

Web Appendices for Online Publication

“Correcting for Heterogeneity in the Use of Subjective Scales”

Contents

Appendix A. Related Literature	2
A.1 Anchoring Vignettes	2
A.2 Response Styles	5
Appendix B. Survey Design	7
B.1 Self-Rating (SR) Question Details	8
B.2 Calibration Question (CQ) Details	9
Appendix C. CQ Diagnostics and Tests	15
C.1 Top- and Bottom-Coding	16
C.2 Variation in Mean CQ Ratings	17
C.3 CQ Response Error Variance	17
C.4 Test of Linear Common Ranking of CQs (Assumption 1')	17
C.5 Test of (Monotonic) Common Ranking of CQs (Assumption 1)	18
C.6 Test for Linear CQ Translation Functions	20
Appendix D. CQ-Only Maximum Likelihood Estimation	20
D.1 Linear Translation Function	21
D.2 Quadratic Translation Function	22
D.3 Two-Wave Specification and the Persistence of Scale Use	22
D.4 Parameters for the Validation Exercise	23
Appendix E. Evidence on the Linearity of the CQ Translation Function	24
E.1 Individual-Level Tests for Quadratic Translation Functions	24
E.2 Estimation of Standard Deviation of Quadratic Coefficient	25
E.3 Cross-Group Average Translation Functions for Likert Scales	25
Appendix F. Semiparametric Estimators and Evaluating Approximation Error of MM Estimator	29
F.1 Semiparametric Estimators	29
F.2 Difference in Coefficients between MM and Semiparametric Estimators	33
F.3 Semiparametric Regression Estimator for the Quadratic Common Ranking Case	33
Appendix G. Test of Proportional Coefficients	36
Appendix H. Comparison of Results Between MM and CHOPIT	38
Appendix I. MM With Likert-Scale Response Options	40
Appendix J. Regressions Using $\hat{w}_i^*$ (Obtained by Inverting the Estimated CQ Translation Function)	42

Appendix K. Simulations	43
K.1 Data-Generating Process	43
K.2 Simulation Results	45
K.3 Discrete Response Scales.	46
K.4 Standard Errors	47
Appendix L. Additional Tables and Figure	49
Appendix M. Proofs	58
M.1 Proof of Proposition 1	59
M.2 Proof of Corollary 1	60
M.3 Proof of Proposition 2	60
M.4 Proof of Proposition 3	63
Appendix References.	65

A Related Literature

Our theoretical framework in Section VII unifies two traditions of scale-use correction, anchoring vignettes and response styles; our MM estimator has advantages over existing estimators in these literatures. In this appendix, we overview these relationships.

A.1 Anchoring Vignettes

In a pioneering paper in the political science literature, King et al. (2004) introduced a theoretical framework and econometric approach to scale-use correction based on same-dimension vignettes. A number of subsequent papers have adopted this approach in political science (e.g., Lin and Lee, 2024) as well as some in other fields (e.g., He et al., 2017), including a few in economics (e.g., Kapteyn et al., 2007; Montgomery, 2022).

King et al.’s theoretical framework is informal and based on two assumptions:

Vignette Equivalence: “the level of the variable represented in any one vignette is perceived by all respondents in the same way and on the same unidimensional scale, apart from random measurement error.”

Response Consistency: “each individual uses the response categories for a particular survey question in the same way when providing a self-assessment as when assessing each of the hypothetical people in the vignettes.”

Vignette Equivalence has two parts. We view our Common Ranking of CQs (Assumption 1) as a formalization of the “on the same unidimensional scale” part, as well as a generalization that

accommodates CQs that are not vignettes. It is not entirely clear how to formalize the other part, “perceived by all respondents in the same way,” but it would seem to be violated by plausible survey-response “biases,” such as an empathy gap (Deaton, 2011; see Section VII.C). Our theoretical framework shows that such an assumption is not needed.

Response Consistency only makes sense for same-dimension vignettes. In that case, every CQ can be identified with a particular situation; for this discussion only, we treat these vignette-described situations as if they were elements of $\mathcal{S}$. We therefore denote such a vignette by the ω notation we also use for the corresponding situation. With this notation, we formalize Response Consistency as: for all $i \in \mathcal{I}$, $\omega \in \mathcal{S}$, and $\omega' \in \mathcal{C}$, $\rho_i^*(\omega') = \rho_i(\omega'|\omega)$. This assumption implies that the CQ reporting function does not depend on a respondent’s own situation ω ; it corresponds to the “slope = 0” scenario in Figure 6. As noted in Section VII.C, this property would be violated by plausible biases and endogeneities, such as self-norming.

Response Consistency, together with the Interpersonal Comparability Presupposition, additionally implies our Assumptions 2 (Common Same-Rating Map) and 3 (SR-CQ Concordance), and is substantially stronger. Starting from the special case of Common Same-Rating Map that we call “self-CQ consistency” (see Section VII.D), Response Consistency additionally specifies how the respondent would self-rate in unobserved, counterfactual situations ω' that are different from the respondent’s own situation ω .

Furthermore, the statement of Response Consistency would seem to impose the restriction that context (which, in our formal framework, can affect scale use) is the same when answering the SR and the CQs. This restriction rules out, for example, that the survey order of the SR and the CQs could matter. Our assumptions do not impose this restriction.

There are more than a dozen papers that report tests of Vignette Equivalence, Response Consistency, or both (e.g., King et al., 2004; Bago d’Uva et al., 2011; Ferrer-i-Carbonell et al., 2011; Kapteyn et al., 2011; Grol-Prokopczyk et al., 2015). Almost all reported tests depend on strong auxiliary assumptions, such as the parametric assumptions of CHOPIT (discussed below) and homoskedasticity of response error variances across respondents (an assumption statistically rejected by Huang et al., 2021, one of the few papers on anchoring vignettes that allows for heteroskedasticity). The tests sometimes reject, but because the tests are opaque (e.g., poor fit from maximum-likelihood estimation), it is generally difficult to discern what drives the rejections. Moreover, for dimensions of practical interest such as job satisfaction, the tests of Response Consistency are not compelling because they rely on vignettes or objective measures of circumstances that do not fully describe the dimension (e.g., objective data relevant to job satisfaction that omit details such as relationships with coworkers).

In summary, our theoretical framework shows that the anchoring-vignette approach is justified under substantially weaker assumptions than Vignette Equivalence and Response Consistency. In

doing so, our framework clarifies that the approach is robust to many plausible violations of the original assumptions.

Turning to econometric strategy, King et al. (2004) introduced the Compound Hierarchical Ordered Probit (CHOPIT) model, which has been used in virtually all applications of the anchoring-vignette method.¹ In terms of our theoretical framework from Section VII, for the SR dimension of interest, there is a representation of the Common Ranking of Situations $y^\star : \mathcal{S} \rightarrow \mathbb{R}$ (analogous to a utility function). Each respondent's situation ω_i is represented as a latent dependent variable $y_i^\star$ corresponding to the level $y^\star(\omega_i)$.² This latent value is modeled as a linear function of the respondent's observed characteristics $\mathbf{x}_i$ and an error term: $y_i^\star = \mathbf{x}_i\delta + e_i^\star$. The set of response categories is discrete: $\mathcal{R} = \{1, \dots, k_{\max}\}$. As in an ordered probit, respondent i self-rates response category k if the latent level falls between thresholds: $r_i^\star = k$ if $\psi_i^{k-1} < \mathbf{x}_i\delta + e_i^\star \leq \psi_i^k$, where $\psi_i^0 = -\infty$ and $\psi_i^{k_{\max}} = +\infty$. CHOPIT differs from a standard ordered probit by also modeling the thresholds as functions of the respondent's observed characteristics: $\psi_i^1 = \lambda^1 \mathbf{x}_i$ and $\psi_i^k = \psi_i^{k-1} + \exp(\lambda^k \mathbf{x}_i)$ for $k \in \{2, \dots, k_{\max} - 1\}$. Each vignette v has a latent level y_v . Respondent i 's rating of a vignette is determined by: $r_{iv} = k$ if $\psi_i^{k-1} < y_v + e_{iv} \leq \psi_i^k$. The model is estimated by maximum likelihood. The model is identified given location and scale normalizations for the latent variable and the assumption that the error terms are independent and normally distributed. Note that in the standard implementation the self-ratings also enter the likelihood for the thresholds, which is the feature we return to in Web Appendix H.

CHOPIT is natural when the response scale has just a few categories (when the naïve regression would be ordered probit), but relative to the MM estimator we introduce, it has drawbacks analogous to those of ordered probit relative to OLS: it is less transparent, heteroskedasticity must be (but rarely is) parametrically modeled for coefficients to remain consistent, and the maximum-likelihood estimation can be poorly behaved (e.g., failing to converge) and unstable when there are more than a few response categories. This last point implies, first, that researchers using data with many response options have to round respondents' data, thereby losing information, before estimating CHOPIT (e.g., Montgomery, 2022). Second, importantly, the fewer the response options, the more severe are the concerns about lack of non-parametric identification, raised by Bond and Lang (2019),

¹King et al. (2004) also introduced an alternative, nonparametric approach: the SRs of respondents who rank all the vignettes in the same way are coded by where they rate themselves relative to the vignettes. But since the nonparametric approach does not allow for response error, respondents whose vignette rankings deviate from the full-sample ranking—which according to the theory cannot happen—need to be recoded somehow to match Common Ranking of CQs (our Assumption 1, their Vignette Equivalence) before analyzing the data. This nonparametric approach has been rarely used in applications because of its serious limitations. Most fundamentally, the goal in most applications is to run regressions or calculate means, which are not meaningfully applicable to ordinal-coded SRs without additional assumptions (such as those of CHOPIT).

²Although $y_i^\star$ plays a role similar to that of the common-scale level $w_i^\star$, it cannot be interpreted as the scale of a reference individual. Instead, the function $y^\star(\cdot)$ is, in effect, chosen by the CHOPIT maximum likelihood estimation so that the parametric assumptions of CHOPIT hold as nearly as possible.

mentioned in Section I.A and again below.

CHOPIT also has two major disadvantages more specific to the context of adjusting for scale-use heterogeneity in subjective survey responses. First, it assumes that there is some function $y^*(\cdot)$ such that the model’s parametric assumptions are correctly specified. This assumption is untestable, and in the related context of ordered probit (without scale-use heterogeneity), Bond and Lang (2019) have argued that results can be sensitive to the distributional assumptions.³ In contrast, key assumptions of the MM estimator, such as a linear CQ translation function, are testable. Second, the CHOPIT model has a large number of scale-use parameters: a length- $|\mathbf{x}_i|$ vector λ^k for each threshold. For example, with ten regressors (including a constant) in $\mathbf{x}_i$ and a 5-point Likert scale, the model has 40 scale-use parameters. Consequently, it is difficult to interpret the estimates and, in particular, assess their plausibility. In contrast, our linear translation function has only the intercept and stretcher parameters, both intuitive.

As mentioned in Section V and detailed in Web Appendix H, we find in our illustrative application that once the publicly available CHOPIT code is modified to estimate the scale-use parameters only from the CQ data, MM and CHOPIT produce similar estimates. In Web Appendix I, we also find that, relative to data from our 0–100 scale, MM produces similar estimates when the response scale has as few as 5 response options—and in simulations, with as few as 3 response options. In light of the above considerations, these results suggest that MM may be preferable even in settings with only a few response options.

A.2 Response Styles

A literature in psychology, marketing, and survey research, dating back at least to Cronbach (1946), studies people’s “response styles” when answering survey questions in general (for reviews, see, e.g., Van Vaerenbergh and Thomas, 2013; Weijters et al., 2016).⁴ Response styles are most often categorized as acquiescence, disacquiescence, extreme responding, and midpoint responding—respectively, the tendencies to give high, low, extreme, and midpoint ratings. In our linear translation

³Examining the distributional assumptions with supplementary information, such as response times (Liu and Netzer, 2023), faces challenges in cases with more than binary response categories.

⁴Three other related literatures in psychology and survey research are response shifts, range-frequency theory, and psychophysics. “Response shifts” refer to changes over time in how respondents map their own quality of life or self-rated health to survey responses (e.g., Sprangers and Schwartz, 1999; Rapkin and Schwartz, 2019). We focus on scale-use heterogeneity, which is considered one among several sources of response shifts (called “scale recalibration”). Range-frequency theory specifies a model of how respondents’ scale use for rating stimuli (e.g., the size of a box) depends both on the range of the stimuli and on their frequency distribution (e.g., Parducci, 1965). Our focus is on measuring and correcting for scale-use heterogeneity, and we remain agnostic about its determinants. Psychophysics, a subfield of psychology, is concerned with estimating reporting functions when the survey question asks about the individual’s perception of some objectively measurable physical stimulus, e.g., loudness of a sound (for a review, see Birnbaum, 1994). In contrast, except for Height and Weight we are concerned with SRs where no measure on an objective scale is available.

function, acquiescence and disacquiescence correspond to high and low intercepts, respectively, while extreme and midpoint responding correspond to high and low stretchers.⁵

There are two main existing approaches to adjusting for response style. First, some methods use the survey questions themselves to assess scale use; for example, the simplest such method is to standardize responses to a set of survey questions at the individual level (for a more sophisticated such method, see Rossi et al., 2001). This approach confounds scale-use heterogeneity with true response heterogeneity. Second, some methods measure scale use with an independent set of “control items,” such as questions randomly sampled from different psychological batteries (e.g., Weijters et al., 2010). This approach implicitly relies on an exclusion restriction: individuals’ responses to the survey question of interest (say, life satisfaction) are independent of their responses to the control items conditional on scale use. That assumption will rarely hold for SR dimensions such as life satisfaction, which are likely correlated with most psychological and economic variables that existing survey questions aim to measure. The CQs we analyze more cleanly identify scale-use heterogeneity because there is no variation in the “true” level of a CQ across respondents.

The response-style literature employs a wide variety of estimators, many of them based on latent factor models (e.g., Van Vaerenbergh and Thomas, 2013). The advantages of the MM estimator relative to these are similar to its advantages relative to CHOPIT: it is simple, transparent, and computationally quick.

Using estimates of response styles to correct for scale-use heterogeneity assumes that scale use does not depend on the dimension—a hypothesis we refer to as *general scale use*. Our data can provide some evidence on this hypothesis. Figure A.1 shows a measurement-error-adjusted correlation matrix of respondents’ scale-use-only benchmarks across dimensions (top-right triangle) and of respondents’ stretchers (bottom-left triangle). General scale use predicts correlations of one for the stretchers and close to one for the benchmarks. For the benchmarks, the strongest pairwise correlations, roughly 0.5, are among the vignette dimensions of Life Satisfaction, Happiness, Worthwhileness and Breathe. The strongest pairwise correlations among the visual CQs are among Darkness, Confidence, and Shape, with smaller magnitudes in the range 0.3–0.4. Between a vignette and a visual CQ, the strongest correlation is lower, at 0.27, between Confidence and Breathe, and most of the pairwise correlations are much smaller in magnitude, close to zero. The patterns of correlation are similar, but smaller in magnitude, for the stretchers. Overall, our data strongly reject general scale use.

A useful property of our MM estimator is that, when used for cross-dimension scale-use correction, it does *not* assume general scale use. It makes the weaker assumption that the projections onto the regressors of the net-of-benchmark SRs, $r_i^* - \hat{\tau}_i \left(\hat{\mathbb{E}} \left(r_i^* \right) \right)$, are perfectly correlated across CQ

⁵The point that the main response styles can be modeled by two parameters corresponding to an intercept and a stretcher has been previously highlighted by, e.g., Greenleaf (1992) and Rossi et al. (2001).

Figure A.1: Correlation Matrix of Scale-Use-Only Benchmarks and of Stretchers Across Dimensions

		Scale-Use-Only Benchmark (Response-Error Corrected)									
		Vignettes						Visual CQs			
		Anxiety	Worthwhileness	Life Satisfaction	No Anxiety	Happiness	Breathe	Pain	Confidence	Darkness	Shape
Stretcher (b)	Anxiety		-0.07	-0.15	N/A	-0.19	-0.13	0.16	-0.02	0.04	0.11
	Worthwhileness	0.32		0.39	0.35	N/A	0.50	0.00	0.20	0.06	0.09
	Life Satisfaction	0.25	0.29		0.30	0.51	0.50	-0.05	0.16	0.03	0.01
	No Anxiety	N/A	0.29	0.21		0.36	0.24	0.04	0.12	0.10	0.01
	Happiness	0.30	N/A	0.27	0.28		0.52	-0.02	0.19	0.03	0.01
	Breathe	0.34	0.37	0.32	0.33	0.37		-0.05	0.27	0.07	0.03
	Pain	0.09	0.09	0.08	0.08	0.09	0.08		0.08	0.12	0.16
	Confidence	0.08	0.14	0.12	0.09	0.12	0.15	0.20		0.40	0.34
	Darkness	0.01	0.03	0.03	0.04	0.03	0.06	0.17	0.24		0.32
	Shape	0.13	0.16	0.11	0.14	0.13	0.17	0.16	0.22	0.18	
		Scale-Use-Only Benchmark (Response-Error Corrected)									
		Stretcher (b)									

Notes: Data: UAS 571 (Survey 1). Sample sizes: 5,037–10,100 by dimension and 2,505–10,100 by estimable pair; N/A denotes pairs with no respondents in common due to CQ randomization. Top-right: pairwise correlations of scale-use-only benchmarks; bottom-left: pairwise correlations of stretchers. All correlations are response-error corrected: for each dimension, we subtract the estimated CQ response-error variance times the sum of squared benchmark or stretcher weights from the corresponding variance before calculating correlations. CQ response-error variances are average squared residual standard errors from respondent-level regressions of the three CQ ratings on the CQs' sample means.

sets. Under that assumption, the scale-use-corrected coefficient vectors are proportional, so every ratio of coefficients is unaffected by which CQs are used. Figure 5 and Web Appendix Table L.4 test this restriction and indicate that it may be a reasonable approximation when the CQ dimensions are closely related (e.g., Life Satisfaction and Happiness).

B Survey Design

This appendix details the survey questions for the Understanding America Study (UAS) data: UAS 571 (Survey 1) and UAS 719 (Survey 2). Additional details can be found in the UAS codebooks (Survey

1: <https://uasdata.usc.edu/survey/UAS+571>; Survey 2: <https://uasdata.usc.edu/survey/UAS+719>). All UAS data are available at <https://uasdata.usc.edu>.

B.1 Self-Rating (SR) Question Details

Table B.1 lists the 26 SRs in Survey 1, in alphabetical order. As noted in Section I.D, Survey 2 includes only the Life Satisfaction SR (for the dimension *How satisfied you are with your life*) and SRs for Height and Weight. Seven dimensions are shown with two versions, separated by a star: the “more is good” version which contains a double-negative (e.g., *You not feeling anxious*), and the “natural direction” version, which has a “more is bad” orientation (e.g., *You feeling anxious*). We randomize at the respondent level (50/50): respondents in the more-is-good treatment see the former version for these seven dimensions, and respondents in the natural-direction treatment see the latter.

Table B.1: Self-Rating Question Prompt and Dimensions in Survey 1 (UAS 571)

<i>Over the past year, on average, how would you have rated. . .</i>
How happy you feel
How much you enjoy your life
How satisfied you are with your life
The extent to which you feel the things you do in your life are worthwhile
The happiness of your family
The overall well-being of you and your family
You being a good person
You being able to support your family financially
You feeling that you have enough time for the things that are most important to you
You having many options and possibilities in your life and the freedom to choose among them
You having people you can turn to in time of need
You not being in physical pain *You being in physical pain
You not being lonely *You being lonely
You not feeling angry *You feeling angry
You not feeling anxious *You feeling anxious
You not feeling sad *You feeling sad
You not feeling stressed *You feeling stressed
You not feeling worried *You feeling worried
Your ability to breathe in and out easily
Your ability to remember things
Your mental health
Your physical health
Your physical safety and security
Your rating of your life on a ladder where the lowest rung is “worst possible life for you” and the highest rung is “best possible life for you”
Your sense of control over your life
Your sense of purpose

We developed this list of dimensions with several considerations in mind. The overall number was determined primarily by survey length constraints. When choosing specific questions, our primary goal was to try to cover dimensions of personal well-being as comprehensively as possible, informed by well-being measurement initiatives from the UK, New Zealand, Gallup, the OECD, and our own previous findings. To increase comprehensiveness, we also included questions to capture three dimensions that Benjamin et al. (2012) found people trade off against happiness and life satisfaction for: sense of purpose, control over one’s life, and family happiness. Our other major consideration was choosing dimensions that had high estimated relative marginal utilities in Benjamin et al. (2014) and/or our pilot data. Finally, we include two dimensions that correspond to a CQ trio, chosen because they are relatively narrow and objective: *Your ability to breathe in and out easily* and *Your ability to remember things*.

B.2 Calibration Question (CQ) Details

This section details the design of calibration questions (CQ) in Surveys 1 and 2: images for visual CQs and vignette text and instructions for vignette CQs.

Table B.2 shows the trio of CQ images for each dimension: Panel (a) shows Darkness, Shape, Confidence, and Pain from Survey 1. Panel (b) shows Height and Weight from Survey 2.

A vignette CQ consists of instructions, the prompt, and the response scale as well as the vignette text itself. Table B.3 shows these elements for each vignette CQ by framing. As noted in Section I.C, vignette framing was randomized at the respondent level in Survey 1 (with the same shown in Survey 2). We refer to the four framings in our UAS data as *Self*, *Nonself/Male*, *Nonself/Female*, and *You (stranger)*. In every framing, we describe the dimension in the prompt in the second person (e.g., *How happy you feel*), to keep the dimension being rated as similar as possible between the SR and corresponding CQs. The vignette text is the same for the Self and the You (stranger) framings, including our redesigned, recommended version of the You (stranger) framing (described further below).

In the *Self* framing, respondents are told to “Imagine everything in your life is the same as it is now, except for the details described in each situation.” We originally adopted this framing because it may be easier for respondents to accurately imagine a situation that is more similar to their own life, but we now realize that it induces respondents to “fill in the blanks” about the vignette person’s life with their own life details. When vignettes have multiple subdimensions and do not describe all of them, such vignette completion can generate violations of Assumption 1 (as explained in the main text’s footnote 26).

In the *You (stranger)* framing, the subject is described as a stranger but rated in the second person. Table B.3 gives two versions of this framing: the version as fielded and the redesigned version

Table B.2: Visual CQ Images

Panel (a): Images for Visual CQs in Survey 1 (UAS 571)					Panel (b): Images for Height and Weight CQs in Survey 2 (UAS 719)				
Dimension	Question Prompt	Low	Medium	High	Dimension	Question Prompt	Low	Medium	High
Darkness	How dark is this circle?				Height	“How would you describe the <u>height</u> of this man, relative to other male adults?”			
Shape	How complex is this shape?				Weight	“How would you describe the <u>weight</u> of this man, relative to other male adults of the same height?”			
Confidence	How confident does this person look?								
Pain	How would you rate this person’s level of pain?								

we would recommend for future use. In the *Nonself* framings, the vignette subject is described with a male or female name.⁶ In the *You (stranger)* and *Nonself* framings, where the dimension uses second-person wording (and *You (stranger)* has second-person in the vignette text itself), the instructions remind respondents that “you” and “your” refer to the stranger.

In the redesigned, recommended version of *You (stranger)*, we modify from the fielded version in two respects. First, we provide explicit instruction to fill in missing details with respondents’ best guess based on the details that are mentioned. If respondents have a common prior, then they will impute the same missing details; in this case, we eliminate the risk (discussed above and in the main text’s footnote 26) that vignette completion leads to violations of Assumption 1. Of course, in practice, respondents will impute different details. Nonetheless, as noted in Section VIII (“Design of CQs”), we believe this instruction is likely to get better consistency across respondents than

⁶Our strategy for choosing male and female names was the following: (i) use the most common names, as reported by the U.S. Social Security Administration, from the birth year corresponding to the average age of MTurk respondents in our pilot data (age 37, birth year 1983: https://www.ssa.gov/oact/babynames/state/top5_1983.htm); (ii) select a subset of names which seemed as much as possible, a priori, to have racial/ethnic neutrality and unambiguous gender connotation; (iii) omit names which were already used in vignettes from other sources: six names from our pilot CQs (Juan, Peter, Sasha, Rachel, James, and Taylor) and 24 names from Kapteyn et al. (2010) vignettes. We randomly assigned the names to CQs.

alternatives, especially because we believe many respondents will impute their best guess even if instructed otherwise. Second, we instruct respondents to use the “same standards” they used to rate themselves. We are concerned that the instructions we actually gave—to rate the vignette “as you would rate it for yourself”—might inadvertently nudge respondents to fill in details of the respondent’s own life—again, potentially leading to violations of Common Ranking of CQs (Assumption 1)—and might cause some respondents, but not others, to try to use the vignette subject’s scale use instead of imagining the *situation* of the vignette subject, inducing violations of Common Same-Rating Map (Assumption 2). We believe that our proposed instruction to use the “same standards” makes self-CQ consistency more likely. Although our assumptions do not require self-CQ consistency—specifically, they accommodate homogeneous self-CQ discrepancy—we worry that self-CQ discrepancies are unlikely to be homogeneous, so our proposed instruction aims to minimize them.

Table L.8, using the test of proportional coefficients (Web Appendix G), finds that we cannot reject the null that different framings yield equivalent MM results. We find no evidence that framing affects coefficient ratios. Because framing was randomized at the respondent level, pooled estimates give an average over framings and increase precision; we therefore pool across framings in all of our analyses.

Table B.3: Design of Vignettes by Framing

Vignette framing	Instruction screen before CQs	Preamble (top of CQ screen)	Vignette text (see Table B.4)	Prompt
Self	So far we have asked you to rate your own life. Now we will ask you to rate situations different from your own life. Here is how to do that: <i>Imagine everything in your life is the same as it is now, except for the details described in each situation.</i>	<u>Rating situations different from your own:</u> <i>Imagine everything in your life is the same as it is now, except for the details described in each situation below.</i> [For all framings: Additional note for “you feeling anxious” only: NOTE THAT FOR ALL THREE QUESTIONS ON THIS SCREEN, A HIGHER RATING MEANS MORE ANXIETY AND A LOWER RATING MEANS LESS ANXIETY. Additional note for “you not feeling anxious” only: NOTE THAT FOR ALL THREE QUESTIONS ON THIS SCREEN, A HIGHER RATING MEANS LESS ANXIETY AND A LOWER RATING MEANS MORE ANXIETY.]	Vignette: Self	If this situation described your life during the past year, on average, how would you have rated “[<i>dimension</i>]”?
You (stranger): fielded version	So far we have asked you to rate your own life. Now we will ask you to rate situations in other people’s lives. Here is how to do that: <i>Rate each situation in a stranger’s life as you would rate it for yourself. (Remember, the words “you” or “your” refer to the stranger. Put yourself in their shoes.)</i>	<u>Rating situations in other people’s lives:</u> <i>Rate each situation in a stranger’s life as you would rate it for yourself. (Remember, the words “you” or “your” refer to the stranger. Put yourself in their shoes.)</i>	Vignette: Self	[same as Self]
Nonself/Male	[same as You (stranger)]	[same as You (stranger)]	Vignette: Male	Over the past year, on average, how would you have rated
Nonself/Female	[same as You (stranger)]	[same as You (stranger)]	Vignette: Female	“[<i>dimension</i>]” in [name]’s life?
You (stranger): redesigned and recommended	So far we have asked you to rate your own life. Now we will ask you to rate other people’s lives using the same standards you used for yourself.	<u>How to rate other people’s lives:</u> • Use the same standards you used for yourself. • Each story describes a stranger. • Don’t assume they are like you. For anything not described in the story, make your best guess of what would be true based on what is described in the story. • As you read, put yourself in their shoes. “You” or “your” refers to the stranger.	Vignette: Self	[same as Self]

Table B.4: Vignette Texts in UAS 571 (Survey 1) and UAS 719 (Survey 2)

Level	Vignette: Self	Vignette: Male	Vignette: Female
Life Satisfaction: How satisfied you are with your life			
Low	You are working as a year-round yard maintenance worker, which doesn't pay well, and is not as pleasant as other occupations you would have liked to have. But you appreciate that you get to work outdoors. You don't have any contact with your parents, siblings, or old friends, but you recently started dating someone you like.	Preston is working as a year-round yard maintenance worker, which doesn't pay well, and is not as pleasant as other occupations he would have liked to have. But he appreciates that he gets to work outdoors. He doesn't have any contact with his parents, siblings, or old friends, but he recently started dating someone he likes.	Brianna is working as a year-round yard maintenance worker, which doesn't pay well, and is not as pleasant as other occupations she would have liked to have. But she appreciates that she gets to work outdoors. She doesn't have any contact with her parents, siblings, or old friends, but she recently started dating someone she likes.
Medium	You have not reached the management position you wanted to reach, but the pay and how much you like the work as a salesperson are acceptable. You don't have any contact with your parents and siblings, which bothers you, but you regularly spend time with your romantic partner's family, with whom you have a good relationship.	Dale has not reached the management position he wanted to reach, but the pay and how much he likes the work as a salesperson are acceptable. He doesn't have any contact with his parents and siblings, which bothers him, but he regularly spends time with his romantic partner's family, with whom he has a good relationship.	Leslie has not reached the management position she wanted to reach, but the pay and how much she likes the work as a salesperson are acceptable. She doesn't have any contact with her parents and siblings, which bothers her, but she regularly spends time with her romantic partner's family, with whom she has a good relationship.
High	After working in the same company for almost 10 years, you finally got your dream job (both financially and in what you like to do): manager of one of the branches. You and your romantic partner have a great relationship, but it bothers you that your relationship with your parents and siblings is badly strained.	After working in the same company for almost 10 years, Roderick finally got his dream job (both financially and in what he likes to do): manager of one of the branches. He and his romantic partner have a great relationship, but it bothers him that his relationship with his parents and siblings is badly strained.	After working in the same company for almost 10 years, Carly finally got her dream job (both financially and in what she likes to do): manager of one of the branches. She and her romantic partner have a great relationship, but it bothers her that her relationship with her parents and siblings is badly strained.
Life Satisfaction 2: How satisfied you are with your life			
Low	You are retired. Your spouse died 3 years ago and you still spend most of your time thinking about them and the times you spent together. You can make ends meet and have enough for your inexpensive hobbies, but have little money for extras such as expensive gifts for your two grandchildren. You are overweight and get tired when walking more than a block.	Paul is retired. His wife died 3 years ago and he still spends most of his time thinking about her and the times they spent together. He can make ends meet and has enough for his inexpensive hobbies, but has little money for extras such as expensive gifts for his two grandchildren. He is overweight and gets tired when walking more than a block.	Susan is retired. Her husband died 3 years ago and she still spends most of her time thinking about him and the times they spent together. She can make ends meet and has enough for her inexpensive hobbies, but has little money for extras such as expensive gifts for her two grandchildren. She is overweight and gets tired when walking more than a block.
Medium	You have been retired for five years. You still miss the contacts with your colleagues and would have liked to keep working part time. You and your spouse are doing OK financially and take a few small vacations every year. For the rest you each lead your own lives and don't do many things together. You have two children but rarely see them.	Mark has been retired for five years. He still misses the contacts with his colleagues and would have liked to keep working part time. He and his wife are doing OK financially and take a few small vacations every year. For the rest they each lead their own lives and don't do many things together. They have two children but rarely see them.	Joanna has been retired for five years. She still misses the contacts with her colleagues and would have liked to keep working part time. She and her husband are doing OK financially and take a few small vacations every year. For the rest they each lead their own lives and don't do many things together. They have two children but rarely see them.
High	You are retired and married. You and your spouse often argue, but recently celebrated your 40th wedding anniversary. You own the house you live in and have a large circle of friends. You play bridge twice a week. You are comfortable financially and go on nice vacations regularly with some friends. Lately you have been suffering from arthritis, which makes work in the house and garden painful.	Bob is retired and married. He and his spouse often argue, but recently celebrated their 40th wedding anniversary. He owns the house he lives in and has a large circle of friends. He plays bridge twice a week. He is comfortable financially and goes on nice vacations regularly with some friends. Lately he has been suffering from arthritis, which makes work in the house and garden painful.	Catherine is retired and married. She and her spouse often argue, but recently celebrated their 40th wedding anniversary. She owns the house she lives in and has a large circle of friends. She plays bridge twice a week. She is comfortable financially and goes on nice vacations regularly with some friends. Lately she has been suffering from arthritis, which makes work in the house and garden painful.

Continued on next page

Table B.4: Vignette Texts in UAS 571 (Survey 1) and UAS 719 (Survey 2) (continued)

Level	Vignette: Self	Vignette: Male	Vignette: Female
Happiness: How happy you feel			
Low	When you are alone, you almost always feel down, and sometimes drag your feet and look at the ground when you walk. You feel better when you are with your friends and often spend time with them. You often complain about your situation and things people have done to you.	When Edwin is alone, he almost always feels down, and sometimes drags his feet and looks at the ground when he walks. He feels better when he is with his friends and often spends time with them. He often complains about his situation and things people have done to him.	When Beth is alone, she almost always feels down, and sometimes drags her feet and looks at the ground when she walks. She feels better when she is with her friends and often spends time with them. She often complains about her situation and things people have done to her.
Medium	You are always smiling when you are with your friends and spend a lot of time with them. However, when you are alone, you often feel down. For some things, you're really grateful but for others, you blame other people or your circumstances.	Henry is always smiling when he is with his friends and spends a lot of time with them. However, when he is alone, he often feels down. For some things, he's really grateful but for others, he blames other people or his circumstances.	Deborah is always smiling when she is with her friends and spends a lot of time with them. However, when she is alone, she often feels down. For some things, she's really grateful but for others, she blames other people or her circumstances.
High	You feel grateful for many things and for the people in your life. You are always smiling when you are around friends or family. However, when you're alone, you sometimes feel down.	Tristan feels grateful for many things and for the people in his life. He is always smiling when he is around friends or family. However, when he's alone, he sometimes feels down.	Kendra feels grateful for many things and for the people in her life. She is always smiling when she is around friends or family. However, when she's alone, she sometimes feels down.
Worthwhileness: The extent to which you feel the things you do in your life are worthwhile			
Low	You aren't known for helping other people; you mostly just take care of yourself. You are planning to start helping more in your community in the next months though. You have to work a lot, but every 2 months you do find time to pursue your hobbies such as going out in nature. You often feel you are not really growing as a person, no matter what you do.	Eddie isn't known for helping other people; he mostly just takes care of himself. He is planning to start helping more in his community in the next months though. He has to work a lot, but every 2 months he does find time to pursue his hobbies such as going out in nature. He often feels he is not really growing as a person, no matter what he does.	Diane isn't known for helping other people; she mostly just takes care of herself. She is planning to start helping more in her community in the next months though. She has to work a lot, but every 2 months she does find time to pursue her hobbies such as going out in nature. She often feels she is not really growing as a person, no matter what she does.
Medium	You usually have to work a lot, but can go out hiking and camping around once a month. You also help out in your community about once a month and sometimes help out strangers, such as homeless people. You sometimes feel you aren't really growing as a person through this.	Stuart usually has to work a lot, but can go out hiking and camping around once a month. He also helps out in his community about once a month and sometimes helps out strangers, such as homeless people. He sometimes feels he isn't really growing as a person through this.	Alyssa usually has to work a lot, but can go out hiking and camping around once a month. She also helps out in her community about once a month and sometimes helps out strangers, such as homeless people. She sometimes feels she isn't really growing as a person through this.
High	Every other week, you go hiking and camping. You are active in your community and help at the soup kitchen serving food to the homeless every other week. While you think these things have an impact on other people, you sometimes question if they help you grow personally at all.	Every other week, Theodore goes hiking and camping. He is active in his community and helps at the soup kitchen serving food to the homeless every other week. While he thinks these things have an impact on other people, he sometimes questions if they help him grow personally at all.	Every other week, Martha goes hiking and camping. She is active in her community and helps at the soup kitchen serving food to the homeless every other week. While she thinks these things have an impact on other people, she sometimes questions if they help her grow personally at all.
Breathe: Your ability to breathe in and out easily			
Low	You have been smoking cigarettes for several years. You have a cough a few times a week that hurts in your chest and throat. You get breathless if you run, and it takes you a few minutes to get back to your normal breathing.	Marc has been smoking cigarettes for several years. He has a cough a few times a week that hurts in his chest and throat. He gets breathless if he runs, and it takes him a few minutes to get back to his normal breathing.	Kelli has been smoking cigarettes for several years. She has a cough a few times a week that hurts in her chest and throat. She gets breathless if she runs, and it takes her a few minutes to get back to her normal breathing.
Medium	You get out of breath whenever you walk faster than you usually do, such as when trying to catch the bus or exercise. This goes away after a minute or so when you slow down.	Luke gets out of breath whenever he walks faster than he usually does, such as when trying to catch the bus or exercise. This goes away after a minute or so when he slows down.	Ann gets out of breath whenever she walks faster than she usually does, such as when trying to catch the bus or exercise. This goes away after a minute or so when she slows down.
High	You are not an athlete, but you're in good condition. You take aerobics classes 3 times a week and go for long walks almost every day, but you are a little short of breath after climbing a flight of stairs.	Terry is not an athlete, but he's in good condition. He takes aerobics classes 3 times a week and goes for long walks almost every day, but he is a little short of breath after climbing a flight of stairs.	Whitney is not an athlete, but she's in good condition. She takes aerobics classes 3 times a week and goes for long walks almost every day, but she is a little short of breath after climbing a flight of stairs.

Continued on next page

Table B.4: Vignette Texts in UAS 571 (Survey 1) and UAS 719 (Survey 2) (continued)

Level	Vignette: Self	Vignette: Male	Vignette: Female
Remember: Your ability to remember things			
Low	You have a spotty memory for details of the past. You have to work hard to remember new things. You sometimes get lost, even in places you should know well. You often forget people's names, but you seldom forget important appointments.	Shaun has a spotty memory for details of the past. He has to work hard to remember new things. He sometimes gets lost, even in places he should know well. He often forgets people's names, but he seldom forgets important appointments.	Barbara has a spotty memory for details of the past. She has to work hard to remember new things. She sometimes gets lost, even in places she should know well. She often forgets people's names, but she seldom forgets important appointments.
Check [†]	You have a pretty good memory for details of the past. However, in order to make sure that you are reading all the questions, please ignore the sentence below and just pull the slider all the way to the left.	Maurice has a pretty good memory for details of the past. However, in order to make sure that you are reading all the questions, please ignore the sentence below and just pull the slider all the way to the left.	Tammy has a pretty good memory for details of the past. However, in order to make sure that you are reading all the questions, please ignore the sentence below and just pull the slider all the way to the left.
High	You have an excellent memory for details of the past, and you are good at memorizing new things. You occasionally forget people's names, and rely on to-do lists to avoid forgetting things you need to do.	Tony has an excellent memory for details of the past, and he is good at memorizing new things. He occasionally forgets people's names, and relies on to-do lists to avoid forgetting things he needs to do.	Denise has an excellent memory for details of the past, and she is good at memorizing new things. She occasionally forgets people's names, and relies on to-do lists to avoid forgetting things she needs to do.
Anxiety / No Anxiety: You feeling anxious / You not feeling anxious			
Low	You seldom get a good night's sleep because you can't turn your mind off, and often have stress nightmares. You are confident about your relationships and your health, but often feel that other things are out of control. You constantly worry about messing up in your work.	Dominic seldom gets a good night's sleep because he can't turn his mind off, and often has stress nightmares. He is confident about his relationships and his health, but often feels that other things are out of control. He constantly worries about messing up in his work.	Naomi seldom gets a good night's sleep because she can't turn her mind off, and often has stress nightmares. She is confident about her relationships and her health, but often feels that other things are out of control. She constantly worries about messing up in her work.
Medium	You sleep OK, but sometimes your thoughts wake you up in the middle of the night. You feel solid about many of the most important things in your life. But you worry a lot, even about little things.	Steve sleeps OK, but sometimes his thoughts wake him up in the middle of the night. He feels solid about many of the most important things in his life. But he worries a lot, even about little things.	Ellen sleeps OK, but sometimes her thoughts wake her up in the middle of the night. She feels solid about many of the most important things in her life. But she worries a lot, even about little things.
High	You usually sleep through the night without too much trouble. You generally feel things are under control. You sometimes worry about what could go wrong if you mess up or if things don't go your way, but not too often.	Marvin usually sleeps through the night without too much trouble. He generally feels things are under control. He sometimes worries about what could go wrong if he messes up or if things don't go his way, but not too often.	Summer usually sleeps through the night without too much trouble. She generally feels things are under control. She sometimes worries about what could go wrong if she messes up or if things don't go her way, but not too often.

Notes: This table lists every version of vignette text used in UAS 571 (Survey 1) and UAS 719 (Survey 2). Each panel heading gives the trio abbreviation used in this paper and the dimension rated in the corresponding SR question prompt. We merged the Anxiety and No Anxiety dimensions since they use the same vignette text. Level labels (Low, Medium, High) are keyed to the “more is good” (No Anxiety) orientation; for the Anxiety dimension the substantive ordering is reversed, as noted in Table C.1.

[†] In place of the medium-level vignette in the *Remember* trio, there is an attention check item. It instructs respondents to move the slider to the far left.

C CQ Diagnostics and Tests

In Sections VII and VIII (“Testing the Theoretical Assumptions for a CQ Type”), we discussed, conceptually, how tests of Assumptions 1–3 could be conducted for CQs in a given dimension. Here, we describe diagnostics and tests that can be conducted with a fixed set of CQs on which data has already been collected. We illustrate using our own data. Table C.1 presents the results of the key diagnostics, which we discuss in turn.

Table C.1: Diagnostics for Calibration-Question Trios (In Addition to Table 4)

CQ dimension	UAS survey	Kind	N	(1)	(2)	(3)	(4)
				Near-endpoint coding % (Low, Medium, High)	SD of CQ means	SD of CQ residuals	Linear Common Ranking test (p -value)
Life Satisfaction	571	Vignette	10,100	(2.9, 2.7, 3.0)	6.94	7.35	0.74
Happiness	571	Vignette	5,063	(3.9, 3.2, 2.8)	11.08	7.83	0.13
Worthwhileness	571	Vignette	5,037	(2.4, 2.8, 3.7)	8.94	7.15	0.43
No Anxiety	571	Vignette	5,062	(4.0, 2.5, 3.1)	11.79	8.51	0.53
Anxiety	571	Vignette	5,038	(2.8, 2.2, 5.2)	14.07	8.22	0.16
Breathe	571	Vignette	10,100	(6.4, 3.9, 4.8)	14.12	9.38	5.2×10^{-14}
Darkness	571	Visual	10,100	(3.9, 1.4, 3.1)	15.52	5.62	0.13
Shape	571	Visual	10,100	(4.6, 2.4, 4.2)	12.38	7.07	0.08
Confidence	571	Visual	10,100	(4.3, 2.4, 3.7)	16.17	8.44	2×10^{-22}
Pain	571	Visual	10,100	(5.0, 1.6, 5.7)	17.09	6.33	2.3×10^{-3}
Remember	571	Vignette	10,100	(3.3, –, 3.5)	18.57	–	–
Life Satisfaction	719	Vignette	2,503	(2.6, 2.3, 2.7)	7.37	6.54	0.24
Life Satisfaction 2	719	Vignette	2,503	(3.0, 2.2, 2.6)	11.13	8.26	1.4×10^{-6}
Height	719	Visual	2,503	(1.4, 2.1, 11.1)	16.53	6.22	0.09
Weight	719	Visual	2,503	(3.0, 1.6, 5.5)	8.33	6.38	3.7×10^{-3}

Notes: Diagnostics for CQ trios in UAS 571 (Survey 1) and UAS 719 (Survey 2). Column 1: percentage of responses ≤ 4 or ≥ 96 , in parentheses for Low, Medium, and High. Column 2: standard deviation of the Low, Medium, and High mean ratings. Column 3: standard deviation of residuals pooled across respondent-level regressions of CQ ratings on a constant and the vector of CQ means. Column 4: Linear Common Ranking Test, the p -value from a heteroskedasticity-robust (HC1) Wald test, in F form, of the joint null on the tested demographics; the dependent variable is the medium CQ rating minus the benchmark implied by the low and high CQs (target: the medium CQ's sample mean), and the regression includes all 25 default demographics, of which Other race/ethnicity and day-of-week indicators are controls excluded from the joint restriction. Anxiety items, ordered high-to-low in the raw data, are reordered here as Low, Medium, High. Remember is shown for completeness but is not a comparable trio because its middle item is an attention check. “–”: diagnostic not applicable with only two substantive CQs, which leave no residual degrees of freedom.

C.1 Top- and Bottom-Coding

Because our econometric model assumes no top- or bottom-coding of CQs, we designed our CQ trios to have as little of top/bottom-coding as reasonably possible (as described in Section I.B). As a diagnostic, we recommend reporting summary statistics on the frequency of responses near the endpoints of the response scale. To illustrate, for each of our CQ trios all of which had 0–100 scales, Column 1 reports the percentages of responses ≤ 4 or ≥ 96 separately for the Low, Medium, and High CQs, in that order. Because most of the CQs resulted from several iterations of survey design, data collection and revision, all CQs, except for the High CQ in the Height trio, are below 6.5% (and all three CQs in half of the trios are below 4.0%). Height, which went through fewer iterations, has 11.1% near-endpoint coding for its High CQ, of which almost all (10.9%) is top-coding. We recommend that anyone considering basing a Height CQ trio on ours modify the image for the High CQ to reduce top-coding.

C.2 Variation in Mean CQ Ratings

Our MM estimator has more statistical precision when there is more variation across the CQs' mean ratings (reported in Table L.3). Table C.1's Column 2 reports the standard deviation of each CQ trio's means. In the tradeoff between increasing this standard deviation and decreasing top/bottom-coding, we prioritized decreasing top/bottom-coding because it is required for the MM estimator to be consistent.

C.3 CQ Response Error Variance

Our MM estimator has more precision when CQ responses have smaller response error variances. Prediction errors from an estimated linear CQ translation function reflect the combination of response errors and deviations from linearity, but the prediction errors primarily reflect response errors when linearity is a good approximation. For a given respondent, the prediction errors are the residuals from a C -observation regression of the respondent's C CQ ratings on the CQ means (in our data, $C = 3$). As a metric for variation in CQ response errors, Table C.1's Column 3 reports the standard deviation of these CQ residuals pooled across respondents and CQs within a trio. Because only one residual degree of freedom remains per respondent (and assuming homoskedastic CQ errors), this pooled SD estimates $\sigma_\eta \sqrt{(C-2)/C}$ rather than σ_η itself; substituting $C = 3$ to back out an estimate of the implied response-error SD, the estimated $\hat{\sigma}_\eta$ is always below 17 points on our 0–100 scale. We report the raw pooled SD because its main use is comparative across trios.

C.4 Test of Linear Common Ranking of CQs (Assumption 1')

Linear Common Ranking is testable if respondents rate at least three CQs. To understand this test, imagine a scatterplot with the CQ means on the horizontal axis and respondents' ratings on the vertical axis. It has only three points. The vertical distance of the point for Medium from the line between the points for Low and High should be due solely to response error if Linear Common Ranking holds. Thus, that vertical distance (a positive number if above, negative if below) should be unpredictable. We implement this test using a modification of our MM estimator. We construct the “scale-use-only benchmark” using only the Low and High CQs, with the benchmark's target level set to the Medium CQ's sample mean. Thus, each respondent's benchmark is that respondent's predicted rating for the Medium CQ, assuming Linear Common Ranking holds. We then regress respondents' Medium CQ rating minus that benchmark on the demographics. Column 4 reports the p -value from a test of the joint null hypothesis that all the coefficients are zero.

For one trio, Remember, we cannot conduct the test because we have only two CQs (the third was used as an attention check). Of the remaining 14, we reject the null hypothesis in five:

Breathe ($p = 5.2 \times 10^{-14}$), Confidence ($p = 2 \times 10^{-22}$), Pain ($p = 2.3 \times 10^{-3}$), Life Satisfaction 2 ($p = 1.4 \times 10^{-6}$), and Weight ($p = 3.7 \times 10^{-3}$). The Breathe trio combines symptom severity with fitness and smoking cues that different demographic groups may weight differently, which is a plausible source of its rejection; it also requires the largest benchmark extrapolation in our data (Web Appendix Table L.3).

More important than whether we can statistically reject Linear Common Ranking is how good an approximation it is substantively. Table C.2 reports the coefficients for the regressions of respondents' Medium CQ rating minus their benchmark on the demographics. The coefficients can be used to assess the robustness of our MM scale-use-adjustment to the Linear Common Ranking assumption. As one robustness metric, we can calculate how the scale-use adjustment would differ if we based it on just the Low and High CQs, rather than all three: for a given right-hand-side variable, the difference equals the regression coefficient in Table C.2 multiplied by the weight on the Medium CQ in the MM's benchmark, shown in Table L.3. For example, the largest coefficient on a demographic dummy variable in Table C.2 is on Hispanic in the Weight regression: 4.19; the benchmark weight for the Medium Weight CQ is 0.325. So using all three Weight CQs for the benchmark instead of just Low and High adds $0.325 \times 4.19 = 1.36$ to the adjustment, meaning it subtracts 1.36 in the scale-use-adjusted regression—an estimate that is roughly equal to the standard error (1.28) around the Weight coefficient in the scale-use-adjusted regression.

Replicating the same analysis for *all* 18×5 coefficients in Table C.2 reveals that in 80 of the 90 cases, using all three CQs for the benchmark instead of just Low and High adds to the adjustment (and hence subtracts in the scale-use-adjusted regression) values that are smaller (in absolute value) than the relevant standard error, and often much smaller (e.g., by an order of magnitude).⁷ In only one of the ten exceptions is the subtracted value larger than twice the standard error.

C.5 Test of (Monotonic) Common Ranking of CQs (Assumption 1)

Testing the weaker assumption of Common Ranking is a bit more complex than testing Linear Common Ranking. As discussed in Section VII.C, the key testable implication of Common Ranking of CQs that is robust to heteroskedasticity is that the ranking of the CQ response means in a trio within demographic groups should be the same as the full-sample ranking. Column 3 of Table 4 reports the results. Except for the Weight trio, where 95% of the demographic groups share the common ranking, every other trio passes this test for 100% of the groups.

Column 5 of Table 4 reports a p -value testing the null that the group means satisfy Common

⁷The ten exceptions, reporting the subtracted (or added, when a minus sign appears) estimate followed by the coefficient's standard error, are: Breathe: Demeaned Age/10 0.38 (0.16), Male -0.69 (0.41), log (HH income) -0.38 (0.25), Black 1.17 (0.68), Asian 1.16 (0.78). Confidence: Male 0.49 (0.37), College grad 0.43 (0.40), Black -0.67 (0.66). Life Satisfaction 2: Male 0.85 (0.75). Weight: Hispanic 1.36 (1.28).

Table C.2: Medium-Benchmark Residual Regressions for Trios Failing the Linear Common-Ranking Test

Demographic variable	Breathe	Confidence	Pain	Life Satisfaction 2	Weight
Demeaned Age / 10	0.77 (0.16)	-0.33 (0.15)	0.05 (0.11)	-0.26 (0.31)	0.33 (0.23)
(Demeaned Age) ² / 100	0.05 (0.08)	0.01 (0.07)	-0.03 (0.05)	0.04 (0.15)	-0.11 (0.10)
Male	-1.40 (0.41)	1.95 (0.37)	0.98 (0.28)	4.10 (0.75)	0.57 (0.58)
Married, not separated	0.56 (0.50)	0.45 (0.45)	-0.25 (0.33)	-0.51 (0.86)	1.61 (0.67)
At least one child under 19 in HH	-0.47 (0.62)	-1.11 (0.55)	-0.58 (0.41)	-1.48 (1.12)	1.12 (0.89)
Log(HH size)	-0.19 (0.52)	-0.19 (0.48)	-0.37 (0.36)	-0.53 (0.95)	-1.76 (0.70)
Log(HH income)	-0.77 (0.25)	0.89 (0.24)	0.35 (0.18)	-0.19 (0.45)	0.22 (0.41)
College grad	-0.30 (0.45)	1.71 (0.40)	-0.14 (0.30)	0.61 (0.80)	0.26 (0.64)
Black/African American	2.38 (0.68)	-2.65 (0.66)	0.78 (0.51)	0.88 (1.13)	2.54 (0.98)
Hispanic/Latino/Spanish	1.11 (0.66)	-1.69 (0.63)	0.85 (0.46)	0.98 (1.48)	4.19 (1.28)
Asian	2.35 (0.78)	-2.42 (0.64)	-0.41 (0.50)	2.33 (1.57)	-0.64 (1.14)
Part-time employed	-0.14 (0.67)	0.49 (0.63)	-0.08 (0.45)	-0.78 (1.28)	-0.57 (0.98)
Unemployed	-1.03 (0.89)	0.68 (0.80)	-0.32 (0.67)	0.63 (1.79)	-1.11 (1.46)
Other employment/Out of labor force	-0.07 (0.59)	0.82 (0.52)	0.28 (0.40)	-1.72 (1.01)	-0.16 (0.81)
Northeast	-0.02 (0.74)	-0.54 (0.64)	-1.14 (0.49)	-0.11 (1.14)	-0.61 (0.88)
South	-0.34 (0.58)	-0.29 (0.51)	-0.46 (0.39)	-1.42 (0.92)	-1.32 (0.71)
West	-1.15 (0.60)	-0.15 (0.52)	-0.58 (0.39)	-2.45 (1.05)	-1.55 (0.84)
Top population-density quartile	0.12 (0.48)	-0.42 (0.43)	-0.17 (0.33)	-0.79 (0.78)	-0.07 (0.63)
Observations	10,100	10,100	10,100	2,503	2,503
R ²	0.010	0.019	0.005	0.026	0.017
F-test based on the above regressors	$F(18, 10074) = 5.77$ $p = 5.2 \times 10^{-14}$	$F(18, 10074) = 8.26$ $p = 2 \times 10^{-22}$	$F(18, 10074) = 2.21$ $p = 2.3 \times 10^{-3}$	$F(18, 2477) = 3.42$ $p = 1.4 \times 10^{-6}$	$F(18, 2477) = 2.13$ $p = 3.7 \times 10^{-3}$

Notes: Coefficients from unweighted OLS regressions, with heteroskedasticity-robust (HC1) standard errors in parentheses. Dependent variable: the medium CQ rating minus the benchmark implied by the low and high CQs, i.e., low CQ + $\varpi \times$ (high CQ – low CQ), where ϖ equates the benchmark mean to the sample mean of the medium CQ. Regressions include the variables shown, Other race/ethnicity, day-of-week indicators, and a constant.

Ranking of CQs, computed as a generalized likelihood ratio test, which we now describe. Index the demographic groups by g , let $n_g \geq 50$ be the number of respondents in group g , and let $\bar{\mathbf{y}}_g$ be its vector of mean CQ ratings. By the central limit theorem, $\bar{\mathbf{y}}_g$ is approximately normal with covariance matrix $\mathbf{V}_g$, the group's sample covariance matrix divided by n_g . The null hypothesis is that each group's population mean CQ vector lies in the cone of vectors respecting the full-sample ranking, with every group free to have its own mean vector and its own covariance matrix. The unconstrained maximum likelihood estimate of a group's mean vector is $\bar{\mathbf{y}}_g$ itself; the constrained estimate $\hat{\boldsymbol{\mu}}_g$ is its projection onto the cone of vectors respecting the full-sample ranking in the metric $\mathbf{V}_g^{-1}$. Twice the log-likelihood difference is, after omitting constants, $\text{LR}_g = (\bar{\mathbf{y}}_g - \hat{\boldsymbol{\mu}}_g)' \mathbf{V}_g^{-1} (\bar{\mathbf{y}}_g - \hat{\boldsymbol{\mu}}_g)$, and the test statistic is $T = \sum_g \text{LR}_g$. A group whose mean ratings already respect the full-sample ranking contributes zero to T , so $T = 0$ —and hence whenever every group respects it, $p = 1.00$. Because the restriction consists of inequalities, T is not asymptotically distributed χ^2 , so its reference distribution is obtained by parametric bootstrap: 2,000 samples are drawn from the constrained estimates $\hat{\boldsymbol{\mu}}_g$ with covariance $\mathbf{V}_g$, and both the constrained and unconstrained fits are recomputed in each.

We note that, in the anchoring vignette literature (e.g., King et al., 2004; Bago d'Uva et al., 2011), tests have been proposed that we do *not* recommend: examining whether disagreement with

the full-sample ranking is predictable by observed demographics. The problem with these tests is that heteroskedasticity of response errors can generate such predictability: demographic groups with larger response error variances will have more respondents who deviate from the full-sample ranking. To examine this issue, we regressed an indicator for failing the strict Low < Medium < High ranking on the 25 default demographics. Much of the predictability we find is plausibly consistent with heteroskedasticity. For example, for all 14 CQ trios, being a college graduate is associated with less disagreement with the full-sample ranking, consistent with college graduates giving less noisy survey responses.

C.6 Test for Linear CQ Translation Functions

Section II.B’s Figure 4 illustrates some diagnostics for assessing linearity of the average translation function across demographic groups: correlation of mean CQ responses across groups and p -value from an F -test of the joint null hypothesis that the TLS regression line’s (intercept, slope) equals (0,1), which tests whether the two groups use the scale identically rather than whether the relationship is linear. In Web Appendix E, we describe and report tests of linearity at the individual level, but we can only conduct these analyses in datasets where each respondent answers at least four CQs.

D CQ-Only Maximum Likelihood Estimation

Some analyses require parameters of the distribution of scale-use parameters and of response error that cannot be recovered from the respondent-level OLS estimates $\hat{a}_i$ and $\hat{b}_i$ because those estimates are contaminated by response error. We estimate these parameters by maximum likelihood, using only respondents’ CQ responses. Estimating scale use from CQ data alone keeps it from being confounded with variation in the common-scale levels of the SR dimension, an issue we discuss in Appendix H in the context of CHOPIT. This appendix describes the baseline (linear) specification and the two extensions used elsewhere in the paper and closes by describing how the resulting parameters feed into the validation analysis reported in Section VI. Throughout this and the subsequent appendices, we write $\mathbb{C} \subseteq \mathcal{C}$ for the particular set of CQs used in a given estimation, so that $C = |\mathbb{C}|$, and we subscript by $\mathbb{C}$ any vector with one element per CQ in that set: $\mathbf{w}_{\mathbb{C}}$ collects those CQs’ common-scale levels, $\mathbf{r}_{i\mathbb{C}}$ respondent i ’s ratings of them, and $\boldsymbol{\theta}_{\mathbb{C}}$ a vector of weights on them.

D.1 Linear Translation Function

For respondent i with CQ rating r_{ic} , the linear CQ translation function in equation (2) is

$$r_{ic} = a_i + b_i w_c + \eta_{ic}.$$

For the MLE, we assume $(a_i, \ln b_i)$ is jointly normal with an unrestricted variance-covariance matrix and a mean vector consistent with the normalizations of our econometric model (Section IV)— $\mathbb{E}(a_i) = 0$ and $\mathbb{E}(b_i) = 1$, the latter implying $\mathbb{E}(\ln b_i) = -\sigma_{\ln b}^2/2$. We assume the response errors η_{ic} are independent across CQs and normally distributed with zero mean and a respondent-specific variance $\sigma_{\eta_i}^2$, where $\ln \sigma_{\eta_i}$ is normally distributed across respondents with mean $\mu_{\ln \sigma_\eta}$ and variance $\sigma_{\ln \sigma_\eta}^2$. This structure lets the response-error variance differ across respondents while holding it constant across the CQs a given respondent answers. The log-likelihood of each respondent's CQ responses is evaluated by hierarchically integrating over the joint distribution of (a_i, b_i) and over σ_{η_i} ; maximizing the likelihood yields consistent estimates of the standard deviations of a_i and b_i , their covariance, and the two parameters of the response-error distribution. In our Survey 1 data, the estimated standard deviations are 30.44 for a_i and 0.38 for b_i .

We estimate two versions of this specification that differ only in the distribution of b_i . The baseline assumes $(a_i, \ln b_i)$ is jointly normal, ensuring $b_i > 0$, as its stretcher interpretation requires. An otherwise identical version, with the same number of parameters, assumes that (a_i, b_i) is jointly normal. Evaluated at the estimates, the latter version fits better in each of the four CQ dimensions in which we compare the two—which span vignette and visual CQs and both surveys—by 629 log points for Life Satisfaction, 261 for No Anxiety, 344 for Weight, and 158 for Height. A substantial share of this gain remains when we restrict the sample to respondents who rate the highest CQ above the lowest, consistent with $b_i > 0$ (98 of 158 log points for Height, 276 of 344 for Weight, 126 of 261 for No Anxiety, and 234 of 629 for Life Satisfaction), indicating that the better fit is not merely an artifact of the normal distribution's ability to accommodate negative stretchers.

Which version we use is determined by how b_i enters the calculation. The semiparametric estimators in Appendix F.1 depend on the full distribution of b_i and therefore use the log-normal version, since a negative stretcher has no interpretation. Calculations that involve b_i only through $\mathbb{E}(b_i^2)$ use the better-fitting normal version. These include the scale-use-corrected standard deviations in Table 3 and Web Appendix Table L.6. This choice matters little in practice. For example, for Life Satisfaction, $\hat{\sigma}_b$ is 0.38 under both versions. Anxiety, whose CQ ratings come closest to order-reversing, shows the largest discrepancy (0.86 under the normal version versus 0.43 under the log-normal version).

Four sets of estimates from this analysis are used elsewhere in the paper. First, the moments $\mathbb{E}(b_i^k)$ enter both the semiparametric estimators of Appendix F.1 and the variance estimator described at the end of this appendix. Second, because response errors are normal conditional on σ_{η_i} , an

unbiased estimated stretcher constructed as a linear combination of a respondent's CQ ratings is normally distributed around the true stretcher with a variance $\theta'_{bC} \theta_{bC} \sigma_{\eta_i}^2$, so the MLE delivers an estimate of this variance; combining this variance estimate with an estimate of the cross-respondent variance $\sigma_b^2 = \text{Var}(b_i)$ and the MLE's distributional assumptions is what allows the semiparametric estimators to construct the conditional moments $\mathbb{E}(b_i^k | \tilde{b}_i)$ by numerical integration. Third, the estimate of $\mathbb{E}(\sigma_{\eta_i}^2)$ is used to net response error out of quantities built from the CQ benchmark (Web Appendix D.4). Fourth, the dimension-specific estimate of $\mathbb{E}(\sigma_{\eta_i}^2)$ is used in Appendix E to correct the dispersion of respondent-level estimates for response error.

D.2 Quadratic Translation Function

Appendix E assesses the linearity assumption by augmenting the translation function with a quadratic term, $r_{ic} = a_i + b_i w_c + d_i w_c^2 + \eta_{ic}$, as in equation (7). Because the standard deviation of respondent-level OLS estimates of d_i is biased away from zero by response error, to obtain a response-error-corrected estimate of σ_d we estimate a CQ-only MLE under this quadratic translation function. We assume that $(a_i, \ln b_i, d_i)$ is jointly normal with a mean vector consistent with the model normalizations ($\mathbb{E}(a_i) = 0, \mathbb{E}(b_i) = 1, \mathbb{E}(d_i) = 0$) and a free variance-covariance matrix, and we assume, as in the baseline specification above, that the response errors η_{ic} are independent and normally distributed with zero mean and individual-specific variance $\sigma_{\eta_i}^2$, where $\ln \sigma_{\eta_i}$ is normally distributed across respondents. The log-likelihood of each respondent's CQ responses is evaluated by hierarchically integrating over the joint distribution of (a_i, b_i, d_i) and over σ_{η_i} ; maximizing this likelihood yields consistent estimates of σ_d as well as the standard deviations of a_i and b_i and their pairwise covariances. Imposing $d_i = 0$ for all i recovers the baseline specification of Appendix D.1.

D.3 Two-Wave Specification and the Persistence of Scale Use

To gauge the persistence of the intercept and stretcher, we extend the baseline specification to accommodate two waves of CQ ratings. Using the sample of about 2,500 respondents from whom we collected retest data (in Survey 2), we model the CQ ratings in wave $t \in \{1, 2\}$ as

$$r_{ic,t} = a_{i,t} + b_{i,t} w_c + \eta_{ic,t},$$

where $(a_{i,t}, b_{i,t})$ are wave-specific scale-use parameters and $\eta_{ic,t}$ is response error. We assume that $(a_{i,1}, a_{i,2}, \ln b_{i,1}, \ln b_{i,2})$ is jointly normal with a free correlation matrix. The variances of a , b , and η remain constant across the waves. The response error is independent across individuals and CQs but may be correlated across waves. The log-likelihood for each respondent is evaluated by hierarchically integrating over this joint distribution; maximizing the resulting likelihood yields estimates of all variances and covariances of $(a_{i,1}, a_{i,2}, \ln b_{i,1}, \ln b_{i,2})$.

In this model, the correlation between $a_{i,1}$ and $a_{i,2}$, which we estimate to be 59.0% (SE = 4.5%), measures the fraction of the intercept’s variance that is persistent across the two waves. Under a permanent-plus-transitory decomposition $a_{i,t} = \alpha_i + u_{i,t}$, with $u_{i,t}$ independent across waves and $\text{Var}(a_{i,t})$ constant, this correlation equals the share of the intercept’s variance that is permanent. The correlation between $b_{i,1}$ and $b_{i,2}$, which we estimate to be 84.8% (SE = 3.2%), measures the fraction of the stretcher’s variance that is persistent. These results indicate that a substantial share—but far from all—of the variance in scale-use parameters reflects persistent individual differences. The stretcher is considerably more persistent than the intercept. The finding that roughly 41% of the intercept’s variance is non-persistent has an important implication for panel-data applications: individual fixed effects, which absorb only the time-invariant component of scale use, leave a sizable transitory component of the intercept uncorrected—in addition to leaving the stretcher entirely uncorrected.

D.4 Parameters for the Validation Exercise

The linear specification supplies most of the parameters needed for the validation exercise of Section VI, which requires the standard deviation of the subjective measure in common-scale units. Assuming independence⁸ between the stretcher b_i and the common-scale level $w_i^\star$, we derive the estimator as follows. Because the benchmark weights sum to one, subtracting the CQ benchmark from the SR differences out the intercept:

$$r_i^\star - \theta'_{sC} \mathbf{r}_{iC} = b_i (w_i^\star - \theta'_{sC} \mathbf{w}_C) + \eta_i^\star - \theta'_{sC} \boldsymbol{\eta}_{iC}.$$

Taking the variance—using independence between b_i and $w_i^\star$ for the first term, and independence of the response errors from everything else for the other two—therefore gives

$$\text{Var}(r_i^\star - \theta'_{sC} \mathbf{r}_{iC}) = \mathbb{E}(b_i^2) \text{Var}(w_i^\star) + [\mathbb{E}(r_i^\star) - \theta'_{sC} \mathbf{w}_C]^2 \left[\mathbb{E}(b_i^2) - 1 \right] + \sigma_{\eta_s}^2 + \theta'_{sC} \boldsymbol{\theta}_{sC} \mathbb{E}(\sigma_{\eta_i}^2),$$

where $\sigma_{\eta_s}^2 = \text{Var}(\eta_i^\star)$. Solving for $\text{Var}(w_i^\star)$ gives the MM estimator of $\text{Var}(w_i^\star)$ as the sample analog of

$$\frac{1}{\mathbb{E}(b_i^2)} \text{Var}(r_i^\star - \theta'_{sC} \mathbf{r}_{iC}) - [\mathbb{E}(r_i^\star) - \theta'_{sC} \mathbf{w}_C]^2 \left(1 - \frac{1}{\mathbb{E}(b_i^2)} \right) - \frac{\sigma_{\eta_s}^2 + \theta'_{sC} \boldsymbol{\theta}_{sC} \mathbb{E}(\sigma_{\eta_i}^2)}{\mathbb{E}(b_i^2)},$$

where $\mathbb{E}(b_i^2)$ and $\sigma_{\eta_i}^2$ are estimated by the CQ-only MLE. The SR response-error variance $\sigma_{\eta_s}^2$, however, is not identified from CQ data, so the validation exercise additionally requires an assumption about SRs: as a baseline assumption, we set $\sigma_{\eta_s}^2$ equal to 0.5 times $\mathbb{E}(\sigma_{\eta_i}^2)$, on the view

⁸The semiparametric estimator in Appendix F relaxes the independence assumption. We employ this simpler estimator in the validation exercise because correcting for scale-use heterogeneity in variances plays a subsidiary role in this paper.

that respondents evaluate their own circumstances more precisely than they evaluate the situations of other people in the CQs. Web Appendix Table L.6 reports results under alternative proportions of 0, 0.25, and 0.75; the qualitative ordering of the same-dimension validation results is unchanged, while a proportion of 0.8 produces negative corrected variances, indicating that ratios of 0.8 or above are incompatible with our model at the estimated nuisance parameters, which is why we report the 0–75% range.

E Evidence on the Linearity of the CQ Translation Function

To gauge the plausibility of the linearity assumption maintained in equation (2), we augment the CQ translation function with a quadratic term:

$$r_{ic} = a_i + b_i w_c + d_i w_c^2 + \eta_{ic}. \quad (7)$$

To be consistent with w_c 's definition as the population mean of respondents' ratings of CQ c , we normalize $\mathbb{E}(d_i) = 0$. We do not restrict the correlation of the coefficient d_i with (a_i, b_i) , and we denote its standard deviation by σ_d . All other assumptions on the scale-use parameters and response error carry over from the linear model. The null hypothesis of linearity corresponds to $d_i = 0$ for all i , or equivalently $\sigma_d = 0$.

E.1 Individual-Level Tests for Quadratic Translation Functions

To conduct an individual-level test of $d_i = 0$, at least four CQ responses per respondent are required: three to estimate a_i , b_i , and d_i , and at least one to provide a degree of freedom for the test. Our Survey 1 includes only three Life Satisfaction vignette CQs per respondent, which is insufficient. We therefore draw on two additional sources of data. First, we use Survey 2, which administered six Life Satisfaction vignette CQs to 2,537 respondents. Second, we use three previously published datasets that contain more than three vignette CQs per respondent: the political-efficacy vignette data of King et al. (2004) ($N = 834$, five CQs on a 1–5 scale), the health vignette data of Kapteyn et al. (2007) ($N = 786$, fifteen CQs on a 1–5 scale), and the life satisfaction vignette data of Montgomery (2022) ($N = 45,253$, six CQs on a 0–10 scale). Notably, the latter three datasets employ Likert scales with a small number of response categories, allowing us to assess whether linearity is a reasonable approximation even in such settings.

For each respondent in each dataset, we estimate d_i by OLS regression of r_{ic} on $\hat{w}_c$, $\hat{w}_c^2$, and a constant (where, as in Section IV, $\hat{w}_c$ is the sample mean rating of CQ c). Table E.1 reports the fraction of respondents who have an estimate of d_i statistically distinguishable from zero at the 5% level. Across all four datasets, the raw fraction ranges from 4.4% to 8.4%, close to the 5% expected

under the null. After applying a false-discovery-rate (FDR) correction via the Benjamini-Hochberg algorithm, no respondent in any dataset retains a significant quadratic coefficient. We therefore cannot reject linearity for any individual respondent. As the next subsection shows, this is consistent with curvature that is precisely estimated at the population level but too small to detect respondent by respondent and quantitatively negligible over the relevant range of the scale.

Table E.1: Tests of the Quadratic Coefficient

	Survey 2 (UAS 719)	King et al. (2004)	Kapteyn et al. (2007)	Montgomery (2022)
Dimension	Life Satisfaction	Political Efficacy	Health	Life Satisfaction
% quadratic coefficients with $p < 0.05$	4.4%	6.1%	8.4%	4.8%
% quadratic coefficients with FDR < 0.05	0%	0%	0%	0%
$\hat{\sigma}_d$	0.00323 (0.000154)	0.0381 (0.00147)	0.0342 (0.00181)	0.0369 (0.000194)

Notes: d_i : respondent i 's coefficient on $\hat{w}_c^2$ from an OLS regression of their CQ ratings r_{ic} on a constant, $\hat{w}_c$, and $\hat{w}_c^2$, where $\hat{w}_c$ is the CQ's sample mean rating. First row: percentage of respondents for whom the test of $H_0 : d_i = 0$ yields $p < 0.05$. Second row: percentage still significant after Benjamini-Hochberg FDR control at 5% within each dataset. Third row: $\hat{\sigma}_d$, the CQ-only maximum-likelihood estimate of the cross-respondent standard deviation of d_i (Appendix D.2), accounting for response error, with standard errors in parentheses.

E.2 Estimation of Standard Deviation of Quadratic Coefficient

The standard deviation of respondent-level OLS estimates of d_i is biased away from zero by response error. To obtain a response-error-corrected estimate of σ_d , we conduct a CQ-only maximum likelihood estimation under the quadratic translation function in equation (7), as described in Appendix D.2. The last row of Table E.1 reports the estimate of σ_d for each dataset (with standard errors in parentheses).

The estimated σ_d 's are small across all datasets, implying little curvature in ratings after allowing a linear approximation to absorb changes in the intercept and slope. One summary statistic is the maximum deviation between the quadratic curve and the line connecting the endpoints of an interval of length Δ in w_c . For respondent i , this deviation is $|d_i|\Delta^2/4$. In Survey 2 data ($\hat{\sigma}_d = 0.00323$), a 10-point interval in w_c therefore implies a maximum deviation of about 0.16 raw-rating points for 95% of respondents, using $|d_i| \leq 2\hat{\sigma}_d$. For roughly comparable 10% intervals on each dataset's respective scale, the corresponding bounds are about 0.003 points on the 1–5 scales of King et al. (2004) ($\hat{\sigma}_d = 0.0381$) and Kapteyn et al. (2007) ($\hat{\sigma}_d = 0.0342$) and about 0.018 points on the 0–10 scale of Montgomery (2022) ($\hat{\sigma}_d = 0.0369$). Together, these findings indicate that deviations from linearity are quantitatively small over the relevant ranges of the CQ scales.

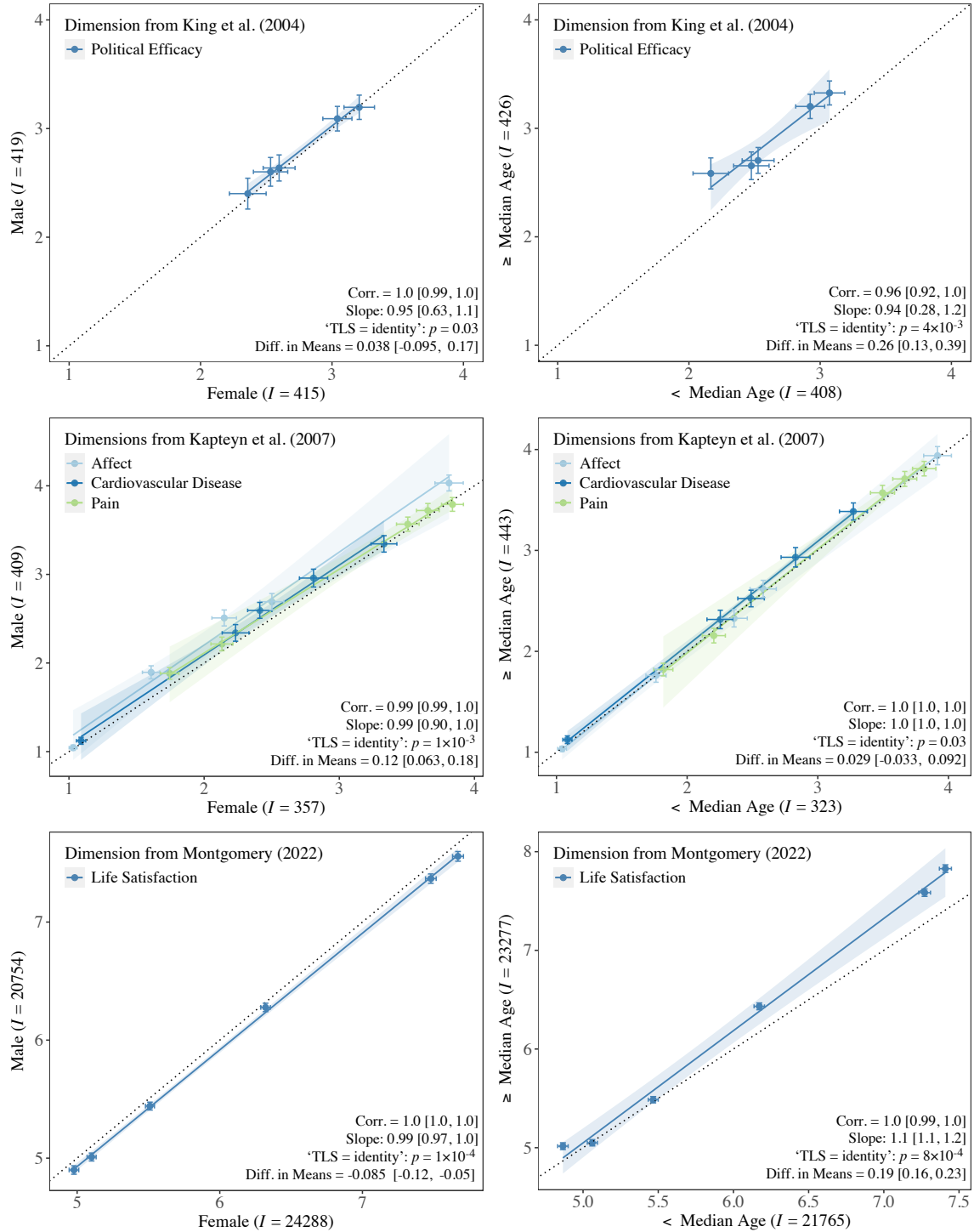
E.3 Cross-Group Average Translation Functions for Likert Scales

To visually evaluate whether linearity is a reasonable approximation for scales with limited response options, like the Likert scales—as used by King et al. (2004), Kapteyn et al. (2007), and Montgomery

(2022)—we provide a graphical test analogous to Figure 4 in the main text. For each dataset, we plot the average CQ ratings for groups defined by gender and age, and we overlay a TLS line. As Figure E.1 shows, the group-level translation functions appear to be well fit by the line despite the restricted number of response categories.

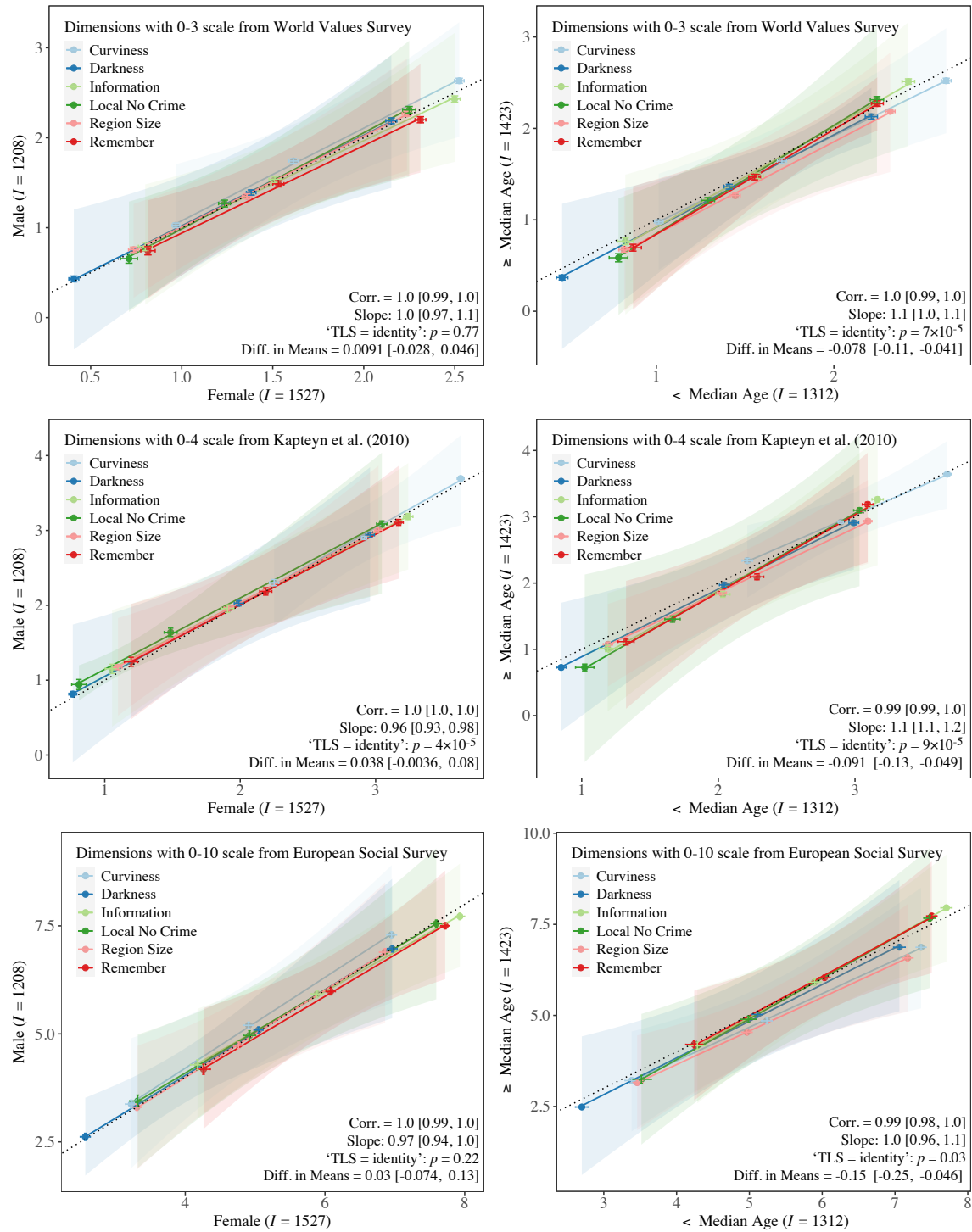
We conduct a similar exercise using MTurk data we collected ($I = 2,735$; for details, see the “Bottomless survey” in Benjamin et al., 2023), in which respondents answered CQs on alternative response scales drawn from existing surveys: the World Values Survey (0–3 scale), Kapteyn et al. (2010) (0–4 scale), and the European Social Survey (0–10 scale). As Figure E.2 shows, the group-level translation functions are again well approximated by straight lines on these alternative scales.

Figure E.1: Translation Functions Across Demographic Groups on Likert-Like Scales



Notes: Each panel: average CQ rating for one group vs. average CQ rating for another, for groups defined by gender and age. Each point: one CQ. TLS fit and within-panel correlation reported in each panel. Sources and scales: King et al. (2004) (political efficacy, 1–5 scale), Kapteyn et al. (2007) (health, 1–5 scale), Montgomery (2022) (life satisfaction, 0–10 scale).

Figure E.2: Translation Functions Across Demographic Groups on Alternative Scales in MTurk Data



Notes: Each panel: average CQ rating for one group vs. average CQ rating for another, for groups defined by gender and age. Each point: one CQ. TLS fit and within-panel correlation reported in each panel. Source: Bottomless survey.

F Semiparametric Estimators and Evaluating Approximation Error of MM Estimator

To evaluate the approximation error of the MM estimator, we use the covariance-with-regressors member of a broader family of semiparametric estimators. The same construction also yields estimators for means, variances, and covariances between SRs. Unlike MM, these estimators do *not* assume independence between the stretcher and the common-scale level and do *not* assume zero coskewness. We derive the family below but implement only the semiparametric regression estimator because regression is the application of interest in the paper.

F.1 Semiparametric Estimators

Means and covariances with observed variables. Since $\text{Cov}(w_i^*, x_i) = \mathbb{E}(w_i^* x_i) - \mathbb{E}(w_i^*) \mathbb{E}(x_i)$, we focus on detailing the estimation of $\mathbb{E}(w_i^* x_i)$ in this section because $\mathbb{E}(w_i^*)$ is a special case of $\mathbb{E}(w_i^* x_i)$ (with $x_i = 1$ for all i), and $\mathbb{E}(x_i)$ can be estimated directly from data.

To estimate $\mathbb{E}(w_i^* x_i)$, we non-parametrically model this conditional expectation:

$$\mathbb{E}((w_i^* - \theta'_C \mathbf{w}_C) x_i | b_i) = A_0 + A_1 b_i + A_2 b_i^2 + \cdots + A_K b_i^K,$$

where the A_k 's are unknown parameters that we will estimate. In principle this polynomial expansion can be made an arbitrarily good approximation by taking K large. In practice, we use $K = 1$, which works well in our simulations.

As inputs into the MM estimator, we construct stretcher estimates $\hat{b}_i$ and scale-use-only benchmarks using a respondent-level OLS regression of respondents' CQ responses on mean CQ responses. For consistency of the MM estimator, the benchmark had to target $a_i + b_i \mathbb{E}(w_i^*)$, which the MM estimator approximates by evaluating the estimated translation function at $\hat{\mathbb{E}}(r_i^*)$. The semiparametric estimator, which models the dependence on the stretcher directly, does not need this target for its benchmark. Its benchmark $\theta'_C \mathbf{r}_{iC}$ is still scale-use-only, estimating $a_i + b_i \theta'_C \mathbf{w}_C$, but the level $\theta'_C \mathbf{w}_C$ is now just a recentering: it is subtracted from w_i^* in the polynomial approximation above and added back exactly below, so any value would do. We use $\mathbb{E}(r_i^*)$ because it is estimable. Instead of the benchmark targeting a particular value, what matters here is that the response errors in the estimated stretcher and benchmark be orthogonal. To achieve this, we construct an alternative (but still unbiased) estimated stretcher $\tilde{b}_i \equiv \theta'_{bC} \mathbf{r}_{iC}$ and benchmark $\theta'_C \mathbf{r}_{iC}$ using these weight vectors:

$$\theta'_C = \begin{bmatrix} \frac{w_{c3} - \mathbb{E}(r_i^*)}{w_{c3} - w_{c1}} & 0 & \frac{\mathbb{E}(r_i^*) - w_{c1}}{w_{c3} - w_{c1}} \end{bmatrix}$$

and

$$\theta_{bC} = \frac{w_{c3} - \mathbb{E}(r_i^*)}{(2w_{c2} - w_{c1} - w_{c3}) \mathbb{E}(r_i^*) + w_{c3}^2 + w_{c1}^2 - w_{c1}w_{c2} - w_{c2}w_{c3}} \begin{bmatrix} \frac{\mathbb{E}(r_i^*) - w_{c1}}{w_{c3} - \mathbb{E}(r_i^*)} \\ \frac{w_{c1} + w_{c3} - 2\mathbb{E}(r_i^*)}{w_{c3} - \mathbb{E}(r_i^*)} \\ 1 \end{bmatrix},$$

where w_{c1} , w_{c2} , and w_{c3} denote the low, medium, and high common-scale levels of the CQs, respectively. Note that this construction requires $w_{c3} \neq \mathbb{E}(r_i^*)$ and a nonzero denominator in θ_{bC} , both of which hold in our data. (Without harming consistency, a different value can be used in place of $\mathbb{E}(r_i^*)$ if either of these nonzero denominator conditions does not hold.)

To estimate the A_k 's, we observe that

$$\begin{aligned} \mathbb{E}((r_i^* - \theta'_C \mathbf{r}_{iC}) x_i | \tilde{b}_i) &= \mathbb{E}((b_i w_i^* + \eta_i^* - b_i \theta'_C \mathbf{w}_C - \theta'_C \boldsymbol{\eta}_{iC}) x_i | \tilde{b}_i) \\ &= \mathbb{E}(b_i (w_i^* - \theta'_C \mathbf{w}_C) x_i | \tilde{b}_i) \\ &= \mathbb{E}(b_i \mathbb{E}((w_i^* - \theta'_C \mathbf{w}_C) x_i | b_i, \tilde{b}_i) | \tilde{b}_i) \\ &= \mathbb{E}(b_i \mathbb{E}((w_i^* - \theta'_C \mathbf{w}_C) x_i | b_i) | \tilde{b}_i) \\ &= \mathbb{E}\left(b_i \left(A_0 + A_1 b_i + A_2 b_i^2 + \dots\right) | \tilde{b}_i\right) \\ &= A_0 \mathbb{E}(b_i | \tilde{b}_i) + A_1 \mathbb{E}(b_i^2 | \tilde{b}_i) + A_2 \mathbb{E}(b_i^3 | \tilde{b}_i) + \dots, \end{aligned} \quad (8)$$

where the second equality holds because the weight vectors are constructed to be orthogonal, $\theta'_C \theta_{bC} = 0$, and because the SR's and the CQs' response errors are independent.⁹ The fourth equality holds because of the independence between (w_i^*, x_i) and the response error.

We estimate the A_k 's by running the regression implied by expression (8). To do so, we construct estimates of the regressors, $\mathbb{E}(b_i^k | \tilde{b}_i)$, by numerically integrating b_i^k with respect to the estimated density of $f(b_i | \tilde{b}_i)$. The density can be found by applying Bayes' rule to $\tilde{b}_i | b_i$, whose distribution is normal (given normal response error) with mean b_i and variance estimable from a CQ-only maximum likelihood estimator. To operationalize the MLE, we use the baseline linear CQ-only MLE described in Appendix D.1. That is, we assume $(a_i, \ln b_i)$ is jointly normal with mean vector $(0, -\sigma_{\ln b}^2/2)$ —so that $\mathbb{E}(a_i) = 0$ and $\mathbb{E}(b_i) = 1$ —and a general variance-covariance matrix, and we assume that response errors are normally distributed with zero mean and individual-specific variance.

⁹Orthogonality of the weights makes the benchmark's response error $\theta'_C \boldsymbol{\eta}_{iC}$ and the stretcher estimation error $\tilde{b}_i - b_i = \theta'_{bC} \boldsymbol{\eta}_{iC}$ conditionally uncorrelated. And because both are linear combinations of a conditionally normal vector, zero covariance implies conditional independence. The variance estimator below achieves conditional independence by different means, pairing a uniformly weighted benchmark with the OLS estimator $\hat{b}_i$, whose weights sum to zero. Both arguments for conditional independence rely on assuming that the η_{ic} are independent across CQs with a common variance for respondent i . If a researcher wishes to avoid making some of our response-error assumptions and has two sets of CQs that each individually identify the scale-use parameters, conditional independence can be guaranteed by constructing the benchmark and $\hat{b}_i$ from disjoint sets of CQs, at the cost of a loss in precision.

The CQ-only MLE also allows us to estimate $\mathbb{E}(b_i^k)$, which together with the estimates of the A_k 's, enables estimating $\mathbb{E}(w_i^\star x_i)$ by plugging in the estimates for their population counterparts into

$$\mathbb{E}(w_i^\star x_i) = A_0 + A_1 \mathbb{E}(b_i) + \cdots + A_K \mathbb{E}(b_i^K) + \theta'_C \mathbf{w}_C \mathbb{E}(x_i).$$

Setting $x_i = 1$ throughout gives the semiparametric estimator of $\mathbb{E}(w_i^\star)$. Combining this estimate with the corresponding estimate of $\mathbb{E}(w_i^\star x_i)$ gives

$$\text{Cov}(w_i^\star, x_i) = \mathbb{E}(w_i^\star x_i) - \mathbb{E}(w_i^\star) \mathbb{E}(x_i).$$

For a vector of regressors, we apply this calculation component by component and premultiply the resulting covariance vector by the sample analog of $\text{Var}(\mathbf{x}_i)^{-1}$ to obtain the semiparametric regression coefficients.

Variances and covariances between SRs. For covariances between SRs, we use the notation d and d' for the two dimensions involved. For both the MM and semiparametric estimators above, the CQ benchmark is targeted to the population mean SR in the sense that its weights, θ_C , imply that $\theta'_C \mathbf{w}_C = \mathbb{E}(r_i^\star)$. Since the common-scale CQ ratings $\mathbf{w}_C$ are determined by the choice of CQs, the benchmark weights are generally nonuniform. For a covariance between two SRs, these question-specific benchmarks need not coincide because the SRs can have different means. To avoid additional complexity, for estimating covariance we instead use the same benchmark for both questions and choose it to be the uniformly weighted CQ average, so $\theta_{dC} = \theta_{d'C} = C^{-1} \mathbf{1}_C$. Under our response-error assumption stated below, this choice also makes the benchmark error orthogonal to the estimation error in the respondent-level OLS estimator $\hat{b}_i$ of the stretcher. The specially constructed estimator $\tilde{b}_i$ used above is therefore unnecessary here.

Our response-error assumption is that, conditional on b_i , the CQ response errors $\eta_{i1}, \dots, \eta_{iC}$ are independent normal variables with mean zero and the same variance for respondent i . We also assume that SR response errors are mean zero and mutually uncorrelated across SR dimensions conditional on b_i and $\mathbf{x}_i$. The response error in the uniformly weighted benchmark is their simple average, $C^{-1} \sum_{c=1}^C \eta_{ic}$, whereas $\hat{b}_i - b_i$ is a weighted sum of the same errors whose weights sum to zero. These two errors therefore have zero covariance conditional on b_i . Because both are linear combinations of the conditionally normal vector $\boldsymbol{\eta}_{iC}$, they are jointly normal conditional on b_i ; zero covariance therefore implies conditional independence.

As in the first-moment estimator, we approximate the relevant conditional expectation by a polynomial in b_i :

$$\mathbb{E}[(w_{id}^\star - \theta'_{dC} \mathbf{w}_C)(w_{id'}^\star - \theta'_{d'C} \mathbf{w}_C) \mid b_i] = A_{0dd'} + A_{1dd'} b_i + \cdots + A_{Kdd'} b_i^K.$$

To see the observable implication of this approximation, subtract the CQ benchmark from each SR

and multiply the resulting quantities:

$$\begin{aligned} & (r_{id}^* - \theta'_{d\mathbb{C}} \mathbf{r}_{i\mathbb{C}}) (r_{id'}^* - \theta'_{d'\mathbb{C}} \mathbf{r}_{i\mathbb{C}}) \\ &= [b_i (w_{id}^* - \theta'_{d\mathbb{C}} \mathbf{w}_{\mathbb{C}}) + (\eta_{id}^* - \theta'_{d\mathbb{C}} \boldsymbol{\eta}_{i\mathbb{C}})] [b_i (w_{id'}^* - \theta'_{d'\mathbb{C}} \mathbf{w}_{\mathbb{C}}) + (\eta_{id'}^* - \theta'_{d'\mathbb{C}} \boldsymbol{\eta}_{i\mathbb{C}})]. \end{aligned}$$

After taking conditional expectations, the cross-terms between common-scale levels and response errors vanish. The product nevertheless contains a mechanical contribution from response error,

$$(\eta_{id}^* - \theta'_{d\mathbb{C}} \boldsymbol{\eta}_{i\mathbb{C}}) (\eta_{id'}^* - \theta'_{d'\mathbb{C}} \boldsymbol{\eta}_{i\mathbb{C}}),$$

because response error enters each benchmarked SR and the same noisy CQ benchmark is subtracted from both. Subtracting the conditional expectation of this response-error product gives

$$\begin{aligned} & \mathbb{E} [(r_{id}^* - \theta'_{d\mathbb{C}} \mathbf{r}_{i\mathbb{C}}) (r_{id'}^* - \theta'_{d'\mathbb{C}} \mathbf{r}_{i\mathbb{C}}) | \hat{b}_i] - \mathbb{E} [(\eta_{id}^* - \theta'_{d\mathbb{C}} \boldsymbol{\eta}_{i\mathbb{C}}) (\eta_{id'}^* - \theta'_{d'\mathbb{C}} \boldsymbol{\eta}_{i\mathbb{C}}) | \hat{b}_i] \\ &= \sum_{k=0}^K A_{kd d'} \mathbb{E} (b_i^{k+2} | \hat{b}_i). \end{aligned} \quad (9)$$

Equation (9) applies directly to a variance. For a covariance, it applies when both SRs use the same CQ set because in that case we can condition on a single estimated stretcher $\hat{b}_i$ (the one estimated in that CQ set). If the two SRs instead use distinct dimension-specific CQ sets, then we would need to condition on both estimated stretchers from the two translation functions; we do not develop that extension here.

We estimate the $A_{kd d'}$'s by regressing the product of the two benchmarked SRs, after subtracting an estimate of the conditional response-error contribution in the second expectation, on the estimated conditional moments $\mathbb{E}(b_i^{k+2} | \hat{b}_i)$, without a constant. The CQ-only MLE supplies the part of the response-error contribution generated by the CQ benchmark. Under the maintained independence assumptions, SR response errors do not contribute when $d \neq d'$. When $d = d'$, the correction also includes the SR response-error variance $\sigma_{\eta_s}^2$, which is not separately identified; implementing the variance estimator therefore requires an assumption such as that used in Appendix D.4 (where, as a baseline, we assumed that the SR response-error variance is 50% of the CQ response-error variance).

The law of iterated expectations then gives

$$\mathbb{E} [(w_{id}^* - \theta'_{d\mathbb{C}} \mathbf{w}_{\mathbb{C}}) (w_{id'}^* - \theta'_{d'\mathbb{C}} \mathbf{w}_{\mathbb{C}})] = \sum_{k=0}^K A_{kd d'} \mathbb{E}(b_i^k).$$

Thus,

$$\text{Cov}(w_{id}^*, w_{id'}^*) = \sum_{k=0}^K A_{kd d'} \mathbb{E}(b_i^k) - [\mathbb{E}(w_{id}^*) - \theta'_{d\mathbb{C}} \mathbf{w}_{\mathbb{C}}] [\mathbb{E}(w_{id'}^*) - \theta'_{d'\mathbb{C}} \mathbf{w}_{\mathbb{C}}],$$

where the means come from the first-moment estimator above. Setting $d' = d$ gives $\text{Var}(w_{id}^*)$.

In the empirical analysis in the next subsection, we use the semiparametric estimator only for $\text{Cov}(w_i^*, x_i)$. Our validation exercise instead uses the simpler MM variance estimator (Appendix D.4) because variance correction plays a subsidiary role there.

F.2 Difference in Coefficients between MM and Semiparametric Estimators

To assess the approximation error underlying the MM estimator, Table F.1 reports the difference (with bootstrapped standard errors) between the semiparametric and MM estimates of $\hat{\gamma}_w$ across six SR dimensions in our data. All of these differences are statistically indistinguishable from zero at the 5% false discovery rate threshold.

Table F.1: Difference in Coefficients Between Semiparametric and MM Estimators

Independent Variables	Difference of $\hat{\beta}_s$ Between Semiparametric and MM					
	Life Satisfaction	Happiness	Worthwhileness	No Anxiety	Anxiety	Breathe
Demeaned age / 10	0.42 (0.24)	0.71 (0.73)	0.12 (0.26)	-0.29 (0.33)	0.10 (0.51)	0.91 (0.44)
(Demeaned age) ² / 100	0.27 (0.10)	0.26 (0.36)	0.25 (0.13)	0.01 (0.16)	0.24 (0.21)	0.26 (0.21)
Log(HH income)	0.25 (0.35)	0.57 (1.20)	0.30 (0.42)	0.00 (0.54)	0.57 (0.82)	0.84 (0.66)
Unemployed	0.38 (1.23)	1.01 (4.26)	-1.32 (1.87)	1.13 (1.78)	-1.86 (2.94)	-0.56 (2.33)
Married, not separated	1.31 (0.74)	-2.18 (2.47)	0.17 (0.90)	-1.46 (1.03)	-0.59 (1.33)	-0.61 (1.24)
College grad	-0.17 (0.68)	3.00 (2.07)	-0.16 (0.75)	1.37 (1.03)	0.16 (1.34)	-1.07 (1.02)
Male	-0.57 (0.58)	-0.17 (1.99)	0.04 (0.79)	-0.61 (0.95)	0.64 (1.16)	-2.29 (1.13)
Obs.	10,100	5,063	5,037	5,062	5,038	10,100

Notes: All regressions also include log household size and indicator variables for being employed part-time, being out of the labor force, having at least one child under 19 in the household, ethnicity (Asian, Black/African-American, Hispanic/Latino/Spanish, ‘Other’), one’s state being in the top quartile of states by population density, U.S. census region, and survey day of the week. Bootstrap standard errors (100 resamples) in parentheses.

F.3 Semiparametric Regression Estimator for the Quadratic Common Ranking Case

Quadratic Common Ranking of CQs means that CQ ratings conditional on i and i ’s situation are quadratic in the common-scale level of the CQ. It is weaker than Linear Common Ranking of CQs, but stronger than (Monotonic) Common Ranking of CQs. It only makes sense over a domain where the slope of the quadratic function is everywhere positive or as an approximation where this property holds in the relevant region of data. We continue to take the sample mean as the reference individual. This is a substantive choice in the quadratic case; changing the reference individual has a non-linear effect on “common-scale” levels.¹⁰ Each respondent’s CQ translation function (before response

¹⁰Indeed, the *definition* of “Quadratic Common Ranking” needs to be conditioned on a choice of the reference individual. To see why, suppose we have Quadratic Common Ranking as defined relative to sample-mean scale use, but we then switched to a reference individual who has different scale use. Then scale use for everyone else would be a quadratic function of the inverse of the reference individual’s quadratic scale-use relative to the sample mean. A

error is added) is given by

$$r_{ic} = a_i + b_i w_c + d_i w_c^2,$$

with $\mathbb{E}(a_i) = 0$, $\mathbb{E}(b_i) = 1$, and $\mathbb{E}(d_i) = 0$. To see that this makes the sample mean the reference scale use, note that these normalizations ensure

$$\mathbb{E}(a_i + b_i w_c + d_i w_c^2) = w_c.$$

To make the expressions below simpler, for any choice of benchmark $\theta'_C \mathbf{r}_{iC}$, define ℓ_i such that

$$r_i^\star - \theta'_C \mathbf{r}_{iC} = \ell_i (w_i^\star - \theta'_C \mathbf{w}_C) + d_i (w_i^\star - \theta'_C \mathbf{w}_C)^2.$$

That is, ℓ_i is the slope of the translation function at $\theta'_C \mathbf{w}_C$. The semiparametric estimator is consistent for any choice of target for $\theta'_C \mathbf{w}_C$. As in the Linear Common Ranking case, in the Quadratic Common Ranking case, we choose $\theta'_C \mathbf{w}_C = \mathbb{E}(r_i^\star)$.

We model the key conditional expectations, the first non-parametrically as a general polynomial in ℓ_i and d_i and the second as linear in ℓ_i and d_i :

$$\begin{aligned} \mathbb{E}((w_i^\star - \theta'_C \mathbf{w}_C) x_i \mid \ell_i, d_i) &= A_{00} + A_{10} \ell_i + A_{01} d_i + A_{20} \ell_i^2 + A_{11} \ell_i d_i + A_{02} d_i^2 + \dots \\ \mathbb{E}((w_i^\star - \theta'_C \mathbf{w}_C)^2 x_i \mid \ell_i, d_i) &= B_{00} + B_{10} \ell_i + B_{01} d_i. \end{aligned}$$

What we ultimately need to know is:

$$\mathbb{E}((w_i^\star - \theta'_C \mathbf{w}_C) x_i) = A_{00} + A_{10} \mathbb{E}(\ell_i) + A_{01} \mathbb{E}(d_i) + A_{20} \mathbb{E}(\ell_i^2) + A_{11} \mathbb{E}(\ell_i d_i) + A_{02} \mathbb{E}(d_i^2) + \dots$$

Because $\mathbb{E}(d_i) = 0$, the calculation can be done without knowing A_{01} .

For the identification of all of the A 's other than A_{01} , we need two additional linear combinations of CQ ratings (beyond the benchmark $\theta'_C \mathbf{r}_{iC}$): (i) $\tilde{\ell}_i$, which is equal to ℓ_i + response error, and (ii) $\tilde{d}_i$, which is equal to d_i + response error. Among such linear combinations, we require them to have response errors that are uncorrelated with the response error of the benchmark. To the extent there are degrees of freedom left, we minimize the sum of response error variances. Note that the zero-correlation constraints and the minimization of the sum of response-error variances use our assumption of homoskedasticity of response errors across CQs in the set.¹¹ We use the notation $(\mathbf{w}_C - [\mathbb{E}(r_i^\star)]\mathbf{1})^{\odot 2}$ for the componentwise square of $\mathbf{w}_C - [\mathbb{E}(r_i^\star)]\mathbf{1}$. Formally, we solve:

quadratic function of the inverse of a quadratic function is not quadratic.

¹¹Note that assuming homoskedasticity across CQs is very different from the frequent assumption in the literature of homoskedasticity across respondents.

$$\begin{aligned}
& \min_{\theta_{\mathbb{C}}, \boldsymbol{\vartheta}_{\mathbb{C}}, \boldsymbol{\varphi}_{\mathbb{C}}} \quad \theta'_{\mathbb{C}} \boldsymbol{\theta}_{\mathbb{C}} + \boldsymbol{\vartheta}'_{\mathbb{C}} \boldsymbol{\vartheta}_{\mathbb{C}} + \boldsymbol{\varphi}'_{\mathbb{C}} \boldsymbol{\varphi}_{\mathbb{C}} \\
& \text{s.t.} \quad \theta'_{\mathbb{C}} \mathbf{1} = 1, \quad \theta'_{\mathbb{C}} \mathbf{w}_{\mathbb{C}} = \mathbb{E} r_i^{\star}, \quad \theta'_{\mathbb{C}} (\mathbf{w}_{\mathbb{C}} - [\mathbb{E} r_i^{\star}] \mathbf{1})^{\odot 2} = 0 \\
& \quad \boldsymbol{\vartheta}'_{\mathbb{C}} \mathbf{1} = 0, \quad \boldsymbol{\vartheta}'_{\mathbb{C}} (\mathbf{w}_{\mathbb{C}} - [\mathbb{E} r_i^{\star}] \mathbf{1}) = 1, \quad \boldsymbol{\vartheta}'_{\mathbb{C}} (\mathbf{w}_{\mathbb{C}} - [\mathbb{E} r_i^{\star}] \mathbf{1})^{\odot 2} = 0 \\
& \quad \boldsymbol{\varphi}'_{\mathbb{C}} \mathbf{1} = 0, \quad \boldsymbol{\varphi}'_{\mathbb{C}} (\mathbf{w}_{\mathbb{C}} - [\mathbb{E} r_i^{\star}] \mathbf{1}) = 0, \quad \boldsymbol{\varphi}'_{\mathbb{C}} (\mathbf{w}_{\mathbb{C}} - [\mathbb{E} r_i^{\star}] \mathbf{1})^{\odot 2} = 1 \\
& \quad \theta'_{\mathbb{C}} \boldsymbol{\vartheta}_{\mathbb{C}} = \theta'_{\mathbb{C}} \boldsymbol{\varphi}_{\mathbb{C}} = 0.
\end{aligned}$$

The first row of constraints is what guarantees that the benchmark is the respondent's translation function evaluated at $\mathbb{E}(r_i^{\star})$. The second row of constraints is what guarantees that $\tilde{\ell}_i = \ell_i + \text{response error}$. The third row of constraints is what guarantees that $\tilde{d}_i = d_i + \text{response error}$.

Note that there are 11 constraints. Each of the three linear combinations of CQ ratings we need—benchmark, $\tilde{\ell}_i$ and $\tilde{d}_i$ —has as many coefficients as there are CQs. With only a trio of CQs, that would be $3 \times 3 = 9$ coefficients in all, not enough to satisfy all 11 constraints. However, with four CQs, the three linear combinations put together contain 12 coefficients—enough to satisfy the 11 constraints. Having more than 4 CQs would leave more degrees of freedom for minimizing the sum of the combination response-error variances while satisfying the 11 constraints.

Now, consider the expectation of the product of the net-of-benchmark SR and a regressor, conditional on both $\tilde{\ell}_i$ and $\tilde{d}_i$:

$$\begin{aligned}
& \mathbb{E} \left((r_i^{\star} - \theta'_{\mathbb{C}} \mathbf{r}_{i\mathbb{C}}) x_i \mid \tilde{\ell}_i, \tilde{d}_i \right) \\
& = A_{00} \mathbb{E} \left(\ell_i \mid \tilde{\ell}_i, \tilde{d}_i \right) + A_{10} \mathbb{E} \left(\ell_i^2 \mid \tilde{\ell}_i, \tilde{d}_i \right) + A_{01} \mathbb{E} \left(\ell_i d_i \mid \tilde{\ell}_i, \tilde{d}_i \right) \\
& \quad + A_{20} \mathbb{E} \left(\ell_i^3 \mid \tilde{\ell}_i, \tilde{d}_i \right) + A_{11} \mathbb{E} \left(\ell_i^2 d_i \mid \tilde{\ell}_i, \tilde{d}_i \right) + A_{02} \mathbb{E} \left(\ell_i d_i^2 \mid \tilde{\ell}_i, \tilde{d}_i \right) + \dots \\
& \quad + B_{00} \mathbb{E} \left(d_i \mid \tilde{\ell}_i, \tilde{d}_i \right) + B_{10} \mathbb{E} \left(\ell_i d_i \mid \tilde{\ell}_i, \tilde{d}_i \right) + B_{01} \mathbb{E} \left(d_i^2 \mid \tilde{\ell}_i, \tilde{d}_i \right) \\
& = A_{00} \mathbb{E} \left(\ell_i \mid \tilde{\ell}_i, \tilde{d}_i \right) + A_{10} \mathbb{E} \left(\ell_i^2 \mid \tilde{\ell}_i, \tilde{d}_i \right) + A_{20} \mathbb{E} \left(\ell_i^3 \mid \tilde{\ell}_i, \tilde{d}_i \right) \\
& \quad + (A_{01} + B_{10}) \mathbb{E} \left(\ell_i d_i \mid \tilde{\ell}_i, \tilde{d}_i \right) + A_{11} \mathbb{E} \left(\ell_i^2 d_i \mid \tilde{\ell}_i, \tilde{d}_i \right) + A_{02} \mathbb{E} \left(\ell_i d_i^2 \mid \tilde{\ell}_i, \tilde{d}_i \right) + \dots \\
& \quad + B_{00} \mathbb{E} \left(d_i \mid \tilde{\ell}_i, \tilde{d}_i \right) + B_{01} \mathbb{E} \left(d_i^2 \mid \tilde{\ell}_i, \tilde{d}_i \right),
\end{aligned}$$

where the first equality holds because the weight vectors are constructed to be orthogonal ($\theta'_{\mathbb{C}} \boldsymbol{\vartheta}_{\mathbb{C}} = \theta'_{\mathbb{C}} \boldsymbol{\varphi}_{\mathbb{C}} = 0$); because the SR's and the CQs' response errors are independent; and because of the independence between $(w_i^{\star}, x_i)$ and the response errors. The expectations of monomials in ℓ_i and d_i conditional on both $\tilde{\ell}_i$ and $\tilde{d}_i$ are straightforward to calculate from the CQ-only MLE extended to jointly estimate ℓ_i and d_i . The coefficients of each term in the polynomial can then be identified by a regression of $(r_i^{\star} - \theta'_{\mathbb{C}} \mathbf{r}_{i\mathbb{C}}) x_i$ on the expectations of monomials in ℓ_i and d_i conditional on both $\tilde{\ell}_i$

and $\tilde{d}_i$:

$$\mathbb{E}(\varrho_i^k | \tilde{\varrho}_i, \tilde{d}_i), \quad \mathbb{E}(d_i^k | \tilde{\varrho}_i, \tilde{d}_i), \quad \mathbb{E}(\varrho_i d_i | \tilde{\varrho}_i, \tilde{d}_i), \quad \mathbb{E}(\varrho_i^2 d_i | \tilde{\varrho}_i, \tilde{d}_i), \quad \mathbb{E}(\varrho_i d_i^2 | \tilde{\varrho}_i, \tilde{d}_i), \dots$$

That regression identifies all of the A 's except A_{01} , which we do not need. Then substitute those estimates into the sample analog of

$$\mathbb{E}[(w_i^* - \theta'_C \mathbf{w}_C) x_i] = A_{00} + A_{10} \mathbb{E}(\varrho_i) + A_{20} \mathbb{E}(\varrho_i^2) + A_{11} \mathbb{E}(\varrho_i d_i) + A_{02} \mathbb{E}(d_i^2) + \dots$$

Subsequent steps of the semiparametric estimator are analogous to those of the Linear Common Ranking case (see Web Appendix F.1).

Remark: Note that, without the assumption that $\mathbb{E}((w_i^* - \theta'_C \mathbf{w}_C)^2 x_i | \varrho_i, d_i)$ is linear in ϱ_i and d_i , some of the A 's necessary to compute $\mathbb{E}[(w_i^* - \theta'_C \mathbf{w}_C) x_i]$ would not be identified for the same basic reason that A_{01} is unidentified: the coefficient of the corresponding monomial involves both an A and a B . However, if $\mathbb{E}((w_i^* - \theta'_C \mathbf{w}_C) x_i | \varrho_i, d_i)$ is linear in ϱ_i and d_i , then we only need A_{00} and A_{10} to calculate $\mathbb{E}[(w_i^* - \theta'_C \mathbf{w}_C) x_i]$. A_{00} and A_{10} are identified even if we non-parametrically model $\mathbb{E}((w_i^* - \theta'_C \mathbf{w}_C)^2 x_i | \varrho_i, d_i)$. Thus, it is enough for identification if *either* $\mathbb{E}((w_i^* - \theta'_C \mathbf{w}_C)^2 x_i | \varrho_i, d_i)$ *or* $\mathbb{E}((w_i^* - \theta'_C \mathbf{w}_C) x_i | \varrho_i, d_i)$ is linear in ϱ_i and d_i . If neither is linear, one could approach identification by looking *also* at $\mathbb{E}((r_i^* - \theta'_C \mathbf{r}_C)^2 x_i | \varrho_i, d_i)$. Identification would then require modeling the heteroskedasticity of response errors as functions of x_i .

G Test of Proportional Coefficients

Throughout our analyses, we test whether two vectors of regression coefficients—for example, scale-use-corrected versus uncorrected coefficient (Table 2), or scale-use-corrected coefficients based on two different CQ sets (Figure 5)—differ by more than a common rescaling. We use the same test throughout, which we call the *test of proportional coefficients*.

Let $\mathbf{b}_A$ and $\mathbf{b}_B$ denote the two vectors of coefficients being compared, where the intercept and any regressors included only as controls are excluded. The null hypothesis is that $\mathbf{b}_A = \lambda \mathbf{b}_B$ for some scalar λ . That is, that the two vectors lie on a common line through the origin, so that every ratio of coefficients is the same in the two vectors. The test statistic is the minimized quadratic form

$$T = \min_{\lambda} (\mathbf{b}_A - \lambda \mathbf{b}_B)' \widehat{\mathbf{V}}_{\lambda}^{-1} (\mathbf{b}_A - \lambda \mathbf{b}_B), \quad \widehat{\mathbf{V}}_{\lambda} = \widehat{\mathbf{V}}_{AA} + \lambda^2 \widehat{\mathbf{V}}_{BB} - \lambda \widehat{\mathbf{V}}_{AB} - \lambda \widehat{\mathbf{V}}_{BA},$$

where the minimization is over both positive and negative λ (and $\widehat{\mathbf{V}}_{BA}$ is the transpose of $\widehat{\mathbf{V}}_{AB}$). We test T using its asymptotic distribution, χ^2 with degrees of freedom equal to the number of coefficients in $\mathbf{b}_A$ (or, equivalently, $\mathbf{b}_B$) minus one, the subtracted degree of freedom corresponding to the free scale factor λ .

The joint covariance matrix of $(\mathbf{b}_A, \mathbf{b}_B)$ is estimated from 500 respondent-level bootstrap draws. Drawing respondents jointly for the two regressions allows their estimation errors to be correlated, as they generally are when the two regressions use overlapping respondents. When the two coefficient vectors are instead estimated on disjoint samples—as in our cross-framing comparisons, where framing is randomized at the respondent level (Web Appendix Table L.8)—the cross-covariance blocks $\widehat{\mathbf{V}}_{AB}$ and $\widehat{\mathbf{V}}_{BA}$ are zero and the weighting matrix reduces to $\widehat{\mathbf{V}}_{AA} + \lambda^2 \widehat{\mathbf{V}}_{BB}$.

Note that because the statistic minimizes over both positive and negative λ , non-rejection is consistent with proportionality at a factor of the unexpected sign. Therefore, the sign of λ is an important component of the test output.

We also report, alongside the p -value, a simple measure of how far two coefficient vectors are from proportional: the correlation between the two regressions' fitted values, $\text{Corr}(\mathbf{x}'_i \hat{\mathbf{b}}_A, \mathbf{x}'_i \hat{\mathbf{b}}_B)$, taken over the same coefficients that enter the test (so the intercept is excluded). We report this, rather than a correlation computed directly from the two coefficient vectors, for the following reasons. First, it is invariant to how the regressors are scaled and parametrized, exactly as the test statistic T is: replacing $\mathbf{x}_i$ by $\mathbf{A}'\mathbf{x}_i$ for any nonsingular $\mathbf{A}$ leaves both unchanged. A correlation of the raw coefficients is not invariant and would therefore move if, for example, age entered in years rather than decades or if a different category were omitted from a set of indicators. Second, it weights each coefficient by how much its regressor actually varies in the data, so that coefficients on regressors that move fitted values little receive correspondingly little weight. Finally, it is closely related to the test. In the special case in which the two regressions have the same regressor second-moment matrix $\mathbf{X}'\mathbf{X}$ —exactly, or in expectation when they are estimated on disjoint but randomly assigned samples, as in our cross-framing comparisons—have homoskedastic errors, and have uncorrelated estimation errors, $\widehat{\mathbf{V}}_\lambda = (\mathbf{X}'\mathbf{X})^{-1} s^2(\lambda)$ for a scalar $s^2(\lambda)$, so that

$$T = \min_{\lambda} \frac{\|\mathbf{X}\mathbf{b}_A - \lambda\mathbf{X}\mathbf{b}_B\|^2}{s^2(\lambda)}$$

is the minimized scaled distance between the two vectors of fitted values, the test statistic T is a decreasing function of their squared correlation, and $\hat{\lambda}$ is the total-least-squares slope from regressing one set of fitted values on the other through the origin, with each scaled by its standard error. (Note, however, that our implementation generalizes this special case—the bootstrap covariance accommodates heteroskedasticity, error in the estimated benchmark, and correlated estimation errors across the two regressions—and in this more general case, the monotonic relationship between the test statistic and the squared correlation is no longer guaranteed to hold.)

H Comparison of Results Between MM and CHOPIT

CHOPIT cannot be run directly on our data with a 0–100 response scale because, with so many response options, it fails to converge. Following Montgomery (2022), we therefore coarsen the responses by rounding our 0–100 data to the nearest multiple of 10 and then mapping them to a 1–11 scale. Using this coarsened variable in all cases, Table H.1 compares regression coefficients from four approaches to estimating the relationship between the Life Satisfaction SR and demographic variables: unadjusted OLS, our MM estimator, the standard CHOPIT model, and a modified version of CHOPIT. Both sets of CHOPIT coefficients are rescaled for comparability with the OLS and MM coefficients.

The standard CHOPIT model (Column 3) produces coefficients that are, in general, much closer to the unadjusted coefficients (Column 1) than the MM coefficients (Column 2). One possible reason is that CHOPIT uses not only the calibration questions (CQs) but also the self-rating (SR) data to identify scale-use parameters. This is potentially problematic because it confounds scale-use heterogeneity with variation in common-scale levels of the SR. Our MM estimator avoids this problem by identifying scale-use parameters solely from CQ data. To investigate how much this issue contributes to the discrepancy between Columns 2 and 3, we modified the publicly available CHOPIT code to identify scale-use parameters only from the CQ data. In general the resulting estimates, in Column 4, are substantially closer to those from MM. For example, the coefficient on log income is 0.17 under MM and 0.19 under Modified CHOPIT, compared with 0.42 under standard CHOPIT (which is nearly identical to the unadjusted estimate of 0.43). Similarly, the coefficients on unemployment and marriage are much closer between MM and Modified CHOPIT than between MM and standard CHOPIT.

Some differences between MM and Modified CHOPIT remain, the largest being the coefficient on college, which is -0.19 under MM but 0.40 under Modified CHOPIT. We do not know the source of these discrepancies, but we suspect they are driven by the parametric assumptions imposed by CHOPIT, such as normality and homoskedasticity of the latent variable, which our MM estimator does not require. Normality and homoskedasticity of the latent variable are untestable without strong assumptions. However, we can test—and we strongly reject—the corresponding properties of the observed SR in our data. The Anderson-Darling test rejects the null hypothesis of normality ($p = 3.7 \times 10^{-24}$). Moreover, the Breusch-Pagan test rejects homoskedasticity in the life satisfaction regressions, both for the raw SR ($p = 7.1 \times 10^{-103}$) and the MM scale-use-adjusted version ($p = 2.3 \times 10^{-20}$).

In light of the differences we find between MM and (Modified) CHOPIT, a natural follow-up question is how CHOPIT fares in the validation exercise of Section VI, which compares scale-use-adjusted subjective measures of height and weight with their objective counterparts. The validation

Table H.1: Comparing Regression Coefficients Between MM and CHOPIT

Independent Variables	Dependent Variable: Life Satisfaction (Discretized)			
	Unadjusted	MM	CHOPIT	Modified CHOPIT
Demeaned age / 10	0.25 (0.02)	0.35 (0.03)	0.22 (0.02)	0.26 (0.02)
(Demeaned age) ² / 100	0.05 (0.01)	0.04 (0.01)	0.05 (0.01)	0.05 (0.01)
Log(HH income)	0.43 (0.03)	0.17 (0.04)	0.42 (0.03)	0.19 (0.01)
Unemployed	−0.69 (0.12)	−0.86 (0.16)	−0.53 (0.11)	−0.94 (0.11)
Married, not separated	0.62 (0.05)	0.70 (0.08)	0.61 (0.06)	0.77 (0.07)
College grad	0.08 (0.05)	−0.19 (0.07)	0.23 (0.06)	0.40 (0.06)
Male	−0.12 (0.04)	−0.07 (0.06)	−0.13 (0.05)	−0.17 (0.06)
Obs.	10,100	10,100	10,100	10,100

Notes: Data: UAS. For comparability with CHOPIT, the dependent variable is discretized to 1–11 for all methods. CHOPIT uses both CQs and SRs to identify scale-use heterogeneity; modified CHOPIT uses CQs only. Scale-use adjustment uses the 3 Life Satisfaction CQs. All regressions also include log household size and indicator variables for being employed part-time, being out of the labor force, having at least one child under 19 in the household, ethnicity (Asian, Black/African-American, Hispanic/Latino/Spanish, ‘Other’), one’s state being in the top quartile of states by population density, U.S. census region, and survey day of the week. CHOPIT coefficients are scaled by the slope from a TLS regression of each method’s predictions on OLS predictions. Bootstrap standard errors (100 resamples) in parentheses.

exercise in Table 3 cannot be replicated directly under CHOPIT because it requires correcting for scale-use heterogeneity in the variance of the subjective measure—a correction that CHOPIT does not provide. To still permit some comparison, Table H.2 reports the correlations between the objective measure and the fitted values of (i) the (scale-use-adjusted) subjective measure for MM, (ii) default CHOPIT, and (iii) Modified CHOPIT, using five sets of CQs: the same-dimension CQs, each of the three visual CQ trios (Darkness, Shape, and Confidence) separately, and all nine of these visual CQs jointly.

As shown in Table H.2, across all five sets of CQs, the correlations under MM, default CHOPIT, and Modified CHOPIT are quantitatively similar, with most pairwise differences statistically indistinguishable from zero at the 5% level (based on bootstrapped standard errors of the relevant differences). For height, MM yields more predictive fitted values than default CHOPIT when scale-use adjustment uses the same-dimension CQs (0.58 vs. 0.52, $p = 6.7 \times 10^{-6}$), while default CHOPIT yields more predictive fitted values when using the Confidence visual CQs (0.45 vs. 0.54, $p = 9.1 \times 10^{-3}$); the remaining height comparisons between MM and default CHOPIT (Darkness, Shape, and all nine visual CQs) are statistically indistinguishable at the 5% level. For weight, MM yields more predictive fitted values than default CHOPIT for the Darkness visual CQs (0.17 vs. 0.10, $p = 0.04$) and the Confidence visual CQs (0.11 vs. 0.06, $p = 0.05$); the same-dimension, Shape, and all-nine-visual comparisons are not statistically distinguishable at the 5% level. Notably, MM and Modified CHOPIT are statistically indistinguishable at the 5% level for every set of CQs for both

Table H.2: Correlations of Objective-Scale Ratings and Fitted Subjective-Scale Ratings under MM and CHOPIT

	Height (Inches)			Weight (Pounds)		
	MM	CHOPIT	Modified CHOPIT	MM	CHOPIT	Modified CHOPIT
Same-Dimension CQs	0.58 (0.04)	0.52 (0.04)	0.52 (0.05)	0.19 (0.04)	0.18 (0.04)	0.18 (0.04)
3 Darkness Visual CQs	0.62 (0.04)	0.60 (0.04)	0.54 (0.10)	0.17 (0.05)	0.10 (0.06)	0.10 (0.05)
3 Shape Visual CQs	0.52 (0.05)	0.55 (0.04)	0.54 (0.06)	0.07 (0.05)	0.04 (0.05)	0.05 (0.05)
3 Confidence Visual CQs	0.45 (0.05)	0.54 (0.05)	0.53 (0.08)	0.11 (0.04)	0.06 (0.05)	0.07 (0.05)
9 Visual CQs	0.56 (0.04)	0.58 (0.04)	0.57 (0.04)	0.11 (0.05)	0.08 (0.06)	0.09 (0.05)

Notes: Correlations between objective-scale measures of height and weight and the fitted (scale-use-adjusted) values of the corresponding SRs under MM, default CHOPIT, and Modified CHOPIT. Default CHOPIT uses both CQs and SRs to identify scale-use heterogeneity; Modified CHOPIT uses CQs only. Data: SRs and Height and Weight CQs from Survey 2; objective-scale measures from respondents' most recent prior UAS reports (Section I.E); visual CQs from Survey 1. Rows: the CQ set used for the scale-use adjustment—same-dimension CQs, each visual CQ trio (Darkness, Shape, Confidence) separately, and all nine visual CQs jointly. For each method, the fitted value is the respondent's regressor vector multiplied by that method's estimated coefficients; the correlations are therefore invariant to the cross-method rescaling used in Table H.1. Bootstrap standard errors (100 resamples) in parentheses.

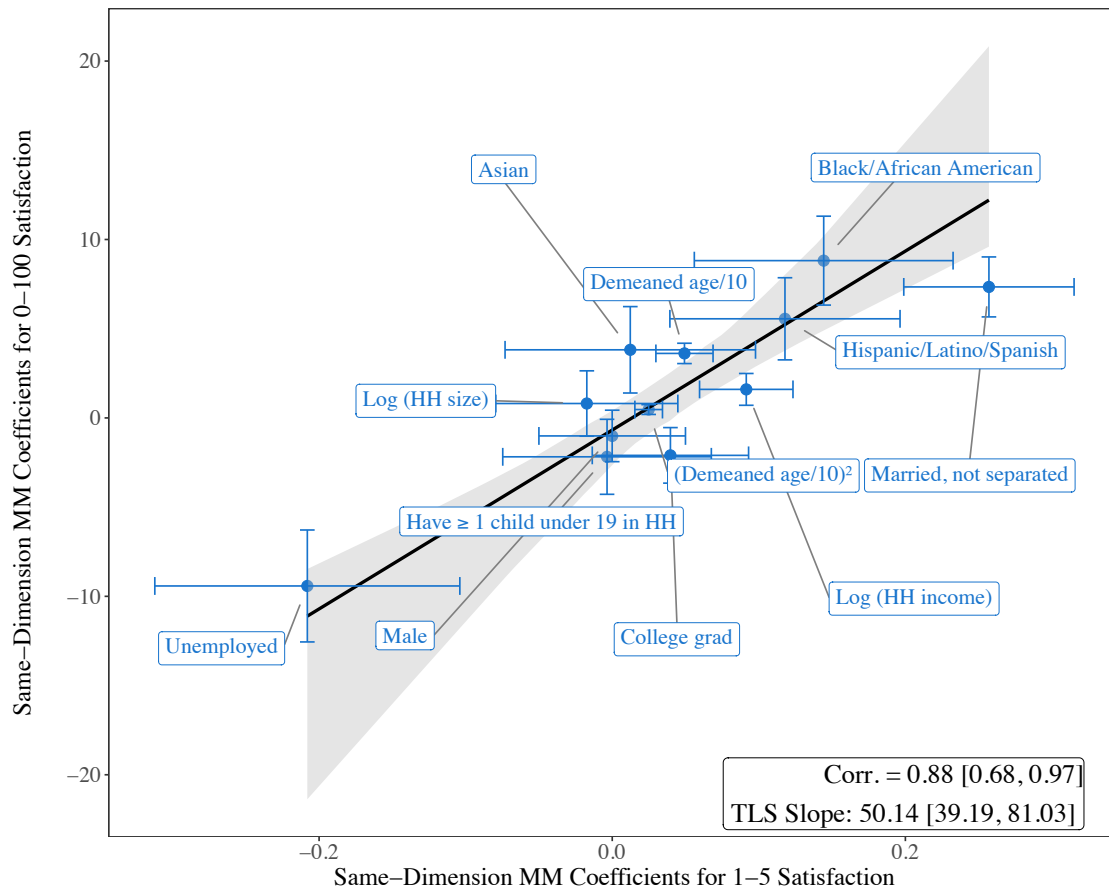
height and weight, consistent with the finding in Table H.1 that identifying scale-use parameters from the CQs alone brings CHOPIT closer to MM. Default and Modified CHOPIT are likewise statistically indistinguishable throughout. Overall, the three methods produce a similar picture in this exercise.

I MM With Likert-Scale Response Options

Our MM estimator, much like OLS, is derived with and most naturally understood with a continuous SR (self-rating) variable. In practice, many SR questions are asked on a Likert scale with just a few response options. To examine how well the MM estimator performs when the dependent variable has a Likert scale, we exploited data from UAS 580, a separate UAS survey fielded in November 2023, which falls within the August 2023–May 2024 fielding period of Survey 1, that elicited life satisfaction on a 5-point Likert scale. Figure I.1 compares the MM-adjusted coefficients obtained from this variable with those obtained from our 0–100 scale. In both cases, the scale-use-only benchmark was constructed from the three Life Satisfaction vignettes we elicited on a 0–100 scale. To put the two on a common scale, we bin the 0–100 CQ ratings into five categories, coding ratings of 20 or below as 1 and each successive 20-point interval as 2, 3, 4, and 5. The MM estimator is then applied using this 1–5 scale, with intercepts and stretchers estimated from the binned CQ ratings and their sample means, and with the benchmark targeting the mean of the 1–5 SR. Because binning compresses the dispersion of the CQ means, the benchmark is noisier on the 1–5 scale than on the 0–100 scale, so the correlation we report is, if anything, a lower bound on the agreement between the two. Each point represents the coefficient on a demographic variable, plotted for the 1–5 scale on the horizontal axis and the 0–100 scale on the vertical axis. The estimated correlation between

the two sets of coefficients is 0.88 (95% CI: [0.68, 0.97]). Although future research should conduct such comparisons more systematically, the strong correlation we find (despite the separation in time between the two Life Satisfaction SRs) suggests that MM produces similar patterns of demographic coefficients regardless of whether the underlying response scale is continuous or has as few as 5 response options. These results suggest that MM may be a practical alternative to CHOPIT even in settings where only a limited number of response categories are available.

Figure I.1: Comparison of Coefficients from 0–100 versus 1–5 Response Scale for Life Satisfaction



Notes: Each point: coefficient on a demographic variable. *x*-axis: MM applied to the 1–5 Life Satisfaction SR from UAS 580; *y*-axis: MM applied to our 0–100 Life Satisfaction SR from Survey 1. Scale-use-only benchmark constructed from the three Life Satisfaction vignettes. TLS line and its 95% CI shown, along with 95% CIs on the point estimates. Correlation between coefficients and its 95% CI shown in the annotation box.

J Regressions Using $\hat{w}_i^*$ (Obtained by Inverting the Estimated CQ Translation Function)

As discussed in Section IV, the most straightforward approach to scale-use correction is at the individual level: invert each respondent's estimated translation function, and use $\hat{w}_i^* = (r_i^* - \hat{a}_i)/\hat{b}_i$ as the dependent variable in a regression in place of r_i^* . However, because $\hat{b}_i$ is estimated from only three CQs, it is noisy and can be near zero, causing $\hat{w}_i^*$ to take on extreme values for some respondents. Table J.1 reports what we found when we took this approach. The first two columns reproduce the unadjusted and (same-dimension) MM estimates from Table 2 for comparison. The third column reports coefficients from regressing $\hat{w}_i^*$ on demographics (excluding the 307 respondents whose estimated stretcher is exactly zero because they gave identical ratings to all three CQs). Even after these algebraically degenerate cases are removed, the estimator is dominated by respondents whose $\hat{b}_i$ is close to zero: no coefficient is statistically distinguishable from zero, and the standard errors are one to two orders of magnitude larger than the corresponding MM standard errors. Winsorizing $\hat{w}_i^*$ (Columns 4 and 5) restores the coefficients to a plausible scale but the estimator remains far less precise than MM (and note that winsorizing itself introduces bias). At 1% winsorizing the standard errors remain roughly four to five times the corresponding MM standard errors. Even winsorizing at 5% they remain roughly twice as large: none of the reported coefficients is statistically distinguishable from zero after false-discovery-rate adjustment (including the coefficient on log income, one of the most robust predictors in the unadjusted and MM-adjusted regressions). These results indicate that, with only a small number of CQs, the naïve inversion approach is far too imprecise to be useful. This problem motivates the MM estimator developed in Section IV.

Table J.1: Regressions Using $\hat{w}_i^\star$

Independent Variables	Dependent Variable: Life Satisfaction				
	(1) Unadjusted	(2) MM	(3) $\hat{w}_i^\star$	(4) $\hat{w}_i^\star$ (1% Winsor)	(5) $\hat{w}_i^\star$ (5% Winsor)
Demeaned age / 10	2.49 (0.18)	3.46 (0.26)	51.56 (36.00)	0.94 (1.13)	0.80 (0.53)
(Demeaned age) ² / 100	0.50 (0.09)	0.43 (0.13)	-17.34 (12.14)	-0.00 (0.57)	0.17 (0.27)
Log(HH income)	4.20 (0.31)	1.71 (0.41)	9.61 (16.33)	2.04 (1.98)	1.78 (0.89)
Unemployed	-6.90 (1.10)	-8.47 (1.42)	-38.15 (102.30)	-4.21 (6.31)	-3.97 (2.81)
Married, not separated	6.09 (0.56)	6.82 (0.77)	-81.81 (145.67)	3.68 (3.35)	1.48 (1.64)
College grad	0.72 (0.49)	-1.93 (0.71)	-107.70 (64.49)	-2.33 (3.13)	0.02 (1.48)
Male	-1.32 (0.46)	-0.72 (0.66)	-19.26 (40.69)	-0.40 (2.97)	-2.29 (1.42)
Obs.	10,100	10,100	9,793	9,793	9,793

Notes: Data: UAS. Scale-use adjustment uses the 3 Life Satisfaction CQs. Smaller sample size in the naïve scale-use adjustments because 307 respondents gave identical ratings to all three CQs, so their stretchers are estimated as zero. $x\%$ Winsor replaces a total of $x\%$ of the sample, $x/2\%$ from each tail. All regressions also include log household size and indicator variables for being employed part-time, being out of the labor force, having at least one child under 19 in the household, ethnicity (Asian, Black/African-American, Hispanic/Latino/Spanish, ‘Other’), one’s state being in the top quartile of states by population density, U.S. census region, and survey day of the week. Heteroskedasticity-robust standard errors in parentheses.

K Simulations

This appendix reports Monte Carlo simulations that assess the finite-sample performance of the MM estimator. The simulations serve two purposes. First, they verify that MM removes scale-use bias while leaving unbiased coefficients undistorted. Second, they quantify how the precision of the MM estimator varies with the number of CQs and the respondent sample size, which is useful for survey design. We also investigate the robustness of MM to discrete (Likert-type) response scales.

K.1 Data-Generating Process

We simulate SRs according to the linear translation function in equation (3):

$$r_i^\star = a_i + b_i w_i^\star + \eta_i^\star,$$

where a_i , b_i , $w_i^\star$, and $\eta_i^\star$ are drawn independently across individuals according to

$$\begin{bmatrix} a_i \\ \ln b_i \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ -\frac{\sigma_{\ln b}^2}{2} \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \rho_{a, \ln b} \sigma_a \sigma_{\ln b} \\ \rho_{a, \ln b} \sigma_a \sigma_{\ln b} & \sigma_{\ln b}^2 \end{bmatrix} \right),$$

$$w_i^\star \sim \mathcal{N}(\mu_w, \sigma_w^2), \quad \eta_i^\star \sim \mathcal{N}(0, \sigma_\eta^2).$$

We calibrate the distribution of b_i using the log-normal rather than the normal version of the CQ-only MLE (Appendix D.1) because the simulations draw individual stretchers, for which negative values have no interpretation. The mean of $\ln b_i$ is set to $-\sigma_{\ln b}^2/2$ so that $\mathbb{E}(b_i) = 1$, ensuring that the mean

of respondents' scales coincides with the common scale. Parameter values are calibrated based on our estimates from the data: $\sigma_a = 30$, $\sigma_{\ln b} = 0.37$, $\rho_{a, \ln b} = -0.93$, $\mu_w = 60$, $\sigma_w = 20$, and $\sigma_\eta = 10$. The negative correlation between a_i and $\ln b_i$ captures the empirical pattern that respondents with higher intercepts tend to have lower stretchers.

For each individual, we also simulate CQ responses. The number of CQs varies across simulation designs, with $C \in \{2, 3, 9, 27, 81\}$. Each individual's CQ responses are generated according to equation (2):

$$r_{ic} = a_i + b_i w_c + \eta_{ic},$$

where the common-scale CQ ratings w_c are drawn once per CQ from $\mathcal{N}(60, 10^2)$, and each respondent's response error for that CQ is drawn from $\eta_{ic} \sim \mathcal{N}(0, \sigma_{\eta_c}^2)$ with $\sigma_{\eta_c}^2 = 208$. We vary the number of respondents with $I \in \{100, 1,000, 10,000\}$.

To construct a right-hand-side variable whose correlation with scale-use parameters we can control, we generate

$$x_i = Aa_i + Bb_i + Dw_i^* + \iota_i,$$

where $\iota_i \sim \mathcal{N}(0, 10^2)$ is independent of all other variables. The parameters A , B , and D govern the partial correlation of x_i with the intercept, the stretcher, and the common-scale level, respectively. We fix $D = 1$ throughout and consider four configurations of (A, B) , shown in Table K.1, that illustrate qualitatively different bias scenarios:

- SR 1 ($A = 0, B = 0$): x_i is uncorrelated with both scale-use parameters. OLS is unbiased, so MM should not distort the estimate.
- SR 2 ($A = 1, B = 0$): x_i loads directly on the intercept only; because $\rho_{a, \ln b} = -0.93$, it remains correlated with the stretcher. OLS is biased upward because respondents with higher x_i tend to have higher intercepts, inflating r_i^* .
- SR 3 ($A = 0, B = 50$): x_i loads directly on the stretcher only but is correlated with the intercept for the same reason. Two opposing channels create OLS bias. First, since $\rho_{a, \ln b} < 0$, higher- x_i respondents have larger b_i but lower a_i , pulling ratings down. Second, since $\mu_w > 0$, larger b_i amplifies the positive mean, pulling ratings up. The first channel slightly dominates under our parameterization, so the net OLS bias is negative.
- SR 4 ($A = 1, B = 50$): x_i is correlated with both scale-use parameters. OLS is biased upward as the intercept effect dominates.

For each combination of C , I , and (A, B) , we run 1,000 Monte Carlo replications. In each replication, we estimate (i) the OLS coefficient from a naïve regression of r_i^* on x_i and (ii) the MM-adjusted coefficient from our method-of-moments estimator using the simulated CQ responses. We report the mean coefficient and the standard deviation of the coefficient across replications.

K.2 Simulation Results

Table K.1 reports the simulation results. We highlight four findings.

Table K.1: Simulation Results: Continuous Response Scale

Num. of CQs	A	B	D	Truth	I = 100		I = 1,000		I = 10,000	
					OLS	MM	OLS	MM	OLS	MM
2	0	0	1	0.80	0.80 (0.11)	0.80 (0.10)	0.80 (0.03)	0.80 (0.03)	0.80 (0.01)	0.80 (0.01)
2	1	0	1	0.29	0.49 (0.06)	0.29 (0.06)	0.49 (0.02)	0.29 (0.02)	0.49 (0.01)	0.29 (0.01)
2	0	50	1	0.46	0.37 (0.10)	0.46 (0.09)	0.37 (0.03)	0.46 (0.03)	0.37 (0.01)	0.46 (0.01)
2	1	50	1	0.54	0.82 (0.07)	0.54 (0.09)	0.82 (0.02)	0.54 (0.03)	0.82 (0.01)	0.54 (0.01)
3	0	0	1	0.80	0.80 (0.10)	0.80 (0.09)	0.80 (0.03)	0.80 (0.03)	0.80 (0.01)	0.80 (0.01)
3	1	0	1	0.29	0.48 (0.06)	0.28 (0.07)	0.49 (0.02)	0.28 (0.02)	0.49 (0.01)	0.29 (0.01)
3	0	50	1	0.46	0.37 (0.10)	0.46 (0.08)	0.37 (0.03)	0.46 (0.03)	0.37 (0.01)	0.46 (0.01)
3	1	50	1	0.54	0.83 (0.07)	0.55 (0.09)	0.82 (0.02)	0.54 (0.03)	0.82 (0.01)	0.54 (0.01)
9	0	0	1	0.80	0.80 (0.10)	0.79 (0.09)	0.80 (0.03)	0.80 (0.03)	0.80 (0.01)	0.80 (0.01)
9	1	0	1	0.29	0.48 (0.06)	0.29 (0.06)	0.49 (0.02)	0.29 (0.02)	0.49 (0.01)	0.29 (0.01)
9	0	50	1	0.46	0.37 (0.10)	0.46 (0.08)	0.37 (0.03)	0.46 (0.03)	0.37 (0.01)	0.46 (0.01)
9	1	50	1	0.54	0.82 (0.07)	0.54 (0.08)	0.82 (0.02)	0.55 (0.03)	0.82 (0.01)	0.54 (0.01)
27	0	0	1	0.80	0.80 (0.11)	0.80 (0.08)	0.80 (0.03)	0.80 (0.03)	0.80 (0.01)	0.80 (0.01)
27	1	0	1	0.29	0.48 (0.06)	0.28 (0.06)	0.49 (0.02)	0.29 (0.02)	0.49 (0.01)	0.29 (0.01)
27	0	50	1	0.46	0.37 (0.10)	0.46 (0.08)	0.37 (0.03)	0.46 (0.03)	0.37 (0.01)	0.46 (0.01)
27	1	50	1	0.54	0.82 (0.07)	0.55 (0.08)	0.82 (0.02)	0.54 (0.02)	0.82 (0.01)	0.54 (0.01)
81	0	0	1	0.80	0.80 (0.10)	0.80 (0.08)	0.80 (0.03)	0.80 (0.03)	0.80 (0.01)	0.80 (0.01)
81	1	0	1	0.29	0.49 (0.06)	0.29 (0.06)	0.49 (0.02)	0.29 (0.02)	0.49 (0.01)	0.29 (0.01)
81	0	50	1	0.46	0.37 (0.09)	0.46 (0.08)	0.37 (0.03)	0.46 (0.02)	0.37 (0.01)	0.46 (0.01)
81	1	50	1	0.54	0.82 (0.07)	0.55 (0.08)	0.82 (0.02)	0.54 (0.02)	0.82 (0.01)	0.54 (0.01)

Notes: Results from 1,000 Monte Carlo replications for each combination of $I \in \{100, 1,000, 10,000\}$ and $C \in \{2, 3, 9, 27, 81\}$. Each cell: mean coefficient (standard deviation across replications in parentheses). “Truth”: population coefficient implied by the data-generating process. OLS: from regressing r_i^* on x_i . MM: method-of-moments estimator from Section IV. A and B govern the correlation of x_i with the intercept and stretcher; $D = 1$ throughout.

First, when x_i is uncorrelated with scale-use parameters (SR 1: $A = B = 0$), OLS is unbiased. The MM estimator also recovers the true coefficient, confirming that the scale-use adjustment does not introduce distortion when no bias is present.

Second, in SRs 2–4, where x_i is correlated with one or both scale-use parameters, OLS is substantially biased. The magnitude and direction of the bias depend on which parameters are correlated with x_i , consistent with the bias formula in equation (4). For example, when x_i loads directly on the intercept only (SR 2), OLS overestimates the true coefficient by roughly 70% (0.49 versus 0.29). When x_i loads directly on the stretcher only (SR 3), the bias goes in the opposite direction: OLS underestimates the true coefficient (0.37 versus 0.46). When both are correlated (SR 4), OLS overestimates by more than 50% (0.82 versus 0.54). In all of these cases, MM removes the bias: the mean MM coefficient closely matches the truth across all sample sizes and numbers of CQs considered.

Third, increasing the number of CQs reduces the standard deviation of the MM estimator, but

not enough to be visible in the table after rounding. For example, the largest gain comes from moving from $C = 2$ to $C = 9$: for SR 4 at $I = 10,000$, the standard deviation drops from 0.0086 to 0.0080, both rounded in the table to 0.01. Further increases in C yield progressively smaller improvements; for instance, the reduction from $C = 9$ to $C = 81$ is minimal. This pattern suggests that, for the purpose of survey design, a modest number of CQs (around 3 to 9) captures most of the precision gains, and little precision is sacrificed by having as few as two CQs. The case for additional CQs beyond two therefore rests mainly on specification testing rather than precision. As noted in Section VIII (“Number of CQs and Sampling Design”), we recommend asking at least three, which enables estimation of the response-error variance and the group-level test of Linear Common Ranking, and at least four if aiming to test Linear Common Ranking (compared to average ratings) at the individual level or to do semiparametric estimation under the assumption of Quadratic Common Ranking.

Fourth, increasing the sample size I reduces the standard deviation of both OLS and MM at approximately the $\sqrt{I}$ rate, as expected for asymptotically normal estimators. Moving from $I = 100$ to $I = 10,000$ (a 100-fold increase) reduces standard deviations by roughly a factor of 10.

K.3 Discrete Response Scales

Although our econometric model assumes a continuous response scale, many existing surveys use Likert-type scales with a small number of response categories. To assess the robustness of MM to discretization, we repeat the simulations but with both the SR r_i^* and the CQ ratings r_{ic} discretized to $K \in \{2, 3, 5, 7, 11\}$ response options. The cutpoints are set at quantiles of the simulated continuous SR distribution, so that each SR response category is used with roughly equal frequency, and the same cutpoints are then applied to the CQ ratings. Because the SR and CQ marginals differ under the data-generating process, the CQ categories are not used with equal frequency; this is the realistic case, since in practice a researcher fixes one response scale for both. We focus on $I = 10,000$ respondents and $C = 3$ CQs, and we run 1,000 replications.

Because discretization changes the scale of the dependent variable and hence the magnitude of coefficients, we normalize all coefficients by dividing by the SR 1 coefficient ($A = B = 0$) at each K .¹² The normalized true coefficient, denoted Truth*, represents the ratio $\gamma_{w\mathbf{x}}/\gamma_{w\mathbf{x}}^{(1)}$, where the superscript (1) refers to SR 1. We use an asterisk throughout to denote a coefficient normalized in this way, so that OLS* and MM* are each divided by that estimator’s own SR 1 coefficient at the same K ; SR 1 therefore equals one by construction for both estimators. Under the continuous DGP, these ratios are 1.00, 0.36, 0.58, and 0.68 for SRs 1–4, respectively; they serve as benchmarks against which to assess the effect of discretization.

¹²This normalization is invariant to any linear transformation of the rating scale: every OLS and MM coefficient is multiplied by the same factor, which cancels in the ratio.

Table K.2 reports the results. Two patterns are evident. First, OLS remains substantially biased under discretization: the normalized OLS coefficients are far from their true values for SRs 2–4 across all values of K . This demonstrates that discretization does not mitigate scale-use bias.

Table K.2: Simulation Results: Discrete Response Scales

	$K = 2$		$K = 3$		$K = 5$		$K = 7$		$K = 11$	
Truth	OLS	MM	OLS	MM	OLS	MM	OLS	MM	OLS	MM
1.00	1.00 (0.01)	1.00 (0.02)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)
0.36	0.64 (0.01)	0.32 (0.01)	0.63 (0.01)	0.33 (0.01)	0.63 (0.01)	0.34 (0.01)	0.63 (0.01)	0.34 (0.01)	0.63 (0.01)	0.34 (0.01)
0.58	0.44 (0.01)	0.58 (0.01)	0.44 (0.01)	0.58 (0.01)	0.45 (0.01)	0.58 (0.01)	0.45 (0.01)	0.57 (0.01)	0.45 (0.01)	0.57 (0.01)
0.68	1.06 (0.01)	0.62 (0.01)	1.05 (0.01)	0.63 (0.01)	1.05 (0.01)	0.64 (0.01)	1.05 (0.01)	0.65 (0.01)	1.05 (0.01)	0.65 (0.01)

Notes: Results from 1,000 Monte Carlo replications with ratings discretized to $K \in \{2, 3, 5, 7, 11\}$ response options using quantile-based cutpoints. Both r_i^* and r_{ic} use the same cutpoints. $I = 10,000$ and $C = 3$ CQs. All coefficients normalized by dividing by the SR 1 ($A = B = 0$) coefficient at each K . Each cell: mean normalized coefficient (standard deviation across replications in parentheses).

Second, MM substantially reduces the bias, though discretization introduces a residual bias whose magnitude varies across SRs. For SR 3, where x_i is correlated directly only with the stretcher, MM recovers normalized coefficients very close to the truth at all K (e.g., 0.58 versus 0.58 at $K = 3$; 0.57 versus 0.58 at $K = 7$). For SRs 2 and 4, the residual bias is larger: SR 2’s MM* ranges from 0.32 ($K = 2$) to 0.34 ($K = 11$) against a truth of 0.36, and SR 4’s MM* ranges from 0.62 ($K = 2$) to 0.65 ($K = 11$) against a truth of 0.68. In both cases, the gap narrows as K increases. What matters for practice, however, is how much of the OLS bias the correction removes, and that share is high at every K we consider: even at $K = 2$, the coarsest scale, MM removes 88% of the OLS bias in SR 2 and 84% in SR 4, rising to 95% and 92% at $K = 11$; in SR 3 the residual bias is negligible throughout. The residual bias under discretization arises because the linear translation function is only an approximation when the response scale is coarse: the mapping from the underlying continuous rating to the discrete response category is a step function rather than a linear function. For example, for SR 4 at $K = 3$, OLS* is 1.05 while MM* is 0.63, compared to a truth of 0.68—the correction closes 86% of the gap. This suggests that the MM estimator is a useful tool even in settings with few response categories, consistent with the empirical evidence on approximate linearity discussed in Sections II.B and IV.

K.4 Standard Errors

We next assess the accuracy of two approaches to standard-error estimation for the MM estimator: heteroskedasticity-robust (HC1) standard errors and bootstrap standard errors (1,000 resamples per replication). For each approach, we compute the ratio of the average estimated standard error to the Monte Carlo standard deviation (MC SD) of the MM coefficient across replications. A ratio close to 1 indicates that the standard errors accurately capture the true sampling variability.

Table K.3 reports these ratios. Both HC1 and bootstrap standard errors perform well across all

sample sizes, numbers of CQs, and data-generating configurations. The bootstrap ratios are generally closer to 1, as expected: because the bootstrap resamples entire individuals, it can account for sources of variation beyond heteroskedasticity, including estimation error for the CQ rating means. The HC1 ratios are also close to 1, though they tend to be slightly above 1 when x_i is correlated with the intercept ($A = 1, B = 0$). The largest HC1 ratio in the table is 1.11 (for $C = 9, A = 1, B = 0, I = 10,000$); outside these cases the HC1 and bootstrap ratios are nearly indistinguishable.

This close agreement extends to our empirical application. In the life satisfaction data analyzed in Section V, the average ratio of HC1 to bootstrap standard errors across the 25 demographic coefficients is 0.99, with a range of $[0.93, 1.02]$. The two approaches thus agree to within about 1% on average, with HC1 only slightly smaller for most coefficients.

On the basis of these findings, we recommend HC1 standard errors as the default for the MM estimator. They are fast to compute and, in both our simulations and our empirical application, closely approximate the more computationally intensive bootstrap.

Table K.3: Ratio of Average Estimated Standard Error to Monte Carlo Standard Deviation for MM

Num. of CQs	$I = 100$					$I = 1,000$		$I = 10,000$	
	A	B	D	HC1	Bootstrap	HC1	Bootstrap	HC1	Bootstrap
2	0	0	1	0.93	0.95	0.96	0.96	1.03	1.03
2	1	0	1	1.07	1.05	1.03	1.00	1.09	1.05
2	0	50	1	0.96	0.95	1.01	0.98	1.03	1.01
2	1	50	1	0.96	0.98	0.99	0.99	1.02	1.02
3	0	0	1	0.96	0.98	1.03	1.03	1.00	1.00
3	1	0	1	1.00	0.98	1.01	0.97	1.03	0.99
3	0	50	1	1.00	1.00	1.06	1.04	1.02	0.99
3	1	50	1	0.98	1.00	0.99	0.99	1.01	1.01
9	0	0	1	0.98	0.98	1.02	1.02	0.98	0.98
9	1	0	1	1.06	1.02	1.07	1.02	1.11	1.05
9	0	50	1	0.93	0.91	1.03	0.99	1.04	1.00
9	1	50	1	1.00	1.00	0.99	0.99	1.02	1.01
27	0	0	1	0.96	0.95	0.99	0.99	1.00	1.00
27	1	0	1	1.01	0.96	1.07	1.02	1.06	1.01
27	0	50	1	0.97	0.95	0.99	0.96	1.04	1.00
27	1	50	1	0.96	0.96	1.00	0.99	0.99	0.99
81	0	0	1	0.99	0.99	1.03	1.02	0.99	0.99
81	1	0	1	1.03	0.98	1.05	0.99	1.03	0.97
81	0	50	1	1.00	0.97	1.04	1.00	1.04	1.00
81	1	50	1	0.95	0.95	1.01	1.01	1.00	1.00

Notes: Ratio of the average estimated standard error to the Monte Carlo standard deviation (MC SD) of the MM coefficient across 1,000 replications, for each combination of $I \in \{100, 1,000, 10,000\}$ and $C \in \{2, 3, 9, 27, 81\}$. “HC1”: heteroskedasticity-robust standard errors. “Bootstrap”: 1,000 resamples per replication. A and B govern the correlation of x_i with the intercept and stretcher; $D = 1$ throughout. Ratio of 1 indicates exact match between estimated standard errors and true sampling variability.

L Additional Tables and Figure

Table L.1: Respondent Demographics

Demographic	Category	UAS 571 Sample (Survey 1)	UAS 719 Sample (Survey 2)	Census (Ages 18+)
Gender	Male	39.1	41.5	49.0
	Female	60.9	58.5	51.0
Marital status	Married, not separated	53.1	56.7	50.2
	Never married	25.9	20.6	31.6
	Other	21.0	22.7	18.2
Highest education level completed	Less than high school	4.1	3.8	10.3
	High school grad	15.3	16.3	27.1
	Some college	34.1	32.4	28.5
	Bachelor's degree	26.6	26.2	21.0
Age	Graduate degree	19.8	21.3	13.1
	18–29	10.3	5.9	20.1
	30–39	18.0	14.5	17.6
	40–49	19.2	19.1	16.1
	50–64	28.0	29.2	23.3
	65 and older	24.4	31.3	22.9
Household income	Less than \$20,000	13.9	12.5	10.5
	\$20,000–\$39,999	15.8	15.3	9.7
	\$40,000–\$59,999	14.0	13.0	11.1
	\$60,000–\$74,999	10.1	9.5	8.5
	\$75,000–\$99,999	12.9	13.1	12.8
	\$100,000–\$149,999	16.4	17.5	19.4
Region	\$150,000 or more	16.8	19.1	28.0
	Midwest	20.2	25.8	20.4
	Northeast	13.0	16.8	17.3
	South	30.5	37.8	38.7
Race	West	36.3	19.6	23.6
	White, and other	67.3	76.3	63.9
	Black	10.7	10.6	11.5
	Hispanic, Latino, or Spanish	14.6	7.6	18.1
Household size	Asian	7.3	5.4	6.5
	1	23.3	22.9	17.4
	2	37.9	39.5	33.0
	3	15.7	15.7	18.5
Employment status	4 and above	23.2	22.0	31.0
	Employed	56.4	57.3	62.3
	Unemployed	6.0	5.6	2.9
Obs.	Not in labor force	37.6	37.1	34.8
		10,100	2,503	

Notes: All numbers are percentages. “Other” marital status includes people who are divorced, widowed, or separated. “Region” is based on state of residence, using the U.S. Census Bureau definition. “Not in labor force” includes Retired, Disabled, Mixed, and “Other Labor Force Status” categories in the UAS data. Race categories are mutually exclusive. “Hispanic, Latino, or Spanish” is based on the Hispanic/Latino indicator; “Black” and “Asian” exclude respondents who also reported Hispanic/Latino; “White, and other” includes White, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and Mixed race categories, but excludes Hispanic/Latino, Black, and Asian. *Sources:* Understanding America Study (UAS); U.S. Census Bureau, 2024 American Community Survey (ACS) 1-Year Estimates and Public Use Microdata Sample (PUMS).

Table L.2: Regression of $\hat{a}_i$ and $\hat{b}_i$ on Demographics

Demographics	$\hat{a}_i$	$\hat{b}_i$
Demeaned age / 10	1.50 ^{†††} (0.19)	−0.02 ^{†††} (0.00)
(Demeaned age) ² / 100	−0.20 [†] (0.09)	0.00 (0.00)
Log(HH income)	−4.77 ^{†††} (0.30)	0.09 ^{†††} (0.01)
Unemployed	−2.01 (1.04)	0.03 (0.02)
Married, not separated	2.29 ^{†††} (0.56)	−0.03 ^{†††} (0.01)
College grad	−5.77 ^{†††} (0.51)	0.10 ^{†††} (0.01)
Male	2.39 ^{†††} (0.48)	−0.05 ^{†††} (0.01)
Part-time employed	−1.50 (0.78)	0.03 (0.01)
Out of labor force	−0.31 (0.66)	0.00 (0.01)
Have ≥ 1 child	0.40 (0.68)	−0.00 (0.01)
Log(HH size)	0.67 (0.58)	−0.01 (0.01)
Asian	8.00 ^{†††} (0.90)	−0.12 ^{†††} (0.02)
Black/African American	11.35 ^{†††} (0.78)	−0.18 ^{†††} (0.01)
Hispanic/Latino/Spanish	7.57 ^{†††} (0.76)	−0.12 ^{†††} (0.01)
Other non-white	11.80 ^{†††} (2.27)	−0.20 ^{†††} (0.04)
4th quartile of population density	0.67 (0.54)	−0.02 (0.01)
Obs.	10,100	10,100

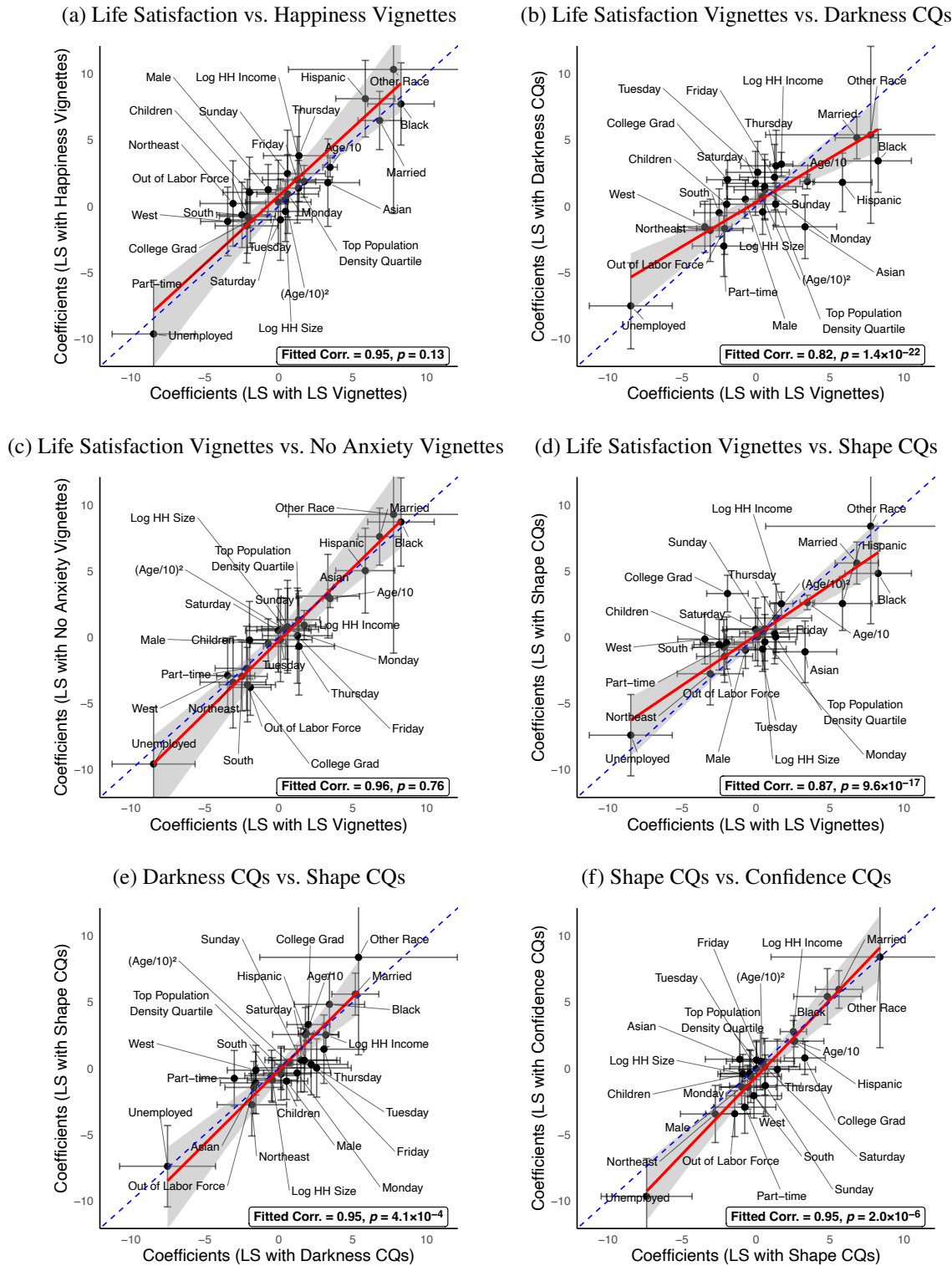
Notes: Sample: 10,100 UAS respondents. $\hat{a}_i$ and $\hat{b}_i$: intercept and slope from regressing respondent i 's 3 Life Satisfaction CQ ratings on the sample means of those 3 ratings. Each regression also includes indicators for region and day of week. [†], ^{††}, ^{†††}: FDR significance at the 10%, 5%, and 1% levels (Benjamini-Hochberg over 25 p -values per column). Heteroskedasticity-robust standard errors in parentheses.

Table L.3: CQ Means and Same-Dimension Benchmark Weights

CQ trio	CQ ratings			Same-dimension benchmark weights			
	Low	Medium	High	SR mean	Low	Medium	High
Life Satisfaction	50.62 (0.21)	59.20 (0.20)	64.35 (0.20)	69.97	-0.587	0.475	1.112
Happiness	40.40 (0.32)	49.45 (0.30)	62.43 (0.29)	69.16	-0.444	0.235	1.208
Worthwhileness	51.55 (0.30)	64.18 (0.27)	68.84 (0.27)	70.63	-0.234	0.484	0.750
No Anxiety	42.96 (0.36)	52.12 (0.30)	66.36 (0.32)	56.91	0.213	0.315	0.473
Anxiety	37.93 (0.34)	56.41 (0.31)	65.55 (0.35)	43.12	0.728	0.253	0.018
Breathe	37.70 (0.25)	55.03 (0.23)	65.68 (0.22)	81.40	-0.749	0.493	1.257
Darkness	25.07 (0.16)	39.91 (0.17)	56.10 (0.19)	69.97	-0.606	0.305	1.301
Shape	39.39 (0.22)	52.43 (0.20)	64.14 (0.20)	69.97	-0.406	0.360	1.046
Confidence	32.00 (0.18)	45.31 (0.19)	64.18 (0.18)	69.97	-0.328	0.252	1.075
Pain	31.91 (0.20)	52.42 (0.19)	65.85 (0.22)	36.07	0.768	0.277	-0.045
Remember	42.42 (0.24)	–	68.69 (0.20)	69.88	-0.045	–	1.045
Life Satisfaction (S2)	48.80 (0.41)	57.58 (0.38)	63.43 (0.38)	72.02	-0.776	0.472	1.304
Life Satisfaction 2 (S2)	39.48 (0.44)	48.45 (0.42)	61.61 (0.39)	72.02	-0.595	0.208	1.386
Height (S2)	46.51 (0.34)	68.19 (0.31)	78.98 (0.39)	61.65	0.430	0.314	0.256
Weight (S2)	50.27 (0.43)	61.01 (0.29)	66.67 (0.51)	58.62	0.378	0.325	0.297

Notes: S2 denotes Survey 2 trios. CQ ratings use a 0–100 response scale. CQ-rating columns: unweighted sample means, with standard errors of the mean in parentheses. Sample sizes: 10,100 (Shape, Darkness, Confidence, Pain, Breathe, Remember, Life Satisfaction); 5,063 (Happiness); 5,062 (No Anxiety); 5,037 (Worthwhileness); 5,038 (Anxiety); 2,503 (All S2 trios). SR mean: unweighted mean, used as the benchmark target. Benchmark weights sum to one and equate the weighted average of the CQ means to that target; they may fall outside the unit interval when the target requires extrapolation. Shape, Darkness, and Confidence have no corresponding SR and use the Life Satisfaction SR mean. Both Life Satisfaction (S2) trios use the mean Survey 2 Life Satisfaction SR. Remember has no medium CQ (its middle item is an attention check). Anxiety items, ordered high-to-low in the raw data, are reordered here as Low, Medium, High.

Figure L.1: Coefficients From Scale-Use Adjusted Life Satisfaction Regression Using Different CQs



Notes: Each panel: pairs of demographic OLS coefficients from two life satisfaction regressions differing only in the set of CQs used for scale-use adjustment (axes labeled accordingly). Bars: 95% CIs. Blue dashed line: 45°; red line: TLS fit, with a 95% confidence band. Fitted Corr.: correlation between the two regressions' fitted values (excluding the intercept); p : test of proportional coefficients (null = the two vectors are proportional; Web Appendix G).

Table L.4: Fitted-Value Correlations and Tests of Proportional Coefficients in Scale-Use-Adjusted (MM) Regressions With Different CQs

CQ dimension	SR dimension				
	Life Satisfaction	Happiness	Worthwhileness	No Anxiety	Anxiety
Life Satisfaction	1.00 (1.00)	0.94 (0.19)	0.86 ($3.9 \times 10^{-4\ddagger\ddagger\ddagger}$)	0.98 (0.06 [†])	0.95 ($3.7 \times 10^{-3\ddagger\ddagger\ddagger}$)
Happiness	0.95 (0.13)	1.00 (1.00)	0.85 (0.07)	0.97 (0.75)	0.94 (0.16)
Worthwhileness	0.87 ($1.9 \times 10^{-7\ddagger\ddagger\ddagger}$)	0.76 ($1.3 \times 10^{-3\ddagger\ddagger\ddagger}$)	1.00 (1.00)	0.90 ($3.5 \times 10^{-4\ddagger\ddagger\ddagger}$)	0.92 (0.12)
No Anxiety	0.96 (0.76)	0.91 (0.22)	0.83 ($9.9 \times 10^{-3\ddagger\ddagger}$)	1.00 (1.00)	-0.78 ($1.8 \times 10^{-8\ddagger\ddagger\ddagger}$)
Anxiety	0.93 ($7.6 \times 10^{-3\ddagger\ddagger}$)	0.92 (0.05 [†])	0.77 ($7.1 \times 10^{-8\ddagger\ddagger\ddagger}$)	-0.78 ($1.8 \times 10^{-8\ddagger\ddagger\ddagger}$)	1.00 (1.00)
Breathe	0.96 ($6.4 \times 10^{-5\ddagger\ddagger\ddagger}$)	0.92 (0.13)	0.86 ($2.1 \times 10^{-4\ddagger\ddagger\ddagger}$)	0.97 ($1.3 \times 10^{-3\ddagger\ddagger\ddagger}$)	0.95 ($1.7 \times 10^{-5\ddagger\ddagger\ddagger}$)
Darkness	0.82 ($1.4 \times 10^{-22\ddagger\ddagger\ddagger}$)	0.79 ($5.7 \times 10^{-7\ddagger\ddagger\ddagger}$)	0.82 ($2.4 \times 10^{-3\ddagger\ddagger\ddagger}$)	0.98 (0.43)	0.94 ($1.2 \times 10^{-9\ddagger\ddagger\ddagger}$)
Shape	0.87 ($9.6 \times 10^{-17\ddagger\ddagger\ddagger}$)	0.83 ($2.9 \times 10^{-6\ddagger\ddagger\ddagger}$)	0.84 ($2.7 \times 10^{-3\ddagger\ddagger\ddagger}$)	0.99 (0.97)	0.97 (0.01 ^{††})
Confidence	0.88 ($8.4 \times 10^{-17\ddagger\ddagger\ddagger}$)	0.81 ($4.3 \times 10^{-8\ddagger\ddagger\ddagger}$)	0.90 (0.12)	0.97 (0.23)	0.94 ($2.5 \times 10^{-9\ddagger\ddagger\ddagger}$)
Pain	0.72 ($4.9 \times 10^{-42\ddagger\ddagger\ddagger}$)	0.68 ($2.3 \times 10^{-14\ddagger\ddagger\ddagger}$)	0.77 ($3.1 \times 10^{-7\ddagger\ddagger\ddagger}$)	0.97 (0.12)	0.94 ($8.8 \times 10^{-8\ddagger\ddagger\ddagger}$)
Remember	0.93 ($5.6 \times 10^{-10\ddagger\ddagger\ddagger}$)	0.92 (0.05 [†])	0.89 ($9.4 \times 10^{-3\ddagger\ddagger}$)	0.97 ($3.1 \times 10^{-3\ddagger\ddagger\ddagger}$)	0.95 ($7.7 \times 10^{-5\ddagger\ddagger\ddagger}$)
Life Satisfaction (S2)	0.92 (0.15)	0.80 ($4.4 \times 10^{-3\ddagger\ddagger}$)	0.77 (0.01 ^{††})	0.92 (0.51)	0.90 (0.44)
Life Satisfaction 2 (S2)	0.90 (0.16)	0.81 (0.07)	0.85 (0.39)	0.94 (0.91)	0.91 (0.18)
Height (S2)	0.92 (0.09)	0.81 ($3.6 \times 10^{-3\ddagger\ddagger\ddagger}$)	0.83 (0.01 ^{††})	0.93 (0.68)	0.93 (0.38)
Weight (S2)	0.92 (0.14)	0.85 (0.04 [†])	0.75 ($2.5 \times 10^{-4\ddagger\ddagger\ddagger}$)	0.94 (0.63)	0.91 (0.22)

CQ dimension	SR dimension			
	Breathe	Remember	Height	Weight
Life Satisfaction	0.92 (0.09)	0.88 ($6.8 \times 10^{-5\ddagger\ddagger\ddagger}$)	0.88 (0.22)	0.80 (0.38)
Happiness	0.80 (0.03 [†])	0.88 (0.19)	0.80 (0.22)	0.51 (0.10)
Worthwhileness	0.64 ($1 \times 10^{-4\ddagger\ddagger\ddagger}$)	0.69 ($1.2 \times 10^{-6\ddagger\ddagger\ddagger}$)	0.71 (0.02 ^{††})	0.33 ($9 \times 10^{-3\ddagger\ddagger}$)
No Anxiety	0.83 (0.02 ^{††})	0.89 (0.15)	0.76 (0.33)	0.68 (0.68)
Anxiety	0.91 (0.09)	0.87 ($2.3 \times 10^{-3\ddagger\ddagger\ddagger}$)	0.77 (0.03 [†])	0.71 (0.04 [†])
Breathe	1.00 (1.00)	0.88 ($4 \times 10^{-4\ddagger\ddagger\ddagger}$)	0.85 ($4.8 \times 10^{-3\ddagger\ddagger}$)	0.74 (0.01 ^{††})
Darkness	0.22 ($1.1 \times 10^{-23\ddagger\ddagger\ddagger}$)	0.63 ($5.2 \times 10^{-15\ddagger\ddagger\ddagger}$)	0.83 (0.21)	0.69 (0.74)
Shape	0.47 ($1.1 \times 10^{-19\ddagger\ddagger\ddagger}$)	0.65 ($5.9 \times 10^{-14\ddagger\ddagger\ddagger}$)	0.82 (0.01 ^{††})	0.69 (0.12)
Confidence	0.39 ($1.2 \times 10^{-22\ddagger\ddagger\ddagger}$)	0.73 ($6 \times 10^{-11\ddagger\ddagger\ddagger}$)	0.83 (0.12)	0.71 ($1.1 \times 10^{-3\ddagger\ddagger\ddagger}$)
Pain	-0.15 ($2.6 \times 10^{-20\ddagger\ddagger\ddagger}$)	0.49 ($9.8 \times 10^{-21\ddagger\ddagger\ddagger}$)	0.86 (0.17)	0.76 (0.60)
Remember	0.78 ($2 \times 10^{-8\ddagger\ddagger\ddagger}$)	1.00 (1.00)	0.87 (0.07)	0.72 (0.61)
Life Satisfaction (S2)	0.75 (0.09)	0.66 ($4.5 \times 10^{-6\ddagger\ddagger\ddagger}$)	0.85 (0.01 ^{††})	0.79 (0.03 [†])
Life Satisfaction 2 (S2)	0.61 (0.32)	0.69 (0.04 [†])	0.88 (0.35)	0.76 (0.80)
Height (S2)	0.72 ($6.5 \times 10^{-4\ddagger\ddagger\ddagger}$)	0.80 ($1.4 \times 10^{-3\ddagger\ddagger\ddagger}$)	1.00 (1.00)	0.91 (0.71)
Weight (S2)	0.79 (0.03 [†])	0.80 ($6.2 \times 10^{-3\ddagger\ddagger}$)	0.94 (0.21)	1.00 (1.00)

Notes: Rows: the CQ trio used to construct the scale-use-only benchmark. Columns: the SR dimension. Each cell: correlation between the fitted values from the MM regression using the row CQ trio and the fitted values from the MM regression using the same-dimension CQ trio (intercept excluded), followed in parentheses by the p -value from a test of proportional coefficients. Data: demographics and SRs from UAS 571, except Height and Weight SRs (which are from UAS 719, 2,503 linked respondents); row CQ trios from UAS 571, except Life Satisfaction (S2), Life Satisfaction 2 (S2), Height, and Weight (which are from UAS 719). Remember uses only two CQs since the nominal middle vignette is an attention check. [†], ^{††}, ^{†††}: significance at the 10%, 5%, and 1% levels after Benjamini-Hochberg FDR adjustment across the non-identity entries (the 9 identity-entry p -values are excluded). Test of proportional coefficients (see Web Appendix G). The direction of alignment is given by the sign of the estimated proportionality factor λ . Expected proportionality slopes are positive in every cell except the two Anxiety–No Anxiety cells, where the opposite-valence SRs imply a negative slope. The estimated slopes have the expected sign everywhere except in Darkness–Breathe and Pain–Breathe. The reported correlation and the estimated λ agree in sign in every cell except Darkness–Breathe, where the correlation is 0.22 while λ is -0.05 , essentially zero. Anxiety–No Anxiety is a special case: because the two SRs are observed in disjoint samples, the table compares the Anxiety SR adjusted with Anxiety CQs to the No Anxiety SR adjusted with No Anxiety CQs, treating the two as equivalent; the same correlation and p -value appear in both cells.

Table L.5: Estimated Proportionality Factors From Tests of Proportional Coefficients in Scale-Use-Adjusted (MM) Regressions With Different CQs

CQ dimension	SR dimension				
	Life Satisfaction	Happiness	Worthwhileness	No Anxiety	Anxiety
Life Satisfaction	1.00 (—)	1.07 (0.07)	1.76 (0.17)	1.11 (0.05)	1.25 (0.07)
Happiness	0.94 (0.06)	1.00 (—)	1.47 (0.20)	1.22 (0.08)	1.09 (0.07)
Worthwhileness	0.68 (0.05)	0.66 (0.09)	1.00 (—)	1.09 (0.09)	1.28 (0.09)
No Anxiety	0.89 (0.06)	0.95 (0.10)	1.45 (0.18)	1.00 (—)	-1.30 (0.12)
Anxiety	1.27 (0.07)	1.50 (0.13)	2.58 (0.36)	-1.30 (0.12)	1.00 (—)
Breathe	0.88 (0.04)	0.92 (0.06)	1.45 (0.13)	1.26 (0.05)	1.14 (0.06)
Darkness	0.81 (0.07)	0.79 (0.09)	1.31 (0.13)	1.12 (0.06)	1.36 (0.05)
Shape	0.87 (0.06)	0.88 (0.09)	1.51 (0.14)	1.24 (0.05)	1.15 (0.05)
Confidence	0.82 (0.05)	0.79 (0.08)	1.24 (0.11)	1.04 (0.05)	1.36 (0.05)
Pain	0.92 (0.10)	0.92 (0.13)	1.34 (0.16)	1.16 (0.05)	1.10 (0.05)
Remember	0.77 (0.04)	0.80 (0.06)	1.12 (0.11)	0.99 (0.05)	1.21 (0.06)
Life Satisfaction (S2)	1.11 (0.09)	1.31 (0.15)	1.87 (0.25)	1.34 (0.13)	1.21 (0.12)
Life Satisfaction 2 (S2)	0.91 (0.09)	0.96 (0.15)	1.48 (0.27)	1.36 (0.12)	1.25 (0.12)
Height (S2)	1.01 (0.08)	1.08 (0.12)	1.55 (0.25)	1.34 (0.13)	1.32 (0.10)
Weight (S2)	1.44 (0.15)	1.58 (0.19)	2.21 (0.54)	1.37 (0.12)	1.52 (0.17)

CQ dimension	SR dimension			
	Breathe	Remember	Height	Weight
Life Satisfaction	1.42 (0.13)	1.75 (0.16)	0.91 (0.14)	0.68 (1.29)
Happiness	1.07 (0.14)	1.50 (0.17)	1.10 (0.21)	0.53 (3.46)
Worthwhileness	0.69 (0.14)	0.94 (0.17)	0.74 (0.17)	0.33 (5.51)
No Anxiety	1.25 (0.17)	1.51 (0.21)	0.84 (0.23)	0.78 (1.82)
Anxiety	2.05 (0.20)	2.66 (0.29)	1.24 (0.30)	2.10 (3.17)
Breathe	1.00 (—)	1.53 (0.13)	1.35 (0.21)	1.14 (4.23)
Darkness	-0.05 (0.57)	1.46 (0.40)	0.87 (0.16)	0.45 (1.24)
Shape	0.64 (0.31)	1.93 (0.28)	0.80 (0.17)	0.42 (2.03)
Confidence	0.48 (1.17)	1.31 (0.18)	0.74 (0.14)	1.97 (3.05)
Pain	-0.75 (0.20)	1.67 (0.86)	0.99 (0.13)	0.68 (0.49)
Remember	0.87 (0.10)	1.00 (—)	1.02 (0.14)	0.48 (2.19)
Life Satisfaction (S2)	1.25 (0.20)	2.17 (0.41)	0.89 (0.14)	0.93 (2.00)
Life Satisfaction 2 (S2)	0.70 (0.21)	1.30 (0.32)	0.84 (0.13)	0.45 (0.69)
Height (S2)	0.96 (0.16)	1.71 (0.27)	1.00 (—)	0.78 (0.44)
Weight (S2)	2.34 (0.44)	2.79 (0.53)	1.15 (0.12)	1.00 (—)

Notes: Rows: the CQ trio used to construct the scale-use-only benchmark. Columns: the SR dimension. Each cell: the estimated proportionality factor $\hat{\lambda}$ from the test of proportional coefficients (see Web Appendix G) comparing the MM coefficients obtained using the row CQ trio ($\mathbf{b}_A$) with those obtained using the same-dimension CQ trio ($\mathbf{b}_B$), intercept excluded, followed in parentheses by its bootstrap standard error (the standard deviation of $\hat{\lambda}$ across the 500 respondent-level bootstrap draws, holding $\hat{\mathbf{V}}_\lambda$ fixed at its bootstrap estimate). The corresponding fitted-value correlations and p -values are reported in Table L.4. Identity entries equal 1 by construction and have no standard error. Data: demographics and SRs from UAS 571, except Height and Weight SRs (which are from UAS 719, 2,503 linked respondents); row CQ trios from UAS 571, except Life Satisfaction (S2), Life Satisfaction 2 (S2), Height, and Weight (which are from UAS 719). Remember uses only two CQs since the nominal middle vignette is an attention check. Expected proportionality factors are positive in every cell except the two Anxiety–No Anxiety cells, where the opposite-valence SRs imply a negative factor; the estimated factors have the expected sign everywhere except in Darkness–Breathe and Pain–Breathe. Anxiety–No Anxiety is a special case: because the two SRs are observed in disjoint samples, the table compares the Anxiety SR adjusted with Anxiety CQs to the No Anxiety SR adjusted with No Anxiety CQs, treating the two as equivalent; the same $\hat{\lambda}$ appears in both cells.

Table L.6: Validation Correlations and Response-Error Sensitivity

	Height				Weight			
	0%	25%	50%	75%	0%	25%	50%	75%
Assumed SR response-error variance								
Unadjusted	0.49 (0.02)	0.49 (0.02)	0.49 (0.02)	0.49 (0.02)	0.23 (0.02)	0.23 (0.02)	0.23 (0.02)	0.23 (0.02)
Unadjusted, Response-Error Corrected	0.49 (0.02)	0.52 (0.02)	0.56 (0.02)	0.60 (0.02)	0.23 (0.02)	0.27 (0.03)	0.32 (0.03)	0.43 (0.05)
Method of Moments (MM), Same-Dimension CQs	0.54 (0.02)	0.57 (0.02)	0.60 (0.02)	0.65 (0.02)	0.34 (0.03)	0.40 (0.04)	0.51 (0.05)	0.88 (0.22)
<i>Method of Moments (MM), Other-Dimension CQs</i>								
Life Satisfaction Vignette CQs	0.42 (0.02)	0.43 (0.02)	0.45 (0.02)	0.48 (0.03)	0.20 (0.02)	0.22 (0.03)	0.24 (0.03)	0.28 (0.03)
Happiness Vignette CQs	0.53 (0.04)	0.55 (0.04)	0.58 (0.05)	0.61 (0.05)	0.25 (0.05)	0.27 (0.05)	0.31 (0.06)	0.35 (0.07)
Worthwhileness Vignette CQs	0.48 (0.03)	0.51 (0.03)	0.53 (0.03)	0.57 (0.03)	0.23 (0.04)	0.26 (0.04)	0.30 (0.05)	0.37 (0.06)
No Anxiety Vignette CQs	0.54 (0.03)	0.56 (0.04)	0.59 (0.04)	0.63 (0.04)	0.24 (0.04)	0.27 (0.05)	0.30 (0.06)	0.36 (0.07)
Anxiety Vignette CQs	0.50 (0.04)	0.52 (0.05)	0.54 (0.05)	0.57 (0.05)	0.28 (0.05)	0.30 (0.05)	0.33 (0.06)	0.38 (0.07)
Breathe Vignette CQs	0.49 (0.02)	0.51 (0.02)	0.53 (0.02)	0.56 (0.03)	0.24 (0.03)	0.27 (0.03)	0.30 (0.03)	0.34 (0.04)
Darkness Visual CQs	0.51 (0.03)	0.54 (0.03)	0.57 (0.03)	0.61 (0.03)	0.20 (0.03)	0.23 (0.03)	0.26 (0.04)	0.32 (0.05)
Shape Visual CQs	0.44 (0.02)	0.46 (0.02)	0.48 (0.02)	0.50 (0.03)	0.17 (0.03)	0.18 (0.03)	0.20 (0.03)	0.23 (0.04)
Confidence Visual CQs	0.44 (0.02)	0.46 (0.02)	0.49 (0.02)	0.52 (0.02)	0.21 (0.03)	0.23 (0.03)	0.27 (0.04)	0.34 (0.05)
Pain Visual CQs	0.53 (0.03)	0.55 (0.03)	0.58 (0.03)	0.62 (0.03)	0.22 (0.03)	0.25 (0.04)	0.28 (0.04)	0.33 (0.05)
Remember Vignette CQs	0.51 (0.02)	0.53 (0.03)	0.56 (0.03)	0.60 (0.03)	0.25 (0.03)	0.28 (0.03)	0.32 (0.04)	0.40 (0.05)
Life Satisfaction (S2) Vignette CQs	0.47 (0.03)	0.50 (0.03)	0.53 (0.03)	0.56 (0.03)	0.22 (0.02)	0.24 (0.03)	0.27 (0.03)	0.33 (0.04)
Life Satisfaction 2 (S2) Vignette CQs	0.52 (0.03)	0.55 (0.03)	0.58 (0.03)	0.62 (0.03)	0.21 (0.03)	0.24 (0.03)	0.28 (0.04)	0.35 (0.05)
Opposite Physical-Dimension CQs (Weight, Height)	0.67 (0.03)	0.72 (0.03)	0.79 (0.03)	0.87 (0.04)	0.25 (0.02)	0.29 (0.03)	0.33 (0.03)	0.42 (0.04)

Notes: Correlations of each SR on the subjective 0-100 scale with the corresponding standardized objective-scale measure; bootstrap standard errors (100 resamples) in parentheses. Each entry is computed as the coefficient from a regression of the dependent variable described below on the objective-scale measure, with both variables standardized, which makes that coefficient their correlation. Section I.D describes the validation measures and same-dimension CQs; Web Appendix Tables B.2a, B.2b, and B.4 display the CQ stimuli. Row labels indicate how the dependent variable is constructed. *Unadjusted*: raw SR divided by its raw SD. *Unadjusted, Response-Error Corrected*: raw SR divided by its response-error-corrected SD. *Method of Moments (MM)*: raw SR minus the indicated CQ benchmark, divided by the MM-corrected SD (Web Appendix D.4). Column labels: the assumed SR response-error variance, as a percentage of the response-error variance estimated from the corresponding same-dimension CQs; at 0%, none is subtracted, so the *Unadjusted, Response-Error Corrected* row equals the *Unadjusted* row. (The *Unadjusted* row is unchanged across assumptions.) Opposite physical-dimension row: Weight CQs for height and Height CQs for weight. Other-dimension CQ measurement parameters use all available respondents for the indicated CQ set, from Survey 1 except for the two Life Satisfaction (S2) trios, which come from Survey 2. Remember excludes its middle, attention-check CQ. Height/weight sample sizes: 2,474/2,481, except Happiness (1,244/1,247), Worthwhileness (1,230/1,234), No Anxiety (1,225/1,232), and Anxiety (1,249/1,249).

Table L.7: Existing Datasets With Calibration-Question Data

Dataset	Sample Size	Dates	CQ Dimensions	Response Scale
UAS 571 (Survey 1)	10,232	2023–2024	Life satisfaction, Happiness, Worthwhileness, Anxiety, No Anxiety, Breathe, Remember; Darkness, Shape, Confidence, Pain (visual)	0–100
UAS 719 (Survey 2)	2,537	2025	Life satisfaction (two trios), Height (visual), Weight (visual)	0–100
MTurk Baseline	3,358	2022	Memory, Internet access, Crime safety (includes visual CQs)	0–100
MTurk Bottomless/Multi-Block	2,735	2022–2023	Various (42 dimensions) and visual CQs	0–100, 0–10, 0–4, 0–3
Gallup World Poll	56,500	2011–2014	Life satisfaction (Cantril Ladder)	0–10
ALP Survey 6	1,113	2006–2007	Global Satisfaction, Income, Job Security	1–5
Dutch Centerpanel	1,978	2003	Work Disability (Pain, Affect, Heart Disease)	1–5
RAND MS Internet Panel	672	2004	Work Disability (Pain, Affect, Heart Disease)	Yes/No
WHO Pilot (Mexico/China)	981	2002	Political Efficacy	1–5
WHO Vision (China/Slovakia)	8,301	2000–2001	Vision Acuity	A–E (None to Extreme)
HRS Disability Vignette Survey	4,639	2007–2008	Disability	Mild to Extreme
SHARE Waves 1 and 2	4,512; 7,646	2004; 2006	Health and Disability	Various

Note: Datasets known to the authors that collected calibration-question data. “Visual”: calibration questions using images rather than text descriptions; all other CQs listed are vignettes.

Table L.8: Cross-Framing Tests of Proportional Coefficients in Scale-Use-Adjusted (MM) Regressions (*p*-Values)

Trio	Wave	Self/You	Self/Male	Self/Female	You/Male	You/Female	Male/Female
Life Satisfaction	UAS 571	0.152	0.084	0.249	0.075	0.394	0.102
Happiness	UAS 571	0.063	0.489	0.445	0.647	0.559	0.857
Worthwhileness	UAS 571	0.584	0.062	0.369	4.5×10^{-3}	0.563	0.630
No Anxiety	UAS 571	0.112	0.799	0.188	0.372	0.022	0.661
Anxiety	UAS 571	0.069	0.273	0.494	0.847	0.661	0.338
Breathe	UAS 571	0.348	0.349	0.042	0.467	0.828	0.478
Remember	UAS 571	0.656	0.484	0.193	0.761	0.730	0.725
Life Satisfaction (S2)	UAS 719	0.420	0.092	0.187	0.037	0.051	0.125
Life Satisfaction 2 (S2)	UAS 719	0.126	0.423	0.343	0.670	0.277	0.835

Notes: Each cell: *p*-value from a test of proportional coefficients (see Web Appendix G) in scale-use-adjusted (MM) regressions across two MM regressions, where the vignettes are framed differently. None is significant at the 10% level after FDR adjustment applied separately within each UAS wave. In each row, the corresponding SR is adjusted using the indicated same-dimension CQ trio (or duo) and regressed on the same demographic controls separately within each framing, restricting the sample to respondents assigned that framing. Remember uses its two substantive CQs, excluding the attention check. Visual CQs are omitted (there was no framing randomization); in UAS 719, framing results are available only for the two Life Satisfaction trios.

M Proofs

This appendix contains proofs of Propositions 1 and 2 and Corollary 1. It also contains a statement and proof of the claim, Proposition 3, that the assumptions are necessary and sufficient for the conclusion of Proposition 1 once Assumption 1 is suitably weakened. For reference, we begin by re-stating the assumptions from the main text, and then we state the weakened assumption, Assumption 1*, before giving the proofs.

Assumption 1. (Common Ranking of CQs) For all $i, i' \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$,

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega').$$

Assumption 2. (Common Same-Rating Map) For all $i, i' \in \mathcal{I}$, $\omega \in \mathcal{S}$, and $c \in \mathcal{C}$,

$$\rho_i^*(\omega) = \rho_i(c|\omega) \iff \rho_{i'}^*(\omega) = \rho_{i'}(c|\omega).$$

Assumption 3. (SR-CQ Concordance) For all $i \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$, with $\rho_i^*(\omega) = \rho_i(c|\omega)$ and $\rho_i^*(\omega') = \rho_i(c'|\omega')$,

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \text{ and } \rho_i(c'|\omega') \geq \rho_i(c|\omega') \iff \rho_i^*(\omega') \geq \rho_i^*(\omega).$$

Assumption 1'. (Linear Common Ranking of CQs) For all $i, i' \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, $c, c', c'' \in \mathcal{C}$, and $\varpi \in \mathbb{R}$,

$$\begin{aligned} \rho_i(c'|\omega) \geq \rho_i(c|\omega) \text{ and } \rho_i(c''|\omega) &= \varpi \rho_i(c'|\omega) + (1 - \varpi) \rho_i(c|\omega) \iff \\ \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega') \text{ and } \rho_{i'}(c''|\omega') &= \varpi \rho_{i'}(c'|\omega') + (1 - \varpi) \rho_{i'}(c|\omega'). \end{aligned}$$

The weakened version of Assumption 1, stated next, is called Assumption 1*. Suppose ω and ω' are the current situations of two respondents. The assumption states that if c and c' are CQs that map to ω and ω' for one of the respondents, then both respondents rank the two CQs the same way at their current situations.

Assumption 1*. (Common Ranking of CQs Along Common Same-Rating Map) For all $i, i' \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$, with $\rho_i^*(\omega) = \rho_i(c|\omega)$ and $\rho_i^*(\omega') = \rho_i(c'|\omega')$,

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega').$$

M.1 Proof of Proposition 1

Proposition 1. *Suppose Assumptions 1–3 hold. For all $i, i' \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$ such that $\rho_i^*(\omega) = \rho_i(c|\omega)$ and $\rho_{i'}^*(\omega') = \rho_{i'}(c'|\omega')$,*

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_i^*(\omega') \geq \rho_i^*(\omega) \text{ and } \rho_{i'}^*(\omega') \geq \rho_{i'}^*(\omega). \quad (\text{P})$$

Proof. Fix $i, i' \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$. For reference, we restate the maintained hypotheses:

$$\rho_i^*(\omega) = \rho_i(c|\omega), \quad (10)$$

$$\rho_{i'}^*(\omega') = \rho_{i'}(c'|\omega'). \quad (11)$$

Applying Assumption 2 to (10) and (11) respectively implies:

$$\rho_{i'}^*(\omega) = \rho_{i'}(c|\omega), \quad (12)$$

$$\rho_i^*(\omega') = \rho_i(c'|\omega'). \quad (13)$$

That is, $\omega \leftrightarrow c$ and $\omega' \leftrightarrow c'$.

We now prove each direction of the equivalence in turn, beginning with the only if ($\Rightarrow$) direction. Suppose the left-hand side of the equivalence holds: $\rho_i(c'|\omega) \geq \rho_i(c|\omega)$. By Assumption 1, the ordering of c' relative to c is common across respondents and situations:

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega), \quad \rho_i(c'|\omega') \geq \rho_i(c|\omega'), \quad (14)$$

$$\rho_{i'}(c'|\omega) \geq \rho_{i'}(c|\omega), \quad \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega'). \quad (15)$$

For respondent i , the same-rating pairs (10) and (13) hold, and both CQ orderings in (14) hold; the left-hand side of Assumption 3 is therefore satisfied, so

$$\rho_i^*(\omega') \geq \rho_i^*(\omega).$$

For respondent i' , the same-rating pairs (11) and (12) hold, and both CQ orderings in (15) hold; Assumption 3 (applied to i') thus gives

$$\rho_{i'}^*(\omega') \geq \rho_{i'}^*(\omega).$$

This is the right-hand side of the equivalence.

Turning to the if ($\Leftarrow$) direction, suppose $\rho_i^*(\omega') \geq \rho_i^*(\omega)$. For respondent i , the same-rating pairs (10) and (13) hold, so Assumption 3 implies

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega),$$

which is the left-hand side of the equivalence. (Symmetrically, $\rho_{i'}^*(\omega') \geq \rho_{i'}^*(\omega)$, together with the same-rating pairs (11) and (12), implies $\rho_{i'}(c'|\omega) \geq \rho_{i'}(c|\omega)$. Then, applying Assumption 1 also gives the left-hand side of the equivalence.) $\square$

M.2 Proof of Corollary 1

Corollary 1. *If Assumptions 1–3 hold, then the Interpersonal Comparability Presupposition holds.*

Proof. Fix respondents $i, i' \in \mathcal{I}$ and situations $\omega, \omega' \in \mathcal{S}$. It suffices to show that $\rho_i^*(\cdot)$ and $\rho_{i'}^*(\cdot)$ rank ω and ω' identically.

By our assumption on $\mathcal{C}$, there exists a CQ $c \in \mathcal{C}$ with

$$\rho_i(c \mid \omega) = \rho_i^*(\omega), \quad (16)$$

that is, $\omega \leftrightarrow c$. By Assumption 2, the same-rating map holds for every respondent; in particular,

$$\rho_{i'}(c \mid \omega) = \rho_{i'}^*(\omega). \quad (17)$$

Likewise, by our assumption on $\mathcal{C}$, there exists a CQ $c' \in \mathcal{C}$ with

$$\rho_{i'}(c' \mid \omega') = \rho_{i'}^*(\omega'), \quad (18)$$

i.e., $\omega' \leftrightarrow c'$, and by Assumption 2 also

$$\rho_i(c' \mid \omega') = \rho_i^*(\omega'). \quad (19)$$

Applying Proposition 1 to equations (16) and (18) yields

$$\rho_i(c' \mid \omega) \geq \rho_i(c \mid \omega) \iff \rho_i^*(\omega') \geq \rho_i^*(\omega) \text{ and } \rho_{i'}^*(\omega') \geq \rho_{i'}^*(\omega). \quad (20)$$

The right-hand side of equation (20) implies that the two SR comparisons realize the same (in)equality relation as one another:

$$\rho_i^*(\omega') \geq \rho_i^*(\omega) \iff \rho_{i'}^*(\omega') \geq \rho_{i'}^*(\omega). \quad (21)$$

Equation (21) is the Interpersonal Comparability Presupposition for the situations ω and ω' . (Equation (21) would also follow from applying Proposition 1 to equations (17) and (19).) $\square$

M.3 Proof of Proposition 2

Proposition 2. *Assumption 1' implies Assumption 1. Moreover, denoting the reference individual by $\bar{i}$, Assumption 1' holds if and only if, for all i , there exist $a_i, b_i : \mathcal{S} \rightarrow \mathbb{R}$ with $b_i > 0$ such that for all $\omega \in \mathcal{S}$ and $c \in \mathcal{C}$,*

$$\rho_i(c \mid \omega) = a_i(\omega) + b_i(\omega)\rho_{\bar{i}}(c \mid \bar{\omega}). \quad (6)$$

Proof. To see that Assumption 1' implies Assumption 1, fix any $i, i' \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$. Applying Assumption 1' with $c'' = c'$ and $\varpi = 1$,

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \implies \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega').$$

Since the respondent–situation pairs (i, ω) and (i', ω') were arbitrary, the same implication holds with the pairs exchanged, and hence

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega').$$

Since c and c' were also arbitrary, the same equivalence holds with c and c' switched:

$$\rho_i(c|\omega) \geq \rho_i(c'|\omega) \iff \rho_{i'}(c|\omega') \geq \rho_{i'}(c'|\omega').$$

Combining the two equivalences delivers each of the cases “>,” “=,” and “<”:

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega').$$

This is Assumption 1.

Next, we prove the equivalence stated in the proposition. Fixing an arbitrary respondent–situation pair (i, ω_i) , for notational compactness below we will often use the notation:

$$r_{ic} := \rho_i(c|\omega_i), \quad r_i^\star := \rho_i^\star(\omega_i), \quad w_c := \rho_{\bar{i}}(c|\bar{\omega}).$$

In this notation, equation (6) in Proposition 2 reads

$$r_{ic} = a_i(\omega_i) + b_i(\omega_i) w_c. \tag{22}$$

We prove each direction of the equivalence in turn, beginning with the only if ($\implies$) direction. Because Assumption 1 holds, the CQ translation function

$$\tau_{\bar{i} \rightarrow i}(w|\omega_i) := \rho_i\left(\left(\rho_{\bar{i}}^{-1}(\cdot|\bar{\omega})\right)(w) \mid \omega_i\right)$$

is well-defined and strictly increasing in w : by definition of $\mathcal{R}_{\bar{i}}(\bar{\omega})$, each common-scale value $w \in \mathcal{R}_{\bar{i}}(\bar{\omega})$ is attained by some CQ, and Assumption 1 guarantees that i 's rating of it orders as w does. Note that $\tau_{\bar{i} \rightarrow i}(w_c|\omega_i) = r_{ic}$ by construction.

Next, we show that the CQ translation function is linear. The set $\mathcal{R}_{\bar{i}}(\bar{\omega})$ of common-scale values contains at least two points; fix $w_0, w_1 \in \mathcal{R}_{\bar{i}}(\bar{\omega})$ with $w_1 > w_0$, and set

$$b_i(\omega_i) := \frac{\tau_{\bar{i} \rightarrow i}(w_1|\omega_i) - \tau_{\bar{i} \rightarrow i}(w_0|\omega_i)}{w_1 - w_0}, \tag{23}$$

which is strictly positive because $\tau_{\bar{i} \rightarrow i}(\cdot|\omega_i)$ is strictly increasing. (Below, we will show that $b_i(\omega_i)$ does not depend on the choice of w_0 and w_1 .)

Now, fix any target common-scale value w_c , and choose

$$\varpi = \frac{w_c - w_0}{w_1 - w_0} \in \mathbb{R}.$$

Let c', c, c'' be CQs whose common-scale ratings are w_1, w_0 , and w_c , respectively; these exist by definition of $\mathcal{R}_{\bar{i}}(\bar{\omega})$. Then

$$\varpi w_1 + (1 - \varpi) w_0 = w_0 + \varpi (w_1 - w_0) = w_0 + (w_c - w_0) = w_c,$$

so, for the reference individual, the premise of Assumption 1' holds: $\rho_{\bar{i}}(c'|\bar{\omega}) \geq \rho_{\bar{i}}(c|\bar{\omega})$ (because $w_1 > w_0$) and $\rho_{\bar{i}}(c''|\bar{\omega}) = \varpi \rho_{\bar{i}}(c'|\bar{\omega}) + (1 - \varpi) \rho_{\bar{i}}(c|\bar{\omega})$. Applying Assumption 1' with $\bar{i}$ in the role of i and i in the role of i' , the same linear relation must hold for respondent i :

$$\rho_i(c''|\omega_i) = \varpi \rho_i(c'|\omega_i) + (1 - \varpi) \rho_i(c|\omega_i).$$

That is, in translation-function form,

$$\tau_{\bar{i} \rightarrow i}(w_c|\omega_i) = \varpi \tau_{\bar{i} \rightarrow i}(w_1|\omega_i) + (1 - \varpi) \tau_{\bar{i} \rightarrow i}(w_0|\omega_i).$$

Substituting $\varpi = (w_c - w_0)/(w_1 - w_0)$ and, from (23), $\tau_{\bar{i} \rightarrow i}(w_1|\omega_i) = \tau_{\bar{i} \rightarrow i}(w_0|\omega_i) + (w_1 - w_0) b_i(\omega_i)$, the two multiples of $\tau_{\bar{i} \rightarrow i}(w_0|\omega_i)$ sum to $\tau_{\bar{i} \rightarrow i}(w_0|\omega_i)$, and the remaining term is $b_i(\omega_i)(w_c - w_0)$. Hence

$$\tau_{\bar{i} \rightarrow i}(w_c|\omega_i) = \tau_{\bar{i} \rightarrow i}(w_0|\omega_i) + b_i(\omega_i) [w_c - w_0]. \quad (24)$$

Setting

$$a_i(\omega_i) := \tau_{\bar{i} \rightarrow i}(w_0|\omega_i) - b_i(\omega_i) w_0$$

and recalling that $\tau_{\bar{i} \rightarrow i}(w_c|\omega_i) = r_{ic}$, equation (24) becomes (22). Since (24) holds at every common-scale value w_c and a linear function is determined by its values at any two points, $a_i(\omega_i)$ and $b_i(\omega_i)$ do not depend on the choice of w_0 and w_1 ; and $b_i : \mathcal{S} \rightarrow \mathbb{R}_{++}$. This completes the proof of the only-if direction.

Turning to the if ($\Leftarrow$) direction, suppose that every respondent–situation pair admits the linear representation (22) with stretcher $b_i(\omega_i) > 0$. Fix pairs (i, ω) and (i', ω') and CQs c, c' , and c'' . We verify the two clauses of Assumption 1', the ranking clause and the linear clause.

We first verify the ranking clause. By (22), differencing the representation for i across two CQs gives

$$\rho_i(c'|\omega) - \rho_i(c|\omega) = b_i(\omega) [w_{c'} - w_c].$$

Since $b_i(\omega) > 0$,

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff w_{c'} \geq w_c. \quad (25)$$

The right-hand side of (25) involves only the common-scale quantities $w_{c'}, w_c$ and is therefore the

same for the pair (i', ω') . Hence

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \implies w_{c'} \geq w_c \implies \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega'),$$

which proves the ranking clause of Assumption 1'.

We next verify the linear clause. By hypothesis, $\rho_i(c''|\omega) = \varpi \rho_i(c'|\omega) + (1 - \varpi) \rho_i(c|\omega)$. Substituting (22) on both sides and collecting terms,

$$a_i(\omega) + b_i(\omega) w_{c''} = a_i(\omega) + b_i(\omega) [\varpi w_{c'} + (1 - \varpi) w_c].$$

Canceling $a_i(\omega)$ and dividing by $b_i(\omega) > 0$ yields

$$w_{c''} = \varpi w_{c'} + (1 - \varpi) w_c. \quad (26)$$

Since (26) contains only common-scale quantities, it holds independently of the pair. Applying the representation for (i', ω') and reversing the steps above,

$$\begin{aligned} \varpi \rho_{i'}(c'|\omega') + (1 - \varpi) \rho_{i'}(c|\omega') &= a_{i'}(\omega') + b_{i'}(\omega') [\varpi w_{c'} + (1 - \varpi) w_c] \\ &= a_{i'}(\omega') + b_{i'}(\omega') w_{c''} = \rho_{i'}(c''|\omega'), \end{aligned}$$

where the middle equality uses (26). This proves the linear clause of Assumption 1'.

□

M.4 Proof of Proposition 3

Before proving Proposition 3, we make explicit a useful fact and prove a useful lemma.

Reversal Fact. For any numbers X, Y, Z, W ,

$$X \gneq Y \iff Z \gneq W \quad \text{implies} \quad Y \gneq X \iff W \gneq Z.$$

Indeed $X \gneq Y \iff Z \gneq W$ abbreviates three separate biconditionals, one for each of “>,” “=,” “<”; reading each with its two sides exchanged (so “>” becomes “<” and vice versa) gives exactly the three biconditionals comprising $Y \gneq X \iff W \gneq Z$.

The following lemma, along with the Reversal Fact above, isolates the recurring “apply the hypothesis to a pair of CQs, then again with the pair reversed, then chain” step used throughout the proof of sufficiency for Proposition 3. The lemma is a consequence of Assumption 1* alone.

Lemma 1. *Suppose Assumption 1* holds. Fix $i \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$ such that $\rho_i^*(\omega) = \rho_i(c|\omega)$ and $\rho_i^*(\omega') = \rho_i(c'|\omega')$. Then for all $i' \in \mathcal{I}$ and $\sigma \in \{\omega, \omega'\}$,*

$$\rho_i(c'|\omega) \gneq \rho_i(c|\omega) \iff \rho_{i'}(c'|\sigma) \gneq \rho_{i'}(c|\sigma).$$

Proof. Applying Assumption 1* directly to $(i, i', \omega, \omega', c, c')$ gives the case $\sigma = \omega'$:

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_{i'}(c'|\omega') \geq \rho_{i'}(c|\omega'). \quad (27)$$

Specializing (27) to $i' = i$ gives

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_i(c'|\omega') \geq \rho_i(c|\omega'). \quad (28)$$

Interchanging the roles of (ω, c) and (ω', c') , Assumption 1* also implies:

$$\rho_i(c|\omega') \geq \rho_i(c'|\omega') \iff \rho_{i'}(c|\omega) \geq \rho_{i'}(c'|\omega).$$

By the Reversal Fact,

$$\rho_i(c'|\omega') \geq \rho_i(c|\omega') \iff \rho_{i'}(c'|\omega) \geq \rho_{i'}(c|\omega). \quad (29)$$

Chaining (28) and (29) (whose middle terms agree) gives, for every $i' \in \mathcal{I}$,

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_{i'}(c'|\omega) \geq \rho_{i'}(c|\omega),$$

which is the case $\sigma = \omega$. Together with (27), this proves the lemma. $\square$

Proposition 3. *Assumptions 1*, 2, and 3 hold if and only if the following statement holds: For all $i, i' \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$ such that $\rho_i^*(\omega) = \rho_i(c|\omega)$ and $\rho_{i'}^*(\omega') = \rho_{i'}(c'|\omega')$,*

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_i^*(\omega') \geq \rho_i^*(\omega) \text{ and } \rho_{i'}^*(\omega') \geq \rho_{i'}^*(\omega). \quad (\text{P})$$

Proof. We begin by proving necessity, and then turn to sufficiency.

Necessity. We show that (P) implies each of the three assumptions in turn.

Step 1: (P) $\Rightarrow$ Assumption 2. Suppose $\rho_i^*(\omega) = \rho_i(c|\omega)$; we will show $\rho_{i'}^*(\omega) = \rho_{i'}(c|\omega)$ (the converse implication follows by exchanging the roles of i and i'). By our assumption on $\mathcal{C}$, there is a CQ $\tilde{c} \in \mathcal{C}$ with

$$\rho_{i'}(\tilde{c}|\omega) = \rho_{i'}^*(\omega). \quad (30)$$

Apply (P) with $\omega' = \omega$ and the roles of i and i' switched. That yields:

$$\rho_{i'}(c|\omega) \geq \rho_{i'}(\tilde{c}|\omega) \iff \rho_i^*(\omega) \geq \rho_{i'}^*(\omega) \text{ and } \rho_i^*(\omega) \geq \rho_i^*(\omega).$$

The right-hand side holds with “=” and fails with “>” and “<,” so $\rho_{i'}(c|\omega) = \rho_{i'}(\tilde{c}|\omega) = \rho_{i'}^*(\omega)$, as claimed.

Step 2: (P) $\Rightarrow$ Assumption 3. Fix $i \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$ satisfying Assumption 3’s premises: $\rho_i^*(\omega) = \rho_i(c|\omega)$ and $\rho_i^*(\omega') = \rho_i(c'|\omega')$. Applying (P) with both respondent roles equal to i yields:

$$\rho_i(c'|\omega) \geq \rho_i(c|\omega) \iff \rho_i^*(\omega') \geq \rho_i^*(\omega). \quad (31)$$

Applying (P) again with the roles of ω and ω' switched yields:

$$\rho_i(c'|\omega') \gtrless \rho_i(c|\omega') \iff \rho_i^*(\omega') \gtrless \rho_i^*(\omega). \quad (32)$$

Since the right-hand sides of (31) and (32) are identical, each of the three statements implies each of the other two. That in turn implies Assumption 3.

Step 3: (P) $\Rightarrow$ Assumption 1*. Fix $i, i' \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, and $c, c' \in \mathcal{C}$ satisfying Assumption 1*'s premises: $\rho_i^*(\omega) = \rho_i(c|\omega)$ and $\rho_i^*(\omega') = \rho_i(c'|\omega')$. By Step 1, Assumption 2 holds, so

$$\rho_{i'}^*(\omega) = \rho_{i'}(c|\omega) \quad \text{and} \quad \rho_{i'}^*(\omega') = \rho_{i'}(c'|\omega'). \quad (33)$$

(P) as stated is:

$$\rho_i(c'|\omega) \gtrless \rho_i(c|\omega) \iff \rho_i^*(\omega') \gtrless \rho_i^*(\omega) \quad \text{and} \quad \rho_{i'}^*(\omega') \gtrless \rho_{i'}^*(\omega).$$

Then the conclusion of Step 2 with i' instead of i says that each of the three statements $\rho_{i'}(c'|\omega) \gtrless \rho_{i'}(c|\omega)$, $\rho_{i'}(c'|\omega') \gtrless \rho_{i'}(c|\omega')$ and $\rho_{i'}^*(\omega') \gtrless \rho_{i'}^*(\omega)$ implies the other two. Thus,

$$\rho_{i'}(c'|\omega') \gtrless \rho_{i'}(c|\omega') \iff \rho_i^*(\omega') \gtrless \rho_i^*(\omega) \quad \text{and} \quad \rho_i^*(\omega') \gtrless \rho_i^*(\omega).$$

Using both of the above two statements,

$$\rho_i(c'|\omega) \gtrless \rho_i(c|\omega) \implies \rho_i^*(\omega') \gtrless \rho_i^*(\omega) \implies \rho_{i'}(c'|\omega') \gtrless \rho_{i'}(c|\omega').$$

But since only one of “>,” “=,” and “<” can hold for each of the three relations, and exactly one relation holds on each side,

$$\rho_i(c'|\omega) \gtrless \rho_i(c|\omega) \iff \rho_i^*(\omega') \gtrless \rho_i^*(\omega) \iff \rho_{i'}(c'|\omega') \gtrless \rho_{i'}(c|\omega').$$

Hence $\rho_i(c'|\omega) \gtrless \rho_i(c|\omega) \iff \rho_{i'}(c'|\omega') \gtrless \rho_{i'}(c|\omega')$, which is Assumption 1*.

Sufficiency. The proof of sufficiency is identical to that in Proposition 1, with only one difference: rather than using Assumption 1, the four CQ orderings (14)–(15) are obtained by applying Lemma 1 to equations (10) and (13) with i as the anchor respondent.

□

Appendix References

Bago d’Uva, Teresa, Maarten Lindeboom, Owen O’Donnell, and Eddy van Doorslaer, “Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity,” *Journal of Human Resources*, 2011, 46 (4), 875–906.

- Benjamin, Daniel J., Kristen Cooper, Ori Heffetz, Miles S. Kimball, and Jiannan Zhou**, “Adjusting for scale-use heterogeneity in self-reported well-being,” NBER Working Paper w31728, National Bureau of Economic Research 2023.
- , **Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones**, “What do you think would make you happier? What do you think you would choose?,” *American Economic Review*, 2012, 102 (5), 2083–2110.
- , —, —, and **Nichole Szembrot**, “Beyond happiness and satisfaction: Toward well-being indices based on stated preference,” *American Economic Review*, 2014, 104 (9), 2698–2735.
- Birnbaum, M. H.**, “Psychophysics,” in “Encyclopedia of Human Behavior,” Vol. 3, Academic Press, 1994, pp. 641–650.
- Bond, Timothy N. and Kevin Lang**, “The sad truth about happiness scales,” *Journal of Political Economy*, 2019, 127 (4), 1629–1640.
- Cronbach, Lee J.**, “Response sets and test validity,” *Educational and Psychological Measurement*, 1946, 6 (4), 475–494.
- Deaton, Angus**, “Comment on ‘Work disability, work, and justification bias in Europe and the U.S.’,” in D. A. Wise, ed., *Explorations in the Economics of Aging*, Chicago, IL: University of Chicago Press, 2011, pp. 312–314.
- Ferrer-i-Carbonell, Ada, Bernard M. S. van Praag, and Ioannis Theodossiou**, “Vignette equivalence and response consistency: The case of job satisfaction,” IZA Discussion Paper 6174, Institute for the Study of Labor (IZA) 2011.
- Greenleaf, Eric A.**, “Improving rating scale measures by detecting and correcting bias components in some response styles,” *Journal of Marketing Research*, 1992, 29 (2), 176–188.
- Grol-Prokopczyk, Hanna, Emese Verdes-Tennant, Mary McEniry, and Márton Ispány**, “Promises and pitfalls of anchoring vignettes in health survey research,” *Demography*, 2015, 52 (5), 1703–1728.
- He, Jia, Janine Buchholz, and Eckhard Klieme**, “Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures,” *Journal of Cross-Cultural Psychology*, 2017, 48 (3), 319–334.
- Huang, Zhiyong, Haoxian Wang, and Wenyuan Zheng**, “An extended hierarchical ordered probit model robust to heteroskedastic vignette perceptions with an application to functional limitation assessment,” *PLOS ONE*, 2021, 16 (3), e0248805.
- Kapteyn, Arie, James P. Smith, and Arthur van Soest**, “Vignettes and self-reports of work disability in the U.S. and the Netherlands,” *American Economic Review*, 2007, 97 (1), 461–473.
- , —, and —, “Life satisfaction,” in Ed Diener, John F. Helliwell, and Daniel Kahneman, eds., *International Differences in Well-Being*, Oxford, UK: Oxford University Press, 2010, pp. 70–104.
- , —, —, and **Hana Vonková**, “Anchoring vignettes and response consistency,” Working Paper WR-840, RAND Labor and Population 2011.
- King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon**, “Enhancing the validity and cross-cultural comparability of measurement in survey research,” *The American Political Science Review*, 2004, 98 (1), 191–207.
- Lin, Nick and Seonghui Lee**, “Anchoring vignettes as a diagnostic tool for cross-national (in)comparability of survey measures: The case of voters’ left-right self-placement,” *British Journal of Political Science*, 2024, 54 (2), 492–502.
- Liu, Shuo and Nick Netzer**, “Happy times: Measuring happiness using response times,” *American Economic Review*, 2023, 113 (12), 3289–3322.
- Montgomery, Mallory**, “Reversing the gender gap in happiness,” *Journal of Economic Behavior & Organization*, 2022, 196, 65–78.
- Parducci, Allen**, “Category judgment: A range-frequency model,” *Psychological Review*, 1965, 72 (6), 407–418.

- Rapkin, Bruce D. and Carolyn E. Schwartz**, “Advancing quality-of-life research by deepening our understanding of response shift: A unifying theory of appraisal,” *Quality of Life Research*, 2019, 28, 2623–2630.
- Rossi, Peter E., Zvi Gilula, and Greg M. Allenby**, “Overcoming scale usage heterogeneity: A Bayesian hierarchical approach,” *Journal of the American Statistical Association*, 2001, 96 (453), 20–31.
- Sprangers, Mirjam A. G. and Carolyn E. Schwartz**, “Integrating response shift into health-related quality of life research: A theoretical model,” *Social Science and Medicine*, 1999, 48 (11), 1507–1515.
- Van Vaerenbergh, Yves and Troy D. Thomas**, “Response styles in survey research: A literature review of antecedents, consequences, and remedies,” *International Journal of Public Opinion Research*, 2013, 25 (2), 195–217.
- Weijters, Bert, Hans Baumgartner, and Maggie Geuens**, “The calibrated sigma method: An efficient remedy for between-group differences in response category use on Likert scales,” *International Journal of Research in Marketing*, 2016, 33 (4), 944–960.
- , **Maggie Geuens, and Niels Schillewaert**, “The stability of individual response styles,” *Psychological Methods*, 2010, 15 (1), 96–110.