chart - Other Likert scales



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

Figure A11: 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 A3: Predictors of the Probability and Cost of Sign-reversal: Robustness Checks

	P(Sign-reversal)			Cost of sign-reversal		
	(1)	(2)	(3)	(4)	(5)	(6)
About the estimation sample:						
Number of observations (logged)	0.026***	0.019**	0.048***	-0.029***	-0.026***	-0.060***
	(0.006)	(0.009)	(0.006)	(0.003)	(0.004)	(0.004)
About the econometric model:						
Number of controls	-0.001	-0.015	0.004	-0.004	-0.003	-0.009**
	(0.009)	(0.014)	(0.006)	(0.003)	(0.006)	(0.004)
Individual FE	0.002	0.120***	-0.000	-0.013*	0.077*	-0.010
	(0.013)	(0.035)	(0.015)	(0.007)	(0.041)	(0.012)
About the independent variable:						
Continuous variable	-0.032***	-0.030***	-0.034***	0.006	0.018***	0.003
	(0.007)	(0.007)	(0.007)	(0.005)	(0.005)	(0.006)
Time-varying variable	-0.078***	-0.084***	-0.068***	0.057***	0.066***	0.040***
	(0.008)	(0.008)	(0.007)	(0.004)	(0.004)	(0.005)
Two-stage least squares	-0.030	-0.051	-0.035	0.021	0.036**	0.011
	(0.032)	(0.031)	(0.029)	(0.018)	(0.015)	(0.021)
Natural experiment	-0.058***	-0.081***	-0.037***	0.034***	0.050***	0.023**
	(0.014)	(0.014)	(0.010)	(0.010)	(0.010)	(0.010)
Macroeconomic indicator	-0.066***	-0.036**	-0.062***	0.037***	0.015	0.047***
	(0.015)	(0.014)	(0.012)	(0.008)	(0.010)	(0.011)
Absolute t-statistics (logged)-0.285***	-0.282***	-0.315***	0.193***	0.196***	0.273***	
	(0.004)	(0.004)	(0.004)	(0.002)	(0.003)	(0.004)
Observations	28,522	28,522	28,522	17,243	17,243	28,522
Journal FE	✓	.	.	✓	.	.
Paper FE	.	✓	.	.	✓	.

Notes: Columns (3) reports marginal effects from probit models and Column (6) reports the coefficients of a linear Cragg hurdle model where we assign a cost of reversal of zero for estimates that cannot be reversed. 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 $***p < 0.01$.

H Additional tables on primary and secondary data

Table A4: 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 E.

Table A5: 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 E.

Table A6: Descriptive Statistics for Additional Datasets

	Benjamin et al.					Prati & Kaiser					LISS				
	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Wellbeing measure															
Life sat. (disc.)	1494	6.69	2.28	0.00	10.00	1931	5.90	2.11	0.00	10.00					
Life sat. (cont.)	1494	6.61	2.54	0.00	10.00	1928	6.60	2.05	0.00	10.00					
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)	1492	10.92	0.78	8.52	13.17	1926	7.33	0.81	5.52	9.10	7801	7.83	0.52	4.61	12.10
Unemployed	1471	0.10	0.30	0.00	1.00	1926	0.09	0.29	0.00	1.00	8548	0.05	0.22	0.00	1.00
Age	1493	45.95	12.79	21.62	83.62	1915	41.19	13.03	18.00	82.00	8548	51.19	17.26	16.00	97.00
Age Squared	1493	2274	1272	467	6992	1915	1866	1183	324	6724	8548	2918	1717	256	9409
Has partner	1494	0.50	0.50	0.00	1.00	1948	0.75	0.44	0.00	1.00	8548	0.59	0.49	0.00	1.00
Higher education	1484	0.64	0.48	0.00	1.00	1948	0.56	0.50	0.00	1.00	8527	0.29	0.46	0.00	1.00
Non-White	1494	0.27	0.44	0.00	1.00	1948	0.12	0.33	0.00	1.00	8548	0.00	0.00	0.00	0.00
Female	1494	0.55	0.50	0.00	1.00	1925	0.50	0.50	0.00	1.00	8548	0.53	0.50	0.00	1.00
Household Size	1494	2.79	1.62	1.00	12.00	1909	2.96	1.31	1.00	9.00	8548	2.56	1.31	1.00	8.00
Has Children	1493	0.35	0.48	0.00	1.00	1948	0.48	0.50	0.00	1.00	8548	0.38	0.49	0.00	1.00
Homeowner	1494	0.00	0.00	0.00	0.00	1948	0.65	0.48	0.00	1.00	8548	0.65	0.48	0.00	1.00

Note: Descriptive statistics for data from Benjamin et al., Prati & Kaiser, and LISS used in Section 3 and Appendix E.