NBER WORKING PAPER SERIES

CORRECTING FOR HETEROGENEITY IN THE USE OF SUBJECTIVE SCALES

Daniel J. Benjamin
Kristen B. Cooper
Ori Heffetz
Miles S. Kimball
Jiannan Zhou

Working Paper 35925
http://www.nber.org/papers/w35925

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2026

This paper supersedes "Adjusting for Scale-Use Heterogeneity in Self-Reported Well-Being" (NBER Working Paper 31728, September 25, 2023) except it analyzes different data. We are grateful to Matt Adler, Angus Deaton, Marc Fleurbaey, Casper Kaiser, Arie Kapteyn, Laura Kubzansky, Richard Lucas, Terra McKinnish, Carolyn Schwartz, Itai Sher, Arthur Stone, Louis Tay, Peter Ubel, Bert Weijters, conference participants at the Bravo Center Workshop on Behavioral Economics at Brown University, the Dartmouth Workshop on Philosophy and Economics, the Hebrew University Federmann Rationality Center's 31st Annual Retreat, the Innsbruck Winter Summit, the Los Angeles Experiments conference, the Normative Economics and Economic Policy webinar, and seminar participants at UCLA Anderson, Baylor University, Cornell University, Chicago Booth, George Mason, Northwestern Kellogg, Oxford Wellbeing Research Centre, Paris School of Economics, Shandong University, Southampton Centre for Experimental Social Sciences, the University of California Santa Barbara, University of Colorado Boulder, University of Connecticut, University of Melbourne, University of Nottingham, University of Pittsburgh, and University of Warwick for helpful comments and discussion; and to Amir Adar, Tal Asif, Keren-Or Barashi Gortler, Yehonatan Caspi, Colby Chambers, Yonathan Cohen, Arshia Hashemi, Dominic Kassirra, Mattar Klein, Tushar Kundu, Dimitriy Leksanov, Rosie Li, Lev Maresca, Josef McCrum, Tuan Nguyen, Jeffrey Ohl, Shenhav Or, Ofri Piltz, Yonatan Rahimi, Becky Royer, Hannah Solheim, Aya Talmon, Pierre-Luc Vautrey, Shira Zadik, and Ruthie Zussman for outstanding research assistance. This work utilized the Summit supercomputer and the Alpine high-performance computing resource at the University of Colorado Boulder and data from the Understanding America Study (UAS), maintained by the Center for Economic and Social Research (CESR) at the University of Southern California (USC), but does not necessarily represent the official views of USC or UAS. We acknowledge support from NIH/NIA grants R01-AG065364 to the Hebrew University of Jerusalem and R01-AG051903 to the University of California Los Angeles for early stages of this research. The authors received IRB approval from the relevant institutions and have no material financial interests that relate to the research described in this paper. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Daniel J. Benjamin, Kristen B. Cooper, Ori Heffetz, Miles S. Kimball, and Jiannan Zhou. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Correcting for Heterogeneity in the Use of Subjective Scales
Daniel J. Benjamin, Kristen B. Cooper, Ori Heffetz, Miles S. Kimball, and Jiannan Zhou
NBER Working Paper No. 35925
October 2026
JEL No. C83, D63, D90, I14, I31

ABSTRACT

Analyses of survey data may be confounded if people use response scales differently. We formalize and generalize the theory behind using "calibration questions," such as vignettes, to correct for scale-use heterogeneity. Finding, in our (N 10,000) and existing data, that the mapping between respondents' scales is approximately linear, we develop a simple estimator for scale-use-corrected regressions. We validate it using subjective-scale ratings of height and weight and illustrate it with life satisfaction regressions, where correction leaves the main qualitative findings intact but can substantially change coefficient ratios, for example tripling the estimated cost of unemployment relative to income.

Daniel J. Benjamin
University of California, Los Angeles
Anderson School of Management
Economics Department
and NBER
daniel.benjamin@anderson.ucla.edu

Kristen B. Cooper
Gordon College
kristen.cooper@gordon.edu

Ori Heffetz
Cornell University
Johnson Graduate School of Management
and The Hebrew University of Jerusalem
and also NBER
oh33@cornell.edu

Miles S. Kimball
University of Colorado Boulder
and NBER
miles.kimball@colorado.edu

Jiannan Zhou
Shandong University
jiannan.zhou@icloud.com

An online appendix is available at http://www.nber.org/data-appendix/w35925
A preregistered research plan is available at
https://osf.io/qk5ta/?view_only=c2c05eb8d51d4c088c363db7a26d2f15

Across economics and other fields, researchers use survey data that are reported on a subjective scale, such as a slider or a 5-point Likert scale. Such data include customer and worker satisfaction, psychological measures, public opinion and political attitudes, self-assessments of pain, disability, and mental health, and—our illustrative example throughout this paper—life satisfaction. Most of this research implicitly assumes that all respondents use the survey response scale in the same way. If instead there is scale-use heterogeneity—if, for example, on a 0–100 scale for life satisfaction, one person's 70 corresponds to another person's 80—it can confound the conclusions researchers draw (as highlighted by, e.g., King et al., 2004; Kapteyn et al., 2007; Oswald, 2008). Economists analyzing panel data often include individual fixed effects, which remove stable mean differences in scale use but not changes in scale use over time or differences in how much of the scale individuals use (i.e., heterogeneity in scale-use variance). Analyses of cross-sectional data in economics typically make no corrections for scale use.

In this paper, we propose a straightforward econometric method to estimate and correct for scale-use heterogeneity in regressions with a subjectively measured dependent variable. In addition, we introduce a theoretical framework that makes precise the assumptions underlying scale-use correction. In turn, the assumptions clarify which survey-response "biases" would be problematic for scale-use correction and which would not. Our framework generalizes and unifies existing approaches, as we discuss further below.

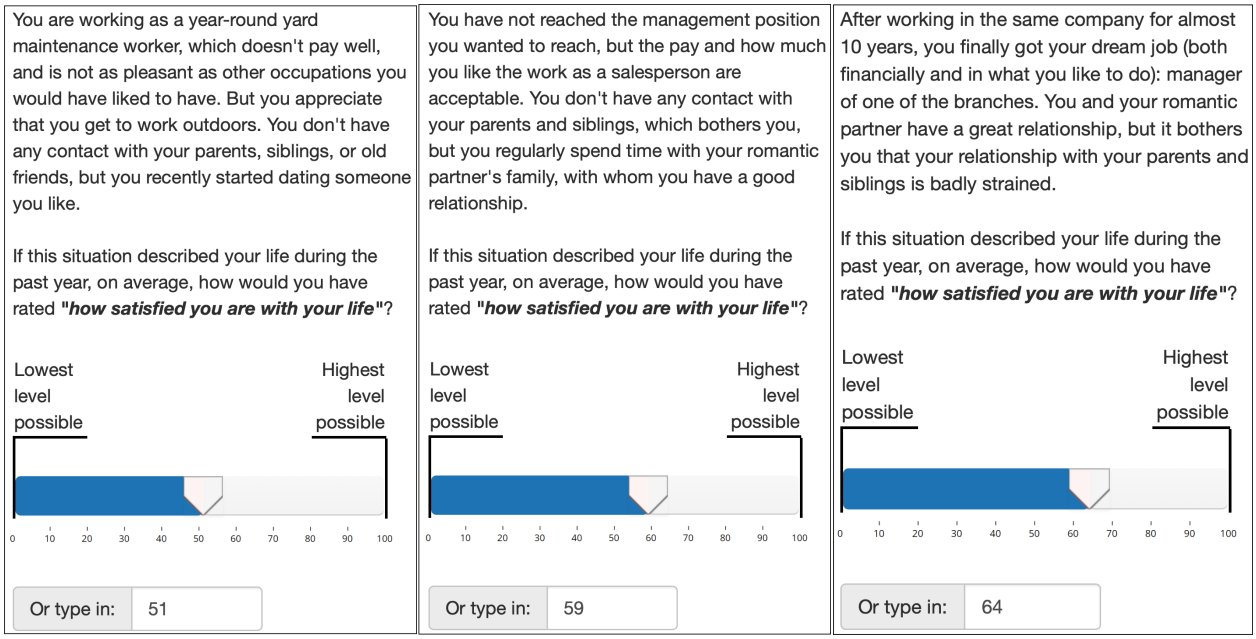
To anchor our methods and theory in a concrete empirical context, we first describe, in Section I, our survey design and sample: ∼10,000 respondents from the Understanding America Study, a demographically diverse web panel. We asked a variety of self-report questions—henceforth *self-ratings* (*SRs*)—including life satisfaction, on a 0–100 response scale. We identify scale use with what we call *calibration questions* (*CQs*): questions asked on the same response scale as the SR but designed so that respondents rate a fixed object. We study two *kinds*: *vignette CQs*, or simply *vignettes*, describe a hypothetical person whom the respondent rates; *visual CQs* are pictures the respondent rates. Panel (a) of Figure 1 shows a trio of vignettes, asked on one screen and designed to elicit a range of responses; Panel (b) shows a trio of visual CQs.

Each kind has an advantage. Vignettes can ask about the same dimension as the SR (life satisfaction vignettes for a life satisfaction SR) and can therefore measure the most relevant scale use. But Deaton (2011) and others have raised concerns about biases in how respondents rate vignettes: for example, respondents may differ in their empathy for the hypothetical person or may fill in unspecified details in ways correlated with their own situation. Visual CQs are much less subject to such biases.

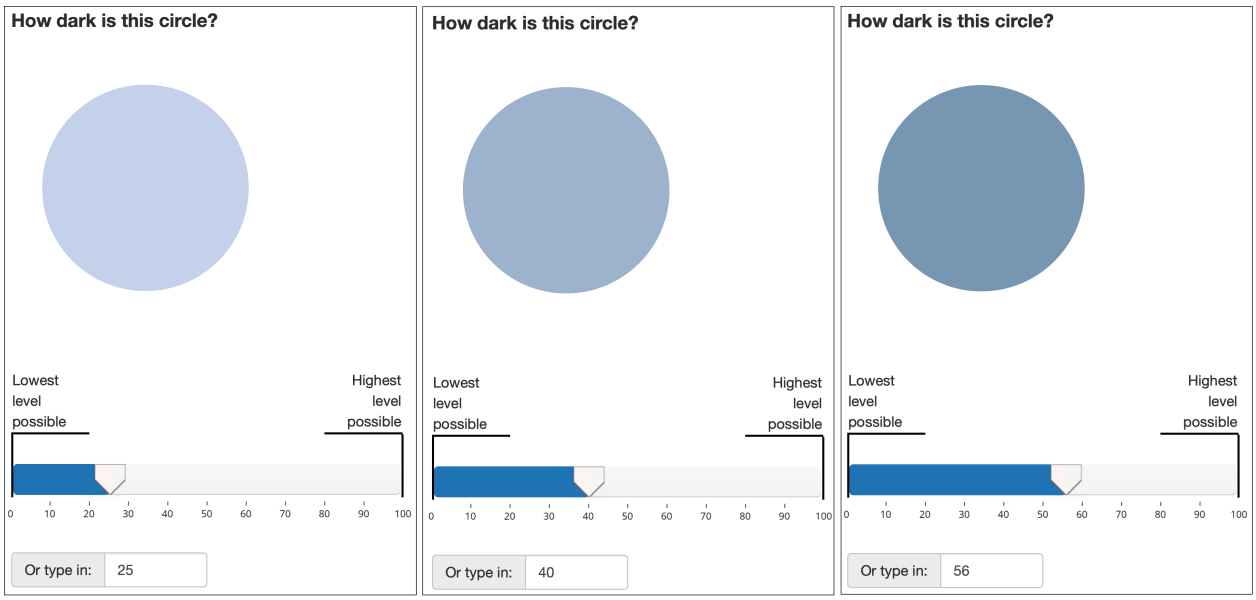
In Section II, we present two types of descriptive evidence. The first is evidence that respondents differ in both the mean and the variance of their scale use—whether they tend to report high or low numbers and how much of the scale they use—and that these differences carry over to their SRs. A

Figure 1: Examples of Calibration Questions (CQs)

(a) Vignette CQ Trio



(b) Visual CQ Trio



respondent's life satisfaction SR is positively associated with their mean rating of life satisfaction vignettes; changes over time in the two move together; the association with mean ratings of visual CQs is positive but weaker; and the standard deviation of a respondent's SRs across several questions is associated with the standard deviation of their CQ ratings. For height and weight, where we have self-reports both on a subjective scale (an SR) and an objective scale, CQ ratings predict the SRs

even after controlling for the objective-scale report.

The second type of evidence concerns how one group of respondents' CQ ratings relate to another group's: across many sample splits, one group's ratings are well approximated by a linear transformation of the other's. We find this relationship in our data and in other vignette datasets. At the individual level, after correcting for multiple hypothesis testing, we cannot reject linearity, and we estimate that any departures from it are quantitatively small. Partly motivated by this evidence, our econometric model assumes linear "CQ translation functions" that map the sample-mean CQ ratings to each respondent's ratings.

Section III introduces our econometric model and derives the bias from failing to correct for scale-use heterogeneity. We assume that respondents' SRs would be interpersonally comparable if translated to a common scale, and we use each respondent's linear CQ translation function to do the translating, so that their scale use is characterized by two parameters, which we call the intercept and stretcher. In a naïve (i.e., uncorrected for scale use) regression of the SR on a set of regressors, the coefficients are biased by heterogeneity in both parameters. An implication is that corrections that address only the intercept, such as fixed effects or (common in psychology) reverse-coding half the questions, do not in general eliminate the bias.

Section IV introduces our method-of-moments (MM) estimator for scale-use-corrected regressions. Each respondent's intercept and stretcher can be estimated from the CQ data by OLS, but when the number of CQs is small—the empirically realistic case—using these noisy estimates to correct each respondent's SR individually performs poorly. The MM estimator instead corrects the regression coefficients. It constructs, for each respondent, a "scale-use-only benchmark": the rating they would give a CQ located at the sample-mean SR. Because that level is the same for every respondent, the benchmark varies across respondents only because of scale use. Regressing the benchmark on the regressors estimates how scale-use heterogeneity biases the naïve coefficients, and subtracting that estimate from the naïve coefficients yields the corrected ones. Under two moment conditions that appear to hold well in our data, the MM estimator is consistent as the number of respondents grows, even with few CQs. In Web Appendix F, we derive a more complex family of semiparametric estimators under different, arguably weaker, assumptions; one is for regressions, and the others estimate means, variances, and covariances of SRs for non-regression applications.

In Section V, we apply our MM estimator to a life satisfaction regression. Correction leaves the most robust findings in the literature intact, including the associations with income, unemployment, and marriage. But it changes the magnitude, significance, and even sign of other coefficients, and it changes coefficient ratios substantially. For example, in the common exercise of estimating the welfare cost of unemployment from the ratio of the unemployment and income coefficients (e.g., Di Tella et al., 2001), correction *triples* the estimated cost, from a ratio of 1.6 to 5.0 (whereas correcting only for the intercept barely moves it, to 1.7). The impact of correction is larger with life satisfaction

vignettes than with visual CQs. We also find that the MM estimator, whose linear model is most natural for a continuous response scale, gives very similar results when life satisfaction is measured on a Likert scale with just a few response categories (Web Appendix I).

In Section VI, we validate our estimator using self-reports of height and weight in objective scales. Correcting the subjective-scale height and weight SRs with same-dimension CQs—CQs that involve rating height and weight—raises their measurement-error-corrected correlations with the objective-scale reports, from 0.56 to 0.60 for height and from 0.32 to 0.51 for weight under our baseline assumptions. Correction with CQs in other dimensions generally helps much less.

In Section VII we develop the theoretical framework underlying our approach to scale-use heterogeneity. SRs are ratings of a respondent's underlying "situation," and CQ responses depend not only on the CQ but may also depend on a respondent's situation. We lay out three assumptions. First, respondents share a common ranking of the CQs. This ensures that every respondent's rating can be translated to a common scale. Second, conditional on a situation, respondents agree on which CQ they would rate equally to that situation. Third, higher situations would map (by equal rating) to higher CQs—which would hold as long as SRs are more responsive to a respondent's situation than their CQ responses are. We discuss the plausibility of these assumptions for different kinds of CQs and how, in principle, they could be tested; and we report some empirical tests in our data. We show that these three assumptions are sufficient—and, once the first is suitably weakened, also necessary—for CQ responses to identify the ranking of respondents' common-scale SRs. Our econometric model then follows from our theoretical framework once we assume that CQ translation functions are linear and add response errors.

In Section VIII, drawing on our theory and evidence, we offer practical recommendations for analyzing existing data and for designing surveys. We recommend same-dimension CQs when they are available and their assumptions are plausible; when they are not, CQs in a closely related dimension are the best alternative, and cross-dimension correction can serve as a robustness check. For survey design, we recommend CQs that respondents rank the same way and rate far apart, without top- or bottom-coding; three CQs rather than the minimum of two, to permit specification tests; and, when asking CQs of the full sample is infeasible, asking them of a subsample.

We conclude in Section IX, discussing how scale-use heterogeneity can matter in studies with random or quasi-random designs and outlining some directions for future research.

This paper is related to the two substantial existing literatures on scale-use correction: anchoring vignettes and response styles. Although scale-use correction using anchoring vignettes (King et al., 2004)—which are the same as what we call same-dimension vignette CQs—has become fairly common in a number of disciplines outside of economics (e.g., Lin and Lee, 2024), the assumptions underlying the approach have been viewed with skepticism (e.g., He et al., 2017). A separate

literature, which has made even less contact with economics,[1] relies on what we call other-dimension CQs to study people's "response styles" (for reviews, see, e.g., Van Vaerenbergh and Thomas, 2013; Weijters et al., 2016). Our theoretical framework nests both approaches as special cases and clarifies the assumptions under which each is reasonable; by generalizing the assumptions of the anchoring-vignette approach, it also shows that the approach is robust to many reasonable concerns. We provide empirical evidence bearing on these assumptions. Relative to typical estimation methods in these literatures, our MM estimator has several advantages: it is simple and transparent; its only parametric assumption, linearity of the CQ translation function, is testable, and it requires no distributional assumptions; and it is robust to heteroskedasticity, including the heteroskedasticity caused by scale-use heterogeneity itself.

Our Web Appendix further discusses our paper's relations to these two literatures and provides details on survey instruments, robustness and supplementary analyses, simulations, and proofs. Our GitHub repository, https://github.com/jiannanzhou/scaleuse, provides R and Stata packages implementing our MM estimator, along with instructions for using them with AI coding assistants.

I Survey Design and Sample

This paper uses two Understanding America Study (UAS) survey modules that we designed: UAS 571 ("Life Aspect Survey") and UAS 719 ("Follow-Up Life Aspect Survey"), which we call Survey 1 and Survey 2. Survey 2 was administered to a subset of Survey 1 respondents 12 to 22 months later.

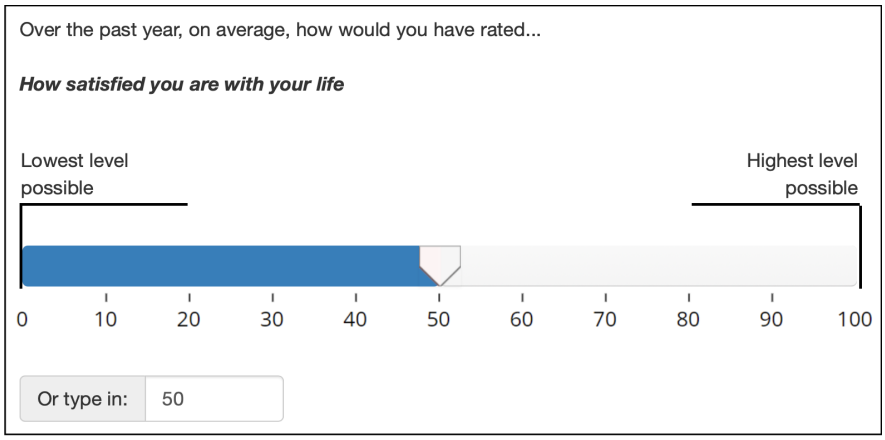
I.A Question Types and Response Scale

The two key types of questions in Surveys 1 and 2 are calibration questions (CQs), illustrated in Figure 1, and self-ratings (SRs), in which respondents rate the level of some dimension of their own life, on average over the past year. We specify the past year to reduce heterogeneity in how respondents interpret the question (Benjamin et al., 2023). Figure 2 shows the Life Satisfaction SR, which appears in both surveys.

CQs and SRs share a response scale: the integers from 0 (labeled "Lowest level possible") to 100 ("Highest level possible"). Respondents either move a slider or type a number; the two automatically adjust to match each other. To discourage inattentive responses, the slider starts at 50 and must be moved, or a number typed, before a response is recorded. The quasi-continuous scale makes our

[1] In the economics literature, the most closely related work we are aware of is Márquez-Padilla and Álvarez (2018), who find that countries' pass/fail grading thresholds in school are related to countries' mean life satisfaction responses, presumably because both reflect cultural norms of scale use; and Decancq and Seré (2025), who find that heterogeneity in how people from different countries rate hotel rooms is related to cross-country differences in national means and variances of happiness ratings.

Figure 2: Example of a Self-Rating (SR) Question



conclusions less dependent on the untestable latent-variable assumptions required when there are only a few response categories (Bond and Lang, 2019).[2] We chose the endpoint labels to suit both SRs and CQs and to sound extreme in order to reduce top- and bottom-coding.

I.B Calibration Question (CQ) Design

CQs always appear in trios: three similar CQs about the same dimension, using the same type of stimulus, and designed to sit at a low, medium, and high level of that dimension. We refer to them as the Low, Medium, and High CQs. Presenting the three together on one screen makes them cognitively simpler than three unrelated CQs, enabling respondents to answer more CQs per minute of survey time.

For visual CQs, we avoided dimensions with a natural objective scale, such as the percentage of a circle that is shaded, which respondents might report on that scale. For vignettes, we began with a set of subdimensions relevant to the rated dimension (e.g., career, relationships, and family for life satisfaction) and wrote the three vignettes so that the subdimensions change monotonically across them. Monotonicity is not required by our econometric approach, but we expected it to make the rating task easier.

To minimize top- and bottom-coding, we designed and pilot-tested the CQs to avoid extreme levels; in particular, every High vignette includes one negative detail and every Low vignette one

[2]A longstanding tradition in psychology maintains that 5-point or 7-point scales are best for surveys because respondents cannot reliably discriminate more finely than that. Evidence suggests that while response scales with more than 5 options are similar to each other along several metrics of psychometric validity, respondents rate them as better in terms of allowing them to express their feelings but worse in ease and quickness of use (e.g., Preston and Colman, 2000). We advocate continuous scales precisely because they maximize how fine-grained respondents can be in expressing their feelings and therefore, in principle, eliminate the problem of inferring a latent variable from discrete response categories. In practice, however, many respondents likely round their answers. Future research should account for such rounding, as in Giustinelli et al. (2022).

positive detail. Web Appendix C shows that top- and bottom-coding were rare for all CQs but one, the High Height CQ.

I.C Survey 1 Description

Both Surveys 1 and 2 (described next) begin with brief instructions and end with exit questions. In both, all randomizations are at the respondent level, independent and uniform. For complete text and images, see Web Appendix B. The flow for Survey 1 is as follows, with focal questions in bold.

(1) **4 trios of visual CQs.** We list here the trio abbreviation and question prompt. ***Darkness***: "How dark is this circle?"; ***Shape***: "How complex is this shape?"; ***Confidence***: "How confident does this person look?" (a man's face); ***Pain***: "How would you rate this person's level of pain?" (a cartoon face). We randomize the order of the trios and the order of CQs within trio.

(2) **26 SRs.** We list here the dimension name and the text rated, for the SRs studied in the paper only (Web Appendix Table B.1 lists all 26). ***Life Satisfaction***: How satisfied you are with your life; ***Happiness***: How happy you feel; ***Worthwhileness***: The extent to which you feel the things you do in your life are worthwhile; ***No Anxiety***: You not feeling anxious; ***Anxiety***: You feeling anxious; ***Breathe***: Your ability to breathe in and out easily; ***Remember***: Your ability to remember things. We randomize the order of SRs and whether a given respondent rates Anxiety or No Anxiety.

(3) Two stated-preference questions, not analyzed in this paper.

(4) **Vignette CQs in 7 dimensions, 5 trios per respondent.** The dimensions are ***Life Satisfaction***, ***Happiness***, ***Worthwhileness***, and ***Anxiety***, which together cover the four dimensions measured by the UK's Office for National Statistics; ***No Anxiety***, to explore more-is-good versus more-is-bad orientation; and ***Breathe*** and ***Remember***, intended to be relatively objective and narrow. Every respondent sees Life Satisfaction, Breathe, and Remember; is randomized to Happiness or Worthwhileness (to save survey time); and sees Anxiety or No Anxiety according to which SR they were assigned. The Medium CQ in the Remember trio is an attention check. We also randomize the order of trios, the order of CQs within trio, and which of four vignette-subject framings the respondent sees.[3] Because we cannot reject that the four framings yield equivalent regression results (Web Appendix Table L.8), we pool across them.

I.D Survey 2 Description

The flow for Survey 2 is as follows, with focal questions in bold.

(1) **3 SRs.** ***Life Satisfaction***: same as in Survey 1; ***Height***: "Relative to adults of your gender, how would you describe **your height** on a scale from 0 to 100, where 0 is the lowest level possible

[3]We refer to the four framings as ***Self*** (as in Figure 1a), ***Nonself/Male*** and ***Nonself/Female***, third-person with a gendered name, and ***You (stranger)***. See Web Appendix Table B.3 for survey instructions and prompts by framing.

and 100 is the highest level possible?"; ***Weight***: "Relative to adults with the **same height** and gender as you, how would you describe **your weight**, on a scale from 0 to 100, where 0 is the lowest level possible and 100 is the highest level possible?"

(2) **2 trios of Life Satisfaction vignette CQs.** One trio is identical to Survey 1's; the other, ***Life Satisfaction 2***, is new. We randomize their order and the order of CQs within trio. Each respondent sees the same vignette-subject framing as in Survey 1.

(3) **Height and Weight CQ trios.** The Height CQs show an AI-generated man standing beneath a basketball hoop and ask, "How would you describe the height of this man, relative to other male adults?" The Weight CQs show an AI-generated man in side profile, with the three men drawn at the same height, and ask, "How would you describe the weight of this man, relative to other male adults of the same height?"

I.E Survey Sample

Survey 1 was fielded between August 30, 2023, and May 9, 2024, with 12,549 UAS panelists invited and 10,232 responses (81.5%). Our main analysis sample is the 10,100 of these with complete responses to the SRs and CQs and complete demographic covariates. Survey 2 was fielded between May 8 and June 29, 2025, to a random sample of 3,077 Survey 1 respondents. There were 2,537 responses (82.5%), including 2,503 with complete demographic covariates. In Surveys 1 and 2, respectively, respondents were compensated $8 and $2, and the average response time was 15 and 6 minutes. See Web Appendix Table L.1 for respondent demographics. Throughout the paper we report unweighted results. Objective-scale self-reports of height (in feet and inches) and weight (in pounds) are each respondent's most recent prior report to the UAS's Health and Retirement Study modules (UAS 20, 95, 185, 396, and 530), which span May 2015 to January 2024.[4]

II Descriptive Evidence: Scale-Use Heterogeneity and Linear CQ Translation Functions

This section presents descriptive evidence on two facts that motivate our approach: scale use varies across respondents and predicts their SRs (Section II.A), and CQ translation functions are approximately linear (Section II.B).

[4]For each measure we take the respondent's most recent non-missing report. Because height is asked far less often than weight, height values are typically older: UAS 396 supplies 77% of the weight values but only 37% of the height values in our validation sample, with 46% of the height values coming from UAS 20 (fielded 2015–2019). For fifteen respondents the most recent weight report comes from the UAS COVID-19 tracking survey rather than these modules.

II.A Scale-Use Heterogeneity

Figure 3 provides suggestive evidence that heterogeneity in CQ ratings is associated with heterogeneity in SR ratings. Panel (a) plots, for each Survey 1 respondent, their Life Satisfaction SR against the mean of their three Life Satisfaction vignette CQs. The correlation, 0.20 (SE = 0.02), is positive: respondents who rate the vignettes higher also rate themselves higher. The magnitude of this correlation, as well as other correlations below, is attenuated by both true variation in life satisfaction and measurement error in the SRs and CQs. Panel (b) shows the analogous scatterplot for *changes*, using the smaller sample who also answered Survey 2 more than a year later. The correlation, 0.13 (SE = 0.03), is smaller and noisier but positive. This finding suggests that scale use itself changes over time, so respondent fixed effects would not fully eliminate confounding from scale-use heterogeneity (a point we develop in Section III.C).

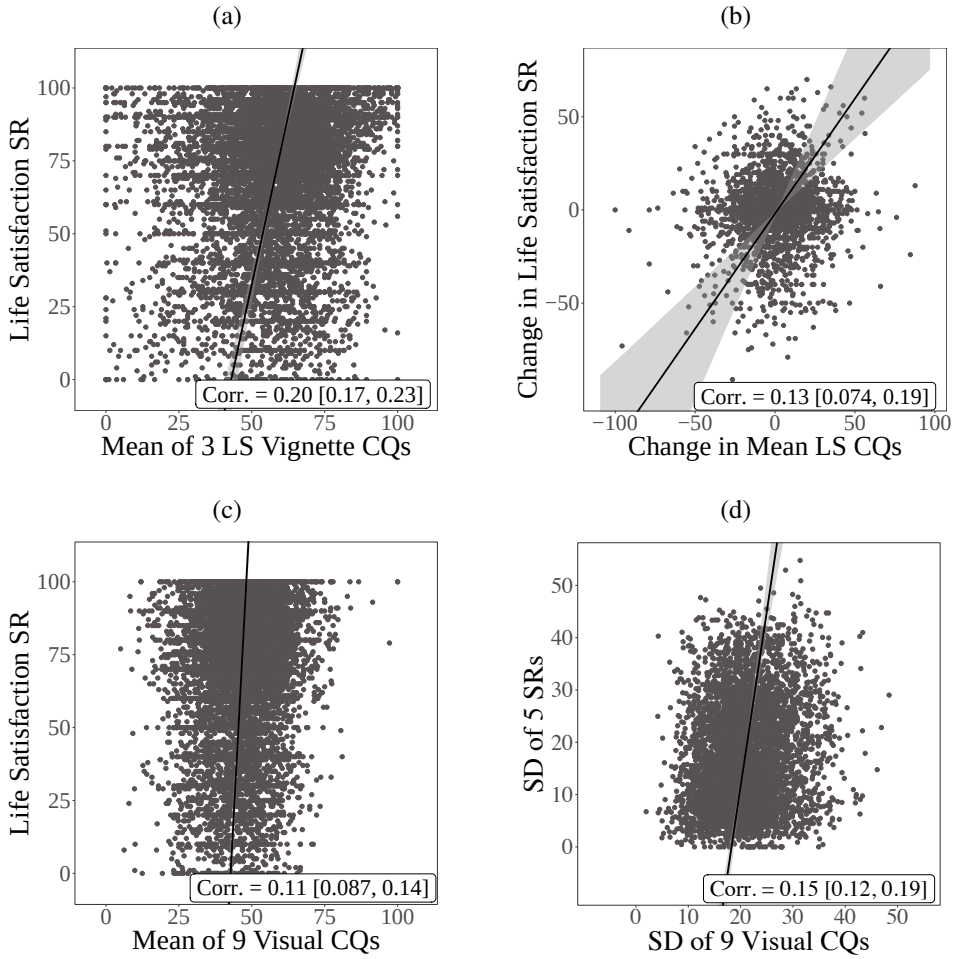
Panel (c) is analogous to Panel (a), but the x-axis is the mean rating of nine visual CQs (the Shape, Darkness, and Confidence trios) rather than of the Life Satisfaction vignettes. The correlation, 0.11 (SE = 0.02), is positive but smaller, consistent with scale use being partly, but not fully, shared across the two kinds of dimension. Panel (d) plots the standard deviation of each respondent's SRs across five dimensions (Life Satisfaction, Happiness, No Anxiety, Breathe, and Remember) against the standard deviation of their ratings of the same nine visual CQs. The correlation, 0.15 (SE = 0.02), is positive, suggesting that respondents also differ in how much of the scale they use, and that this difference carries over to their SRs.

Although suggestive that scale use confounds comparisons of SRs across respondents, none of the evidence in Figure 3 is definitive because heterogeneity in scale use could be correlated with true variation in the underlying dimension. For stronger evidence, we turn to the two dimensions for which we have self-reports both on our 0–100 scale (the SRs) and on objective scales: height and weight. Table 1 shows that scale use predicts the SRs even after controlling for the objective-scale reports. Each column regresses one SR on the respondent's ratings of the three same-dimension CQs, controlling for gender and cubics in log height (inches) and log weight (pounds). The top panel shows that five of the six CQ coefficients are positive and statistically distinguishable from zero. The exception, the High Height CQ, has a small negative coefficient; it is the most top-coded of the six CQs (see Web Appendix C), which compresses its informative variation. The bottom panel shows that adding the CQs raises R^2 from 0.288 to 0.332 for height and from 0.067 to 0.109 for weight.

II.B CQ Translation Functions

The purpose of CQs is to measure scale use: by asking respondents who use the scale differently to rate the same CQs, we learn how to "translate" one respondent's scale into another's. A respondent's

Figure 3: Evidence of Scale-Use Heterogeneity in Self-Ratings



Notes: LS = Life Satisfaction. Each point = one respondent. Sample (by panel): N = 10,100 (a, c) from Survey 1; 2,503 (b; respondents who answered follow-up Survey 2 over one year later); 5,062 (d). X-axis (Panels c, d): mean and standard deviation of nine visual CQs from three trios (Shape, Darkness, and Confidence). Y-axis (Panel d): standard deviation of respondent's SRs across five dimensions (Life Satisfaction, Happiness, No Anxiety, Breathe, and Remember). Black lines: total least squares (TLS) regressions; 95% confidence regions in gray (hardly visible in Panels a, c).

CQ translation function, which maps a common-scale rating (which we take to be the mean rating across respondents) to the respondent's own rating, is central to our econometric and theoretical models in later sections. Here we provide descriptive evidence on it.

Because translation functions estimated at the respondent level are noisy, Figure 4 instead compares average CQ ratings across groups of respondents. Each panel compares one pair of groups defined by a demographic or socioeconomic variable (e.g., below- versus above-median age). Each of its 32 points plots the two groups' mean ratings of one CQ against each other, color-coded by

Table 1: Scale Use Predicts Self-Ratings

	Height SR (0–100)	Weight SR (0–100)
Low CQ	0.16 (0.02)	0.18 (0.02)
Medium CQ	0.15 (0.03)	0.09 (0.04)
High CQ	-0.04 (0.02)	0.07 (0.02)
Controls	Yes	Yes
Obs.	2,471	2,471
R^2	0.332	0.109
R^2 (Controls only)	0.288	0.067

Notes: Each column regresses the indicated SR on the respondent's ratings of the three same-dimension CQs (Low, Medium, and High), controlling for gender and cubics in log height (inches) and log weight (pounds). Heteroskedasticity-robust standard errors in parentheses.

dimension: the 20 vignette CQs (seven trios, less the Remember attention check) and the 12 visual CQs (four trios). The relationship traced by the points is the average translation function from one group's scale to the other's.

We highlight four observations. First, within every trio, the two groups rank the three CQs identically (consistent with an assumption of our theoretical framework, Common Ranking of CQs; Section VII.C). Second, many points lie off the identity line, and the figure's tests reject the identity line in most panels, suggesting scale-use differences across groups. Third, the cross-group translation functions differ across dimensions, a point we return to in Section VIII ("Same- vs. Cross-Dimension Considerations"). Fourth, the cross-group translation functions are roughly linear. In Web Appendix E, we test linearity at the individual level using the 2,537 Survey 2 respondents, who each answered six Life Satisfaction vignettes. We reject the null at the 5% level for 4.4% of respondents, about the rate expected under the null, and no rejection survives false-discovery-rate correction. A complementary maximum-likelihood analysis estimates the dispersion of the quadratic terms and implies that any departures from linearity are quantitatively small.

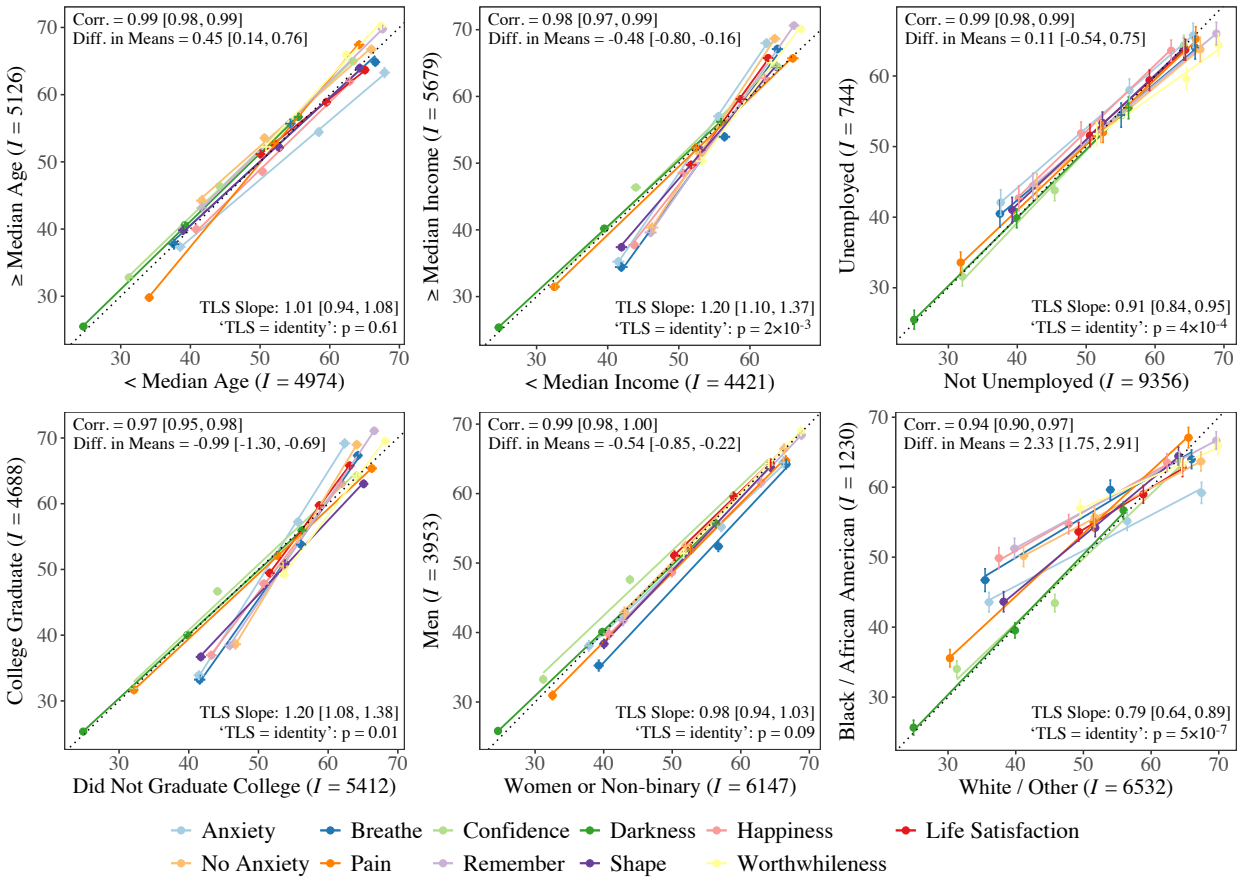
These observations also hold in other vignette datasets that we analyze in Web Appendix E, including those of King et al. (2004), several of which use Likert scales with few response categories. Linearity thus appears to be a reasonable approximation even for such scales, a conclusion further supported by Likert-scale data we collected on a convenience sample (Web Appendix Figure E.2).

III Econometric Model and Bias from Scale-Use Heterogeneity

III.A The Problem of Scale-Use Heterogeneity

Suppose a researcher wishes to run a regression in which the dependent variable, a response to a self-rating (SR) question, is measured on a subjective scale but is *not* subject to scale-use

Figure 4: Estimated Average CQ Translation Functions Across Demographic Groups



Notes: Each point = mean for a single CQ for both groups, with 95% CIs (capped bars); 32 CQs in total. Sample: N=10,100 respondents. Colored solid lines: TLS regression lines fit separately within each CQ trio (one line per trio). Dotted line: 45°. Corr.: Pearson correlation between the x- and y-axis means across CQs. Diff. in Means: y- minus x-axis mean CQ ratings. TLS slope: a single aggregate TLS line fit through all CQ points (not the per-trio lines). 'TLS = identity' p-value: F-test of the joint hypothesis that this line's (intercept, slope) = (0,1). Square brackets: 95% CIs.

heterogeneity: it is measured on a common scale across respondents. Our running example will be a regression of life satisfaction on a vector of sociodemographics, denoted $\mathbf{x}_i$. We use a superscript $\star$ to indicate an SR rating (as opposed to a CQ rating). Denoting the SR on the common scale by $w_i^\star$, the desired regression is

$$w_i^\star = \gamma_0 + \mathbf{x}_i \boldsymbol{\gamma}_w + \upsilon_i^\star, \tag{1}$$

where $\boldsymbol{\gamma}_w$ is the coefficient vector of interest. Unfortunately, this regression is infeasible because $w_i^\star$ is unobserved. Instead, researchers typically run the feasible but naïve regression of an SR that *is* subject to scale-use heterogeneity, denoted $r_i^\star$, on $\mathbf{x}_i$, yielding the coefficient vector $\boldsymbol{\gamma}_r$. In Section III.C, we derive the bias of this naïve regression, and in Section IV, we describe how researchers can use CQs to recover a consistent estimate of $\boldsymbol{\gamma}_w$. We turn first, however, to defining the "CQ translation function" because it is an ingredient in the subsequent analyses.

III.B The CQ Translation Function

As Section VII justifies formally, we presuppose a common scale for both CQs and SRs. We denote the common-scale rating of CQ c by w_c. Respondent i's *CQ translation function* maps any common-scale CQ rating w_c to i's own rating r_{ic}. Section VII.E shows that, under our assumptions and absent response error, scale-use correction is nonparametrically identified whatever the shape of this function. For reasons explained below, however, our econometric methods assume it is linear:[5]

$$r_{ic} = a_i + b_i w_c + \eta_{ic}, \tag{2}$$

where the *intercept* a_i and *stretcher* b_i parameterize scale-use heterogeneity, and η_{ic} is a response error for CQs (not part of the translation function). As we will discuss in Section VII, any respondent's scale could be chosen as the common scale. For econometric analysis, a convenient choice is the population mean rating: $w_c := \mathbb{E}(r_{ic})$, which implies $\mathbb{E}(a_i) = 0$ and $\mathbb{E}(b_i) = 1$. We assume $\mathbb{E}(\eta_{ic}|w_c, w_i^\star, \mathbf{x}_i, a_i, b_i) = 0$ but allow for arbitrary heteroskedasticity, including dependence of the variance of η_{ic} on $(w_c, w_i^\star, \mathbf{x}_i, a_i, b_i)$. Indeed, we generally expect respondents with larger stretchers to have larger response-error variance, since stretching the scale stretches the errors along with it.

We assume a linear CQ translation function for three reasons. First, its simplicity makes it a natural starting point. Second, as discussed in Section II.B, it is a good approximation empirically. Third, a linear CQ translation function is attractive theoretically because, by ensuring that r_{ic} is measured on an interval scale, it implies that ratios of scale-use-corrected regression coefficients are invariant to whose scale is chosen as the common scale.[6]

III.C The Bias from Scale-Use Heterogeneity

To calculate the bias from the naïve regression of $r_i^\star$ on $\mathbf{x}_i$, we need to know the relationship between each respondent's reported SR $r_i^\star$ and their unobserved common-scale SR $w_i^\star$. In our theoretical analysis in Section VII, we show that, under our assumptions, we can use the CQ common scale in place of the (unidentified) SR common scale without disturbing the ranking of respondents' situations. Each respondent's SR is then located on that scale by their own CQ translation function:

$$r_i^\star = a_i + b_i w_i^\star + \eta_i^\star, \tag{3}$$

[5]Throughout this paper, "linear" is used in the sense standard in econometrics, allowing an intercept; a mathematician would say "affine."

[6]In light of our analysis in Section II.B, we anticipate that a linear CQ translation function will generally be a good approximation empirically. Nonetheless, for illustrative purposes, in Web Appendix F.3 we show how, with at least four CQs and under some additional assumptions, our semiparametric estimator can be extended to the case of a quadratic CQ translation function. Once we move away from a linear CQ translation function, however, ratios of coefficients in $\boldsymbol{\gamma}_w$ are no longer invariant to choice of scale; $w_c := \mathbb{E}(r_{ic})$ becomes a substantive choice, albeit one that dovetails with viewing regression as a linear approximation around the sample mean.

where the intercept a_i and stretcher b_i are the same as in equation (2), and we assume that the response error $\eta_i^\star$ satisfies $\mathbb{E}(\eta_i^\star | w_i^\star, \mathbf{x}_i, a_i, b_i) = 0$ but can have arbitrary heteroskedasticity.

The bias from running the naïve regression is then:

$$\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_w = \text{Var}\left(\mathbf{x}_i\right)^{-1} \left[\text{Cov}\left(\mathbf{x}_i, r_i^\star\right) - \text{Cov}\left(\mathbf{x}_i, w_i^\star\right)\right],$$

where, by substituting $r_i^\star = a_i + b_i w_i^\star + \eta_i^\star$, each element of the vector in square brackets is given by:

$$\begin{aligned}\text{Cov}\left(x_i, r_i^\star\right) - \text{Cov}\left(x_i, w_i^\star\right) = \text{Cov}&\left(x_i, a_i + b_i \mathbb{E}\left(w_i^\star\right)\right) \\ &+ \mathbb{E}\left[\left(b_i - \mathbb{E}\left(b_i\right)\right)\left(x_i - \mathbb{E}\left(x_i\right)\right)\left(w_i^\star - \mathbb{E}\left(w_i^\star\right)\right)\right]. \qquad (4)\end{aligned}$$

The first term on the right-hand side captures differences in scale use across values of x_i. The term $a_i + b_i \mathbb{E}(w_i^\star)$ is i's CQ translation function evaluated at the mean common-scale level of the SR; it summarizes i's scale use, and its sample analog is the "benchmark" at the heart of our estimator in Section IV. As a concrete example, if x_i is a dummy for being male, and if males have different a_i's or b_i's than females, then the first term generates a bias. The second term is the (unstandardized) coskewness between the stretcher, the x_i variable, and the common-scale level of the SR. Intuitively, all respondents are equally weighted in $\text{Cov}\left(x_i, w_i^\star\right)$, but respondents with larger stretchers get more weight in $\text{Cov}\left(x_i, r_i^\star\right)$. Because the bias for a given regressor x_i is premultiplied by $\text{Var}\left(\mathbf{x}_i\right)^{-1}$, it affects not only the coefficient on x_i but also any other regressors correlated with it.

Equation (4) makes clear why two common practices only partially address the bias from scale-use heterogeneity. Individual fixed effects in panel data remove the intercept but not the stretcher, and even for the intercept they remove only its stable component. That component is far from the whole: in Web Appendix D.3, using CQ ratings from our two surveys, we estimate that only 59.0% (SE = 4.5%) of the variance in the intercept is persistent (for the stretcher, 84.8%, SE = 3.2%).[7]

Reverse-coding some items in a multi-question battery, common in psychology, likewise addresses only the intercept. More subtly, it corrects only for the component of the intercept that shifts all of a respondent's responses up or down the scale, not for a component that differs between positively and negatively worded items (e.g., Anxiety versus No Anxiety).

[7]The consequences can be severe for estimating the difference in trends between two groups in panel data with individual fixed effects, even in the best case for fixed effects, in which both scale-use parameters are constant over time. In that case, differencing removes a_i but not b_i, so each group's measured trend is its true trend in $w_i^\star$ multiplied by the group's mean stretcher (assuming that, within group, the stretcher is uncorrelated with the change in $w_i^\star$). Suppose the common-scale level rises by 2 points for group A and by 1 point for group B, so that A's true trend exceeds B's by 1 point, and suppose A's mean stretcher is 0.5 while B's is 1.5. The measured trends are then 1 and 1.5, so the estimated difference in trends is −0.5: opposite in sign to the truth.

IV Econometric Correction Using Calibration Questions

The core idea of scale-use correction is to estimate respondents' CQ translation functions and then use them to adjust the analysis of SRs. In Section VII, we detail the assumptions that imply the existence of CQ translation functions and justify using them to adjust SRs; here, we take them as given. In Section VIII, we discuss *which* CQs should be used for estimating the translation functions—for example, vignettes or visual CQs; here, we take as fixed some set of CQs, denoted $\mathbb{C}$, indexed by $c = 1, 2, ..., C$. We denote the set of individuals $\mathcal{I}$ and its size I.

As a preliminary to scale-use correction, we use the CQ responses to estimate each respondent's intercept a_i and stretcher b_i. We do so in two steps. First, for each CQ c, we consistently (as the number of individuals in the sample grows, $I \to \infty$) estimate its common-scale rating w_c by the mean response across individuals: $\hat{w}_c = \frac{1}{I}\sum_i r_{ic}$. In our sample of I = 10,100 respondents, the standard errors are extremely small (averaging 0.4% of $\hat{w}_c$).

Second, we estimate each respondent i's CQ translation function by an OLS regression of r_{ic} on $\hat{w}_c$, with one observation per CQ. By equation (2), the intercept and slope of this regression are estimates $\hat{a}_i$ and $\hat{b}_i$. Intuitively, $\hat{a}_i$ measures how high i rates the CQs relative to the sample mean, and $\hat{b}_i$ how much i spreads the ratings out relative to the sample mean. Assuming negligible estimation error in $\hat{w}_c$, these estimators are unbiased, and they are consistent as the number of CQs grows ($C \to \infty$).

In practice, however, although assuming $I \to \infty$ may be reasonable, we anticipate that C will generally be small. In that case, $\hat{a}_i$ and $\hat{b}_i$ will have substantial estimation error, which acts as measurement error when they are used as proxies for a_i and b_i. For example, in our data with three Life Satisfaction vignettes, the mean standard errors of $\hat{a}_i$ and $\hat{b}_i$ are 52.47 and 0.90, respectively, larger than our estimates of the standard deviations of a_i and b_i themselves, 30.44 and 0.38 (Web Appendix D.1).

For some analyses, this measurement error does not generate bias.[8] However, for our main purpose, correcting for scale-use heterogeneity, it means that the most straightforward approach, inverting the estimated translation function to recover each individual's common-scale rating, is unlikely to be fruitful (unless C is quite large). That is, consider the estimator

$$\hat{w}_i^\star = \frac{r_i^\star - \hat{a}_i}{\hat{b}_i} \underset{\substack{I\to\infty \\ C\to\infty}}{\xrightarrow{p}} w_i^\star + \frac{\eta_i^\star}{b_i}.$$

[8] Here are three examples. First, in Web Appendix Table L.2, we use $\hat{a}_i$ and $\hat{b}_i$ as dependent variables in regressions to study demographic correlates of scale use. Second, in Web Appendix D.1, with additional assumptions we estimate the standard deviations of a_i and b_i from the CQ responses by maximum likelihood, correcting for response error. Our estimates of 30.44 and 0.38, respectively, imply substantial scale-use heterogeneity in our sample. Third, combining the results of these two analyses, we find that the fraction of scale-use heterogeneity explained by our measured demographic variables is only 4.0% and 4.1% of the total variances in a_i and b_i, respectively.

In the large-sample limits of both individuals and CQs, this estimator converges to $w_i^\star$ plus response error. However, the division by $\hat{b}_i$, which is noisy and may be close to zero, means that the number of CQs would have to be large to avoid outliers and large biases. Indeed, in Web Appendix J, we report regressions with $\hat{w}_i^\star$ as the dependent variable, and we find estimates that remain too imprecise to be useful even after winsorizing. Therefore, we seek an econometric strategy that is consistent as $I \to \infty$ but does not rely on large C and does not involve dividing by $\hat{b}_i$.[9]

Our strategy starts from the coefficient vector $\hat{\boldsymbol{\gamma}}_r$ of the naïve regression of $r_i^\star$ on $\mathbf{x}_i$ and subtracts an estimated scale-use adjustment, $\hat{\boldsymbol{\gamma}}_{\text{adj}}$, to obtain an estimate of $\boldsymbol{\gamma}_w$, the coefficient vector from the infeasible regression (1). Before deriving what we call our method-of-moments (MM) estimator and stating its key assumptions, we list the five steps for implementing it, given $\hat{a}_i$ and $\hat{b}_i$:

Step (i): Estimate the population mean of the SR, $\mathbb{E}\left(r_i^\star\right)$, by its sample mean, $\hat{\mathbb{E}}\left(r_i^\star\right) \equiv \frac{1}{I}\sum_i r_i^\star$.

Step (ii): For each respondent i, use $\hat{a}_i$ and $\hat{b}_i$ to construct the *estimated CQ translation function*:

$$\hat{\tau}_i(w) = \hat{a}_i + \hat{b}_i w.$$

Step (iii): For each respondent i, plug the sample mean of SRs into the estimated CQ translation function to obtain what we call the *(scale-use-only) benchmark*:

$$\hat{\tau}_i\left(\hat{\mathbb{E}}(r_i^\star)\right) = \hat{a}_i + \hat{b}_i\hat{\mathbb{E}}\left(r_i^\star\right).$$

The benchmark is an estimate of the rating respondent i would give a CQ at common-scale level $\hat{\mathbb{E}}\left(r_i^\star\right)$. Because that level is held fixed across respondents, variation in the benchmark reflects scale use alone.[10]

Step (iv): Regress the benchmark $\hat{\tau}_i\left(\hat{\mathbb{E}}(r_i^\star)\right)$ on $\mathbf{x}_i$. The estimated coefficient vector is the scale-use adjustment $\hat{\boldsymbol{\gamma}}_{\text{adj}}$. The standard errors can be used for confidence intervals and hypothesis tests on the scale-use adjustment.

Step (v): Calculate the scale-use-corrected coefficient vector $\hat{\boldsymbol{\gamma}}_w = \hat{\boldsymbol{\gamma}}_r - \hat{\boldsymbol{\gamma}}_{\text{adj}}$. To find confidence intervals and conduct hypothesis tests on $\hat{\boldsymbol{\gamma}}_w$, use the standard errors from regressing $r_i^\star - \hat{\tau}_i\left(\hat{\mathbb{E}}(r_i^\star)\right)$ on $\mathbf{x}_i$.

Notice that the measurement errors in $\hat{a}_i$ and $\hat{b}_i$ do not generate bias in the MM estimator because they are on the left-hand side of the regressions in steps (iv) and (v) (and the estimation error in $\hat{\mathbb{E}}\left(r_i^\star\right)$ asymptotically vanishes). As a default, we recommend using heteroskedasticity-robust (HC1)

[9] Another straightforward econometric strategy that does not work is controlling for the CQ responses r_{ic} when regressing $r_i^\star$ on $\mathbf{x}_i$. There are two problems. First, response error in the CQ responses gives rise to noisy regressors, which are incomplete controls. Second, if r_{ic} is correlated with $w_i^\star$, then the estimator is not consistent for $\boldsymbol{\gamma}_w$, even after measurement-error correction.

[10] Conditional on the mean SR targeted, the benchmark is a linear combination of CQ ratings. Web Appendix Table L.3 gives the weights in those linear combinations for same-dimension benchmarks in Survey 1.

standard errors because they are quick to compute and we find they provide a close approximation to bootstrap standard errors (Web Appendix K.4).

Intuitively, the MM estimator starts from the naïve regression coefficients and subtracts an approximation to the bias from scale-use heterogeneity. To derive it, begin with the bias formula in equation (4). Then, make two approximations: (i) in the first term, replace $\mathbb{E}\left(w_i^\star\right)$ with $\mathbb{E}\left(r_i^\star\right)$; (ii) assume the second term is equal to zero. These two approximations are exact under two moment conditions: (i) $\text{Cov}(b_i, w_i^\star) = 0$, which implies $\mathbb{E}(r_i^\star) = \mathbb{E}(w_i^\star)$; and (ii) zero coskewness, $\mathbb{E}\left[(b_i - \mathbb{E}(b_i))\left(\mathbf{x}_i - \mathbb{E}(\mathbf{x}_i)\right)\left(w_i^\star - \mathbb{E}(w_i^\star)\right)\right] = 0$. With these approximations, we have the vector of moment conditions $\text{Cov}\left(\mathbf{x}_i, r_i^\star\right) - \text{Cov}\left(\mathbf{x}_i, w_i^\star\right) = \text{Cov}\left(\mathbf{x}_i, a_i + b_i\mathbb{E}(r_i^\star)\right)$. Replacing each term by its sample counterpart, premultiplying by the inverse sample variance of $\mathbf{x}_i$, and rearranging yields the MM estimator: $\hat{\boldsymbol{\gamma}}_w = \hat{\boldsymbol{\gamma}}_r - \hat{\boldsymbol{\gamma}}_{\text{adj}}$, where $\hat{\boldsymbol{\gamma}}_{\text{adj}}$ is the coefficient vector from regressing the benchmark $\hat{\tau}_i\left(\hat{\mathbb{E}}(r_i^\star)\right) = \hat{a}_i + \hat{b}_i\hat{\mathbb{E}}\left(r_i^\star\right)$ on $\mathbf{x}_i$.[11]

The MM estimator has several attractive features. It is computationally quick and transparent, with the scale-use adjustment coming directly from the regression of the benchmark on $\mathbf{x}_i$. Because the benchmark is computed at the individual level, the standard errors account for within-group scale-use heterogeneity. Its only parametric assumption, linearity of the CQ translation function, is testable, as illustrated in Section II.B, and it requires no distributional assumptions. Although linearity is most natural when the number of response categories is large, results in Web Appendix I and simulations in Web Appendix K.3 indicate that the estimator continues to perform well with as few as three response categories.

Finally, we find no evidence that its two approximations matter in practice. In Web Appendix F.1, we develop a semiparametric estimator for $\boldsymbol{\gamma}_w$ that does not impose them. It is substantially more complex and relies on different, arguably weaker, assumptions: mainly an assumption about the distribution of b_i, a first-order approximation, and normal, mutually independent response errors. Across six SR dimensions, all differences between the MM and semiparametric coefficients are statistically indistinguishable from zero at the 5% false-discovery-rate threshold (Web Appendix Table F.1), suggesting that any violations of the moment-condition assumptions are too small to

[11] Since the stretcher is multiplicative, it is perhaps surprising that *subtracting* the benchmark can eliminate the bias. Part of the explanation is that the benchmark itself multiplies the estimated stretcher by the SR mean. An alternative derivation of the MM estimator provides insight into the importance of choosing a benchmark at the SR mean (rather than subtracting any other average of CQ ratings): subtract equation (2) from equation (3), $r_i^\star - r_{ic} = b_i\left(w_i^\star - w_c\right) + \eta_i^\star - \eta_{ic}$, and consider the covariance with $\mathbf{x}_i$: $\text{Cov}\left(\mathbf{x}_i, r_i^\star - r_{ic}\right) = \text{Cov}\left(\mathbf{x}_i, w_i^\star\right) + \text{Cov}\left(\mathbf{x}_i, b_i\right)\mathbb{E}\left(w_i^\star - w_c\right) + \mathbb{E}\left[(b_i - \mathbb{E}\left(b_i\right))\left(\mathbf{x}_i - \mathbb{E}\left(\mathbf{x}_i\right)\right)\left(w_i^\star - \mathbb{E}\left(w_i^\star\right)\right)\right]$. The first term on the right-hand side is the object of interest, and the last term is the coskewness, assumed to be zero by moment condition (ii). The middle term is bias from the stretcher due to its covariance with $\mathbf{x}_i$. While the intercept is eliminated by subtracting r_{ic} regardless of which CQ c is used, eliminating the stretcher bias requires subtracting the rating of a CQ, r_{ic}, that has common-scale rating $\mathbb{E}\left(w_i^\star\right)$. Because a respondent's estimated CQ translation function can be evaluated at any common-scale rating, we can predict each respondent's rating of a CQ even when no actual CQ has that common-scale rating. Under moment condition (i), the scale-use-only benchmark is that predicted rating.

detect in our data. In simulations, the MM approximation is tight across a range of data-generating processes (Web Appendix K).

A subtle but important point concerns the scale on which the MM estimator measures the dependent variable. Because we adopted the CQ common scale for the SRs, the scale of $w_i^\star$ depends on the set of CQs: different CQ sets define different common scales, related to one another by strictly increasing but not necessarily linear transformations (Section VII.E). The $w_i^\star$ scale also differs from the $r_i^\star$ scale even in the absence of scale-use heterogeneity, in which case the two are related by a linear rescaling. A linear change of scale leaves *ratios* of the coefficients (other than the constant) unchanged. Therefore, to assess whether scale-use correction changes the coefficients, the relevant question is whether these ratios change. The same question is relevant for assessing whether a change of CQ set changes the coefficients, but only approximately: because the change of scale between two CQ sets need not be linear, the ratios are invariant only to the extent that a linear approximation of that change over the range of the data is accurate. We generally expect such an approximation to be reasonable, so we treat the comparison of ratios as informative in that case as well. In Web Appendix G, we develop a test of proportional coefficients, which we use in subsequent sections.

V Illustrative Application: Life Satisfaction Regression

One commonly used SR is self-reported well-being (SWB). Economists' most frequent use of SWB data is a regression of an SWB measure, often life satisfaction, on demographic and other variables (for surveys of this literature, see Clark et al., 2008; Helliwell and Barrington-Leigh, 2010; Krueger and Stone, 2014; Layard and De Neve, 2023); well-known findings on age, gender, religiosity, and children come from such regressions. When the regressors include income, coefficient ratios have been used to price non-market goods, including, for example, unemployment, clean air, and the life of a relative (e.g., Di Tella et al., 2001; Levinson, 2012; Deaton et al., 2010).[12]

Table 2 reports six versions of such a *life satisfaction regression.* All six use the same regressors: those shown in the table (age, income, employment status, marital status, education, and sex) and those listed in the table note (from log household size and ethnicity to when and where the survey was taken). Column 1 reports a "standard" regression, unadjusted for scale-use heterogeneity, of respondents' Life Satisfaction SR (on the 0–100 scale) on these regressors. It replicates some typical findings from the literature: life satisfaction is strongly positively associated with log income and being married (and not separated) and is strongly negatively associated with being unemployed.

[12]Our illustrative application follows standard practices in the literature, which largely focuses on associations, rather than on well-identified causal effects. Beyond concerns about causal identification, the economic interpretation of such regressions and their coefficient ratios requires a number of additional assumptions unrelated to scale use. These issues have been discussed elsewhere (e.g., Benjamin et al., 2024); we do not address them here.

(Findings related to age, sex, and education vary across studies.)

Table 2: Life Satisfaction Regression and Scale-Use Adjustment

	Dependent Variable: Life Satisfaction					
	Unadjusted	Intercept-Only Adj.	Method of Moments (MM)			CHOPIT
		Same Dimension	Cross Dimension		Same Dim.	Same Dim.
			Visual CQs	Vignette CQs		
	(1)	(2)	(3)	(4)	(5)	(6)
Demeaned age / 10	2.49††† (0.18)	2.33††† (0.21)	2.24††† (0.22)	2.31††† (0.21)	3.46††† (0.26)	2.63††† (0.23)
(Demeaned age)^2 / 100	0.50††† (0.09)	0.48††† (0.11)	0.50††† (0.11)	0.41††† (0.10)	0.43††† (0.13)	0.60††† (0.11)
Log(HH income)	4.20††† (0.31)	4.05††† (0.33)	3.28††† (0.37)	2.23††† (0.34)	1.71††† (0.41)	5.03††† (0.36)
Unemployed	−6.90††† (1.10)	−6.87††† (1.14)	−7.78††† (1.32)	−7.10††† (1.24)	−8.47††† (1.42)	−6.26††† (1.21)
Married, not separated	6.09††† (0.56)	5.94††† (0.62)	5.39††† (0.67)	5.83††† (0.65)	6.82††† (0.77)	7.29††† (0.70)
College grad	0.72 (0.49)	0.55 (0.57)	2.58††† (0.59)	−2.45††† (0.58)	−1.93†† (0.71)	2.74††† (0.65)
Male	−1.32†† (0.46)	−1.71††† (0.53)	−0.61 (0.55)	1.36†† (0.55)	−0.72 (0.66)	−1.55†† (0.61)
Obs.	10,100	10,100	10,100	10,100	10,100	10,100
p-value: proportional to Col. (1)	1.00	1.0×10^{-3}	6.6×10^{-7}	6.9×10^{-34}	2.8×10^{-31}	1.1×10^{-14}

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 (as a state-level proxy for urban/rural), U.S. Census region, and survey day of the week. Unemployed coefficient: base category is full-time employed. Intercept-only adjustment (Column 2) subtracts the simple average of respondent's CQ ratings. Same Dimension (Columns 2, 5, and 6): Life Satisfaction CQ trio. Cross-Dimension Visual CQs (Column 3): Confidence, Shape, and Darkness. Cross-Dimension Vignette CQs (Column 4): Breathe and Remember. CHOPIT (Column 6): Estimated using the package from King et al. (2004) (except we discretized the dependent variable and rescaled the coefficients; for details, see Web Appendix H). Heteroskedasticity-robust standard errors in parentheses. †††, ††, †: significance at the 1%, 5%, and 10% levels, FDR-adjusted. Bottom row: p-value from proportionality-of-coefficients test, using 500 joint respondent-level bootstrap draws; null = column's vector of non-constant coefficients is proportional to Column 1's (unadjusted) vector.

Column 2 applies an intercept-only adjustment: it subtracts the simple average of the respondent's ratings of the Life Satisfaction vignette CQs from their Life Satisfaction SR before regressing it on the demographics. This adjustment implicitly assumes that every respondent's stretcher equals one. Columns 3–5 apply our MM estimator with three different sets of CQs: the three (cross-dimension) visual CQ trios (Confidence, Shape, and Darkness; Column 3), the two (cross-dimension) vignette CQ trios seen by all respondents (Breathe and Remember; Column 4), and the same-dimension Life Satisfaction trio (Column 5). We view Columns 3–5 as applying increasingly dimension-specific correction, on the view that scale use for the Breathe and Remember vignettes is likely closer to scale use for the Life Satisfaction vignettes than is scale use for the visual CQs. Column 6 is a Compound Hierarchical Ordered Probit (CHOPIT), proposed by King et al. (2004): an ordered probit whose cutpoints depend on the demographics (see Web Appendix A.1). We estimate it with the King et al. package, using the Life Satisfaction vignette CQs, and rescale the coefficients for comparability with the other columns (Web Appendix H). We highlight four findings.

First, scale-use correction, with any of the specifications in Columns 2–6, leaves the most robust findings from the literature intact qualitatively. The associations with income, unemployment, and marriage noted above hold in every column, as do the strongly positive coefficients on age and age^2, which some recent studies find and others do not (Blanchflower et al., 2025).

Second, other coefficients change across columns in magnitude, in whether they are statistically distinguishable from zero, and even in sign. The coefficients on being a college graduate and on

being male are each statistically distinguishable from zero in some columns and not others, in some negative and in others positive.

Third, coefficient ratios—the relevant output for scale-use-corrected regressions (see Section IV)—can change substantially with scale-use correction. The table's bottom row reports the p-value from a test that each column's coefficient vector is proportional to Column 1's (Web Appendix G; covariance matrix from 500 bootstrap draws). Every adjusted column rejects proportionality decisively.

Some of these changes matter substantively. Consider the ratio of the unemployment coefficient to the log-income coefficient, which prices unemployment in income—a major application of SWB data (Di Tella et al., 2001; Frijters et al., 2020). Unadjusted (Column 1), the ratio, in absolute value, is $6.90/4.20 \approx 1.6$. The intercept-only correction barely moves it, to 1.7 (Column 2).[13] MM correction raises it to 2.4, 3.2, and 5.0 as the CQs become more dimension-specific (Columns 3–5), while CHOPIT lowers it to 1.2 (Column 6). The price of marriage in income behaves similarly: 1.5 in Columns 1 and 2; 1.6, 2.6, and 4.0 in Columns 3–5; and 1.4 in Column 6.[14]

Columns 5 and 6 provide the cleanest comparison between MM and CHOPIT. Signs and significance mostly agree, but magnitudes often differ in varying directions, which is why the coefficient ratios differ so much. We show in Web Appendix H that the differences mainly arise because the existing CHOPIT implementation estimates scale-use heterogeneity from both SR and CQ data, so heterogeneity in the SR's common-scale level can confound the scale-use estimates; MM uses only the CQ data. When we modify CHOPIT to use only the CQ data, its estimates move much closer to MM's for most coefficients, though the college-graduate coefficient still has the opposite sign. Even so, the modified CHOPIT retains major disadvantages relative to MM: it relies on untestable distributional assumptions, and its many scale-use parameters are hard to assess for plausibility (see Web Appendix A.1).

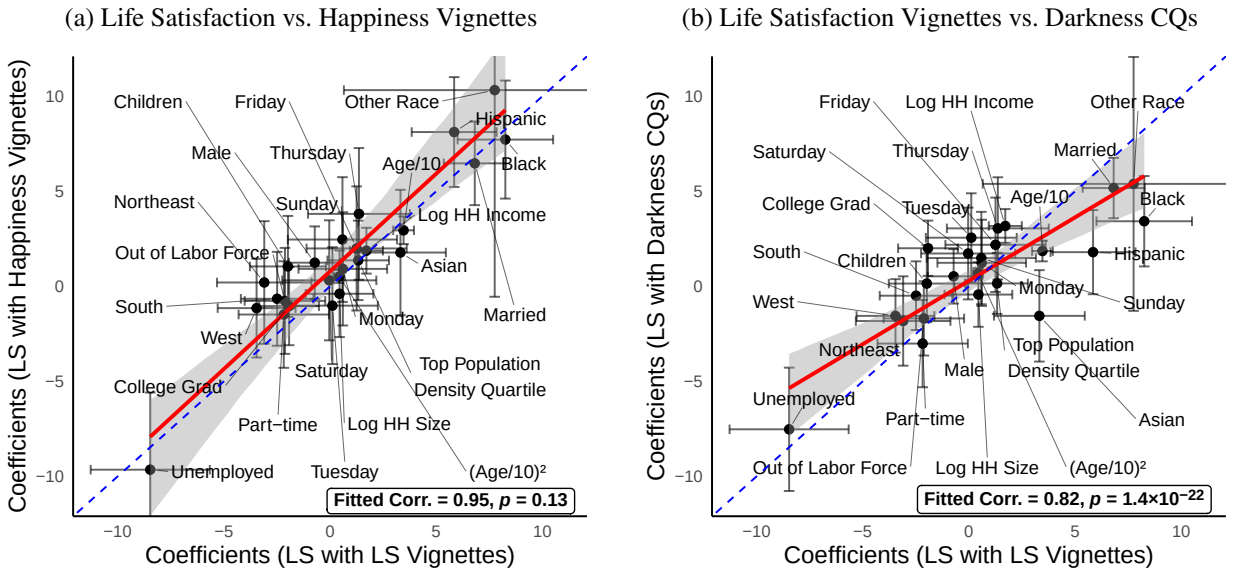
Fourth, the choice of CQs matters, and it matters less when the CQ dimension is closely related to the SR dimension. Figure 5 illustrates. Each panel plots against each other the 25 non-constant coefficients from two MM-adjusted regressions that differ only in the CQ trio used (the regressions are like those in Columns 3–5; the Life Satisfaction one is identical to Column 5), along with the

[13] In general, the intercept-only correction can be expected to correct for less of the scale-use heterogeneity than the MM estimator does. Intuitively, it effectively corrects for vertical shifts of the CQ translation function at the mean of the CQs, rather than at the (much higher) mean of the SRs, as the MM does. These shifts differ due to stretcher heterogeneity. Indeed, the MM estimator neutralizes stretcher heterogeneity *because* it adjusts at the mean of the SRs (as explained in footnote 11).

[14] These differences are statistically distinguishable from zero at standard thresholds. Using a sandwich covariance that accounts for estimation of the benchmark weights and for the correlation between the two regressions, the delta-method estimate of the change in the unemployment-to-income ratio between Columns 1 and 5 is 3.30 (SE 1.49, $p = 0.03$; 95% CI $[0.38, 6.22]$), and in the marriage-to-income ratio it is 2.53 (SE 1.05, $p = 0.02$; 95% CI $[0.46, 4.59]$). A Fieller-type test, which is better behaved when the denominator is imprecisely estimated, rejects more strongly in both cases ($p = 4.3 \times 10^{-5}$ and $p = 3.8 \times 10^{-7}$, respectively).

correlation between the two regressions' fitted values and the p-value from the proportionality test. Replacing the Life Satisfaction vignettes with Happiness vignettes (Panel a) leaves the coefficient ratios statistically indistinguishable with a fitted correlation of 0.95, whereas replacing them with the visual Darkness CQs (Panel b) yields a strong rejection of proportionality and a fitted correlation of 0.82.

Figure 5: Coefficients From Scale-Use-Adjusted Life Satisfaction Regression Using Different CQs



Notes: Each panel: pairs of demographic coefficients from two scale-use-adjusted 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; see Web Appendix G). Sample: 10,100 respondents, except the Happiness-vignette regression in panel (a), which uses the 5,063 respondents shown that trio.

Web Appendix Table L.4 repeats the exercise for all 15 CQ trios and all nine SR dimensions (Web Appendix Table L.5 shows the estimated constants of proportionality and Web Appendix Figure L.1 plots four additional pairs). In the Life Satisfaction regression, proportionality is rejected for every Survey 1 visual trio and for every other-dimension vignette trio except Happiness and No Anxiety, and fitted-value correlations are correspondingly lower for the visual CQs (0.72–0.88) than for the other-dimension vignettes (0.87–0.96).[15] Other SR dimensions look similar, except No Anxiety, whose coefficients are better matched by visual CQs than by the other vignettes.

Proportional coefficients across CQ sets is exactly the condition under which cross-dimension correction reproduces same-dimension results, so these findings have a direct practical implication,

[15]The four Survey 2 trios, which include the Height and Weight visual trios, do not fit this pattern: their correlations are 0.90–0.92 and none is rejected. Their tests are also the least powered in the table, with 2,503 linked respondents versus 5,000 to 10,100 for the Survey 1 trios.

to which we return in Section VIII: when same-dimension CQs are unavailable, CQs in a closely related dimension can be a reasonable substitute, but CQs in a distant dimension generally cannot.

VI Validation

As a validation test (inspired by Oswald, 2008), we examine whether our Height and Weight SRs, rated on the 0–100 scale, become more strongly related to height and weight reported on objective scales (inches and pounds) after scale-use correction.[16] The logic is as follows. If scale-use heterogeneity and the response errors we model were the only sources of discrepancy between an SR and its objective-scale counterpart, if the objective-scale measure were error-free, and if the two were linearly related, then after scale-use and response-error correction their correlation would be one. Each premise is an approximation (for example, height SRs may reflect implicit comparisons with idiosyncratic reference groups), so we expect a correlation below one, but closer to one after correction. Because scale-use correction changes the units of the SR, we report correlations rather than regression coefficients. And because response error in the SR attenuates correlations, we correct the SR's standard deviation for response error before standardizing.

Table 3 reports the correlations. The top row uses the unadjusted SR: 0.49 (SE = 0.02) for height and 0.23 (0.02) for weight. The middle row corrects the SR's standard deviation for response error but not for scale-use heterogeneity. It isolates the effect of the response-error correction, so that the bottom row's further increase can be attributed to scale-use correction. Because the SR's response-error variance is not separately identified from the variance of the common-scale SR, we assume as a baseline that it equals 50% of the response-error variance estimated from the same-dimension CQs. Mechanically, the correction raises the correlations, to 0.56 (0.02) for height and 0.32 (0.03) for weight.

The bottom row applies scale-use correction using the same-dimension (Height and Weight) CQs. As in Section IV, we subtract the scale-use-only benchmark from the SR. Standardizing additionally requires the standard deviation of the SR in common-scale units, which we estimate with a method-of-moments estimator (Web Appendix D.4) that assumes the stretcher b_i is independent of the common-scale level $w_i^\star$ and nets out response error under the same 50% assumption. The correlations rise further, to 0.60 (0.02) for height[17] and 0.51 (0.05) for weight.

[16] Each objective-scale measure is the respondent's most recent prior report to the UAS rather than a contemporaneous one. The gap between it and the Survey 2 SR is substantial and varies widely: median 5.1 years for height (interquartile range 2.4–8.8) and 2.2 years for weight (1.9–2.7), because the UAS asks about height far less often than weight. The long gaps for height are unlikely to matter, since adult height is stable in adulthood, but intervening weight change makes the objective-scale weight measure a noisier proxy for weight at the time of the SR. Because this affects every row of Table 3, it attenuates the levels of the correlations more than the comparison between rows.

[17] Among our CQs, the High Height CQ has by far the most top-coding (11.1%; Web Appendix Table C.1) and, likely for this reason, a small negative coefficient in Table 1. When we use a benchmark based on only the Low and Medium

Table 3: Correlations of Subjective-Scale Self-Ratings (SR) with Objective-Scale Measures

	Height (Inches)	Weight (Pounds)
Unadjusted SR	0.49 (0.02)	0.23 (0.02)
Unadjusted, Response-Error Corrected SR	0.56 (0.02)	0.32 (0.03)
Method of Moments (MM) Corrected SR	0.60 (0.02)	0.51 (0.05)
Obs.	2,474	2,481

Notes: Correlations of self-reports on the 0–100 subjective scale—SRs—for height and weight with the corresponding objective-scale measures. 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 SRs and same-dimension (visual) CQs; Web Appendix Table B.2b displays the CQ stimuli. Section I describes the objective-scale measures. 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 corresponding same-dimension scale-use-only benchmark, divided by the MM-corrected SD. Second and third rows assume SR response-error variance is 50% of the corresponding CQ response-error variance (see text; Web Appendix Table L.6 reports alternative assumptions). Bootstrap standard errors (100 resamples) in parentheses.

In sum, scale-use correction raises the response-error-corrected correlation modestly for height (from 0.56 to 0.60) and substantially for weight (from 0.32 to 0.51). Web Appendix Table L.6 shows that this conclusion holds qualitatively across the full admissible range of alternative response-error assumptions (0%–75%; at 80%, the correction implies negative variances). It also shows what happens when other-dimension CQs are used instead: with one exception, they yield lower corrected correlations than the same-dimension CQs, and in most cases no higher than with no scale-use correction at all.[18] Together with Section V's finding that the choice of CQ dimension matters, this supports using same-dimension CQs when they are available, as we discuss in Section VIII.

VII Theoretical Framework

In this section, we introduce three assumptions and show that they are sufficient—and, once the first is suitably weakened, also necessary—for CQ responses to identify the ranking of respondents' common-scale SRs. We end by showing that a strengthening of the first assumption implies a linear CQ translation function, which, with response error added, is the econometric model of Section III. We discuss when each assumption may fail, and we emphasize that the assumptions are weaker than they might appear: they permit many complications one might expect to confound scale-use correction. Each is in principle testable, and we report some tests using our data.

Height CQs, the corrected correlation is 0.63 (0.02).

[18]The exception is correcting the Height SR with the Weight CQs, which yields a correlation of 0.79 under the baseline 50% assumption. (Conversely, correcting the Weight SR with the Height CQs yields 0.33, below the same-dimension 0.51.)

VII.A Formal Setup

We refer to the attribute rated in a question as its *dimension*; examples include life satisfaction, overall health, and darkness of a circle. SRs can only be about certain dimensions (e.g., life satisfaction but not darkness of a circle), whereas CQs can be about any dimension.

We divide the factors that may affect ratings into two categories. Those relevant to the SR dimension form the respondent's *situation* vector, ω, whose elements we call *subdimensions*; those that affect only scale use form the respondent's *context* vector, which includes, for example, previous survey questions and questions on the same screen. The situation may also affect scale use, for example if it includes past situations, reference points, or mood. We refer to a respondent's situation during the survey as *current*, denote i's current situation by ω_i, and assume it is the same when answering the SR and CQ questions. Denoting the set of respondents by $\mathcal{I}$, we define $\mathcal{S} := \{\omega_i : i \in \mathcal{I}\}$ as the set of respondents' current situations.

We do *not* assume that context is the same when answering the SR and CQ questions; the survey order of the SR and CQs, for example, could affect ratings. Formally, context always includes the individual's identity i and the times at which the respondent answers the SR and each CQ (or, for the hypothetical CQs introduced below, would answer them); its remaining elements are functions of these. Because our analyses are cross-sectional, we suppress the times, and $i \in \mathcal{I}$ should be read as jointly indexing the individual and their sequence of contexts.

We denote the set of response-scale options for the survey questions, assumed to have at least two elements, by $\mathcal{R} \subseteq \mathbb{R}$. For example, for a 5-point Likert scale, the elements of $\mathcal{R}$ are indexed by the integers 1–5, whereas for a slider, $\mathcal{R}$ is continuous. Some or all response-scale options could have labels attached to them, as long as the options are ordered monotonically. We denote by $\mathcal{R}_i^\star \subseteq \mathcal{R}$ the ordered subset of the response options that i would use in at least one situation $\omega \in \mathcal{S}$ (e.g., on our 0–100 scale, only the range 40–70 or only multiples of 5).

Following Oswald (2008), we refer to the mapping $\rho_i^\star : \mathcal{S} \to \mathcal{R}_i^\star$ from any situation $\omega \in \mathcal{S}$ to how respondent i would rate their own situation if it were ω as i's (*self-*) *reporting function*. To place survey responses on a common scale, we arbitrarily choose someone, denoted $\bar{\imath}$, as the *reference individual*. We call that person's rating of a situation ω that situation's *SR-common-scale rating*, with value $w = \rho_{\bar{\imath}}^\star(\omega)$.

VII.B Interpersonal Comparability and the Self-Rating (SR) Translation Function

When researchers study SRs, which lack an objective scale, they nonetheless assume that the questions ask about an underlying quantity that is interpersonally comparable.

Interpersonal Comparability Presupposition (Common Ranking of Situations).

$$\textit{For all } i, i' \in \mathcal{I} \textit{ and } \omega, \omega' \in \mathcal{S}, \quad \rho_i^\star(\omega) \gtreqless \rho_i^\star(\omega') \iff \rho_{i'}^\star(\omega) \gtreqless \rho_{i'}^\star(\omega').$$

This assumption states that all respondents rank the situations in $\mathcal{S}$ identically. In an analogy to utility theory, Common Ranking of Situations means that any two individuals have the same "iso-rating curves" but may assign different ratings to those curves. The Interpersonal Comparability Presupposition is inherent to any analysis that involves ranking levels across individuals: there must be a (unidimensional) level to rank.

We do not need to assume the Interpersonal Comparability Presupposition because it will be implied by our assumptions below, but we highlight it here for two reasons. First, it helps clarify when the exercise of comparing SRs across respondents is sensible at all. The assumption could be violated if respondents differ in how they interpret the wording of the question (as King et al., 2004, point out in a related context). For example, one respondent might interpret a life satisfaction question as asking about life up until now, while another interprets it as including expectations about the future (Benjamin et al., 2023). It could also be violated if respondents differ in their weights when aggregating across subdimensions. For example, critics of the literature on life satisfaction sometimes argue that, because of heterogeneity in how people aggregate across subdimensions of life satisfaction, people would rank situations differently; if so, life satisfaction fails the Interpersonal Comparability Presupposition.[19] The Interpersonal Comparability Presupposition is most plausible for narrower and more objective dimensions, such as ***Your ability to breathe in and out easily***, because for such dimensions, there is less scope both for different interpretations of question content and for different aggregation weights.

Second, the Interpersonal Comparability Presupposition implies the existence of a strictly monotonically increasing "translation function" that maps from the common-scale rating of any situation to how the respondent would self-rate in that situation. Fixing the reference individual $\bar{\imath}$, we define the *SR translation function*, $\tau_i^\star : \mathcal{R}_{\bar{\imath}}^\star \to \mathcal{R}_i^\star$, as $\tau_i^\star(w) := \rho_i^\star\left((\rho_{\bar{\imath}}^\star)^{-1}(w)\right)$, where $(\rho_{\bar{\imath}}^\star)^{-1}(w)$ picks out any ω such that $\rho_{\bar{\imath}}^\star(\omega) = w$. In order for this definition to correspond to a well-defined function, whenever some set of situations is mapped by $\rho_{\bar{\imath}}^\star(\cdot)$ to a constant rating, say 70, the identical set must be mapped by $\rho_i^\star(\cdot)$ to a constant rating, say 80. This is implied by the Interpersonal Comparability Presupposition.

The fundamental problem of scale-use correction is that the SR translation functions are not

[19] Defenders sometimes argue that this heterogeneity is a feature, not a bug: in terms of our framework, they are assuming that the only element of ω is the respondent's (already aggregated) rating of life satisfaction. In that view, two respondents with the same facts about their life but different life satisfaction ratings are in different situations, and the Interpersonal Comparability Presupposition is satisfied if they agree that a respondent with a higher rating than another has higher life satisfaction. But by treating the ratings themselves as interpersonally comparable, this view assumes away the problem of scale-use heterogeneity.

identified because we never observe how a respondent (including the reference individual) would rate another respondent's situation. Although our assumptions imply that each SR translation function exists and is strictly increasing, a surprising result of our theoretical analysis will be that scale-use correction can be justified *without* any further assumptions about these functions.

VII.C Calibration Questions (CQs) and the CQ Translation Function

Calibration questions (CQs) provide a way to translate the response scale across respondents because respondents rate the same CQs, rendering the "CQ translation function" observable. Our analysis encompasses both *same-dimension CQs*—vignettes about the SR dimension or, as for our Height and Weight CQs, images of it—and *other-dimension CQs*, which may likewise be vignettes or visual CQs (e.g., the darkness of a circle).

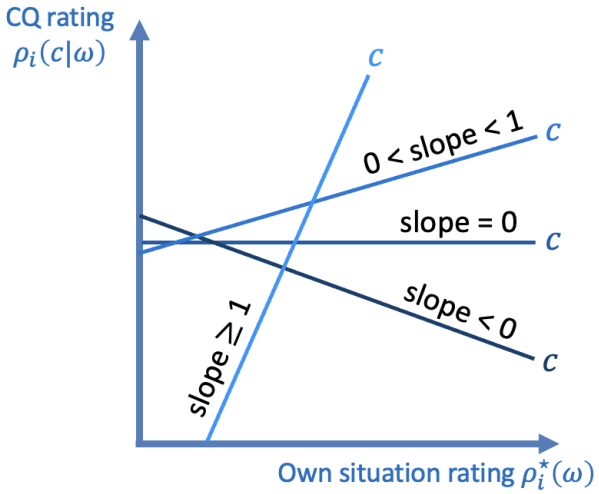
In a given dimension, we denote a CQ by $c \in \mathcal{C}$. We allow CQ ratings to depend on both the respondent i and the respondent's situation ω. We refer to the mapping $\rho_i(\cdot|\omega) : \mathcal{C} \to \mathcal{R}$ from CQ c to respondent i's rating in situation ω as the *CQ reporting function*, and denote it $\rho_i(c|\omega)$. In the theory below, we will make use of how respondents would rate particular CQs that are not actually asked in the survey; for such a CQ c, $\rho_i(c|\omega)$ denotes how i would rate c, in the context in which i would answer it, were it asked.

We assume that $\mathcal{C}$ includes the CQs asked in the survey and, for every respondent $i \in \mathcal{I}$ and situation $\omega \in \mathcal{S}$, at least one CQ c (possibly hypothetical) that i, when in situation ω, would rate the same as ω: $\rho_i(c|\omega) = \rho_i^\star(\omega)$. We denote the ordered subset of the response options that i, when in situation ω, would use to rate at least one CQ in $\mathcal{C}$ by $\mathcal{R}_i(\omega) \subseteq \mathcal{R}$, and we assume that for all ω, $\mathcal{R}_i(\omega)$ has at least two elements.

By allowing the CQ reporting function to depend on the respondent i and the respondent's situation ω, our framework accommodates a number of possible systematic differences in how people rate CQs. To analyze how the CQ rating depends on the situation, it is helpful to graph a *rating curve*: for a fixed respondent i and CQ c, the x-axis is the respondent's rating of the situation $\rho_i^\star(\omega)$, and the y-axis is the respondent's CQ rating, $\rho_i(c|\omega)$. Rating curves are actually correspondences (because two situations that are rated the same could cause the respondent to rate CQ c differently) and could be discontinuous, non-monotonic, or non-linear; we graph them as linear functions of a continuously varying ω to focus on the core issues.

Each line in Figure 6 illustrates a possible rating curve corresponding to a particular CQ c. In the "slope = 0" scenario, the CQ rating does not depend on ω. It is most plausible for a visual CQ such as Darkness, whose ratings are unlikely to depend on the circumstances relevant to, say, Life Satisfaction. (The anchoring-vignette literature discussed in Web Appendix A, which studies only vignettes, implicitly assumes this scenario.) The "slope < 0" rating curve could arise from

Figure 6: Four Possible Rating Curve Cases



self-norming if the CQ is in the same dimension as the SR and the respondent uses their own situation (at least partially) as a reference point when rating the vignette; the higher i's own situation, the lower the vignette is rated. The "0 < slope < 1" rating curve could arise from what we call *vignette completion* if the CQ is a same-dimension vignette: the respondent "fills in" missing vignette details by making some assumption about those details; if a higher own situation leads to a higher extrapolated vignette situation, then the rating curve is upward sloping. Finally, the "slope ≥ 1" rating curve can arise from *complete overwriting*: ignoring details of the CQ altogether and instead reporting one's own-situation rating—as if the CQ situation were the same as one's own. In the greater-than-one case, the CQ rating is more responsive to the respondent's own situation than their SR is. We view complete overwriting and beyond as implausible. Assumption 3 below will rule it out. Our scale-use-correction framework accommodates the other three scenarios, as long as our Assumptions 1 and 2 below are satisfied.

Allowing the rating of a CQ to depend on i is of course necessary in order to accommodate scale-use differences across respondents. But since i jointly indexes the respondent and their contexts, the framework also accommodates effects of context on CQ ratings. For example, in criticizing the assumptions of King et al. (2004), Deaton (2011) points out that respondents may have an *empathy gap*: those who have not experienced the situation described in a vignette (e.g., having cancer) cannot fully appreciate how they would rate it if they experienced it themselves. An empathy gap makes CQ ratings depend on the respondent's experience, which in our framework is part of the situation if current and part of the context if past. Either way, it violates the anchoring-vignette literature's implicit assumptions but not ours.

Our first assumption is that CQ responses are interpersonally comparable:

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

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

Assumption 1 is analogous to the Interpersonal Comparability Presupposition but for CQs rather than SRs. The conditions under which it will fail are also analogous—i.e., interpersonal differences in interpretation or in weighting of subdimensions—but it could additionally fail if differences in respondent context or situation generate differences in CQ ranking. For example, suppose the sample of respondents have different health conditions, and we aim to study how serious each condition is for life satisfaction. Now, suppose one vignette character gets migraines and another has kidney stones. Assumption 1 would fail if respondents with experience of migraines rate the first vignette lower in life satisfaction while respondents with experience of kidney stones rate the second lower (an example of the empathy gap described by Deaton, 2011).

Fixing the reference individual $\bar{\imath}$ and $\bar{\imath}$'s current situation $\bar{\omega}$, we define the *CQ translation function* $\tau_i(\cdot|\omega) : \mathcal{R}_{\bar{\imath}}(\bar{\omega}) \to \mathcal{R}_i(\omega)$ by $\tau_i(w|\omega) := \rho_i\left((\rho_{\bar{\imath}})^{-1}(w|\bar{\omega})|\omega\right)$; it is the mapping from the common-scale rating of a CQ, w, to the rating that i would assign to that CQ when in situation ω. Assumption 1 implies that the CQ translation function is well-defined. In the special case of a linear CQ translation function, for any CQ c,

$$\tau_i(w_c|\omega) = a_i(\omega) + b_i(\omega)w_c,$$

where a respondent's intercept a_i and stretcher b_i may depend on the respondent's situation.

In principle, Assumption 1 is testable, ideally with a rich set of CQs varying independently along every subdimension. In practice, it can be tested only with the CQs for which data exist, and every test has limitations. Column 1 of Table 4 reports, for each CQ dimension, the fraction of respondents who share the full-sample ranking (Column 2 reports sample sizes). Respondents do not rank the CQs unanimously, especially for broad dimensions such as Life Satisfaction and Worthwhileness. But this comparison assumes no response error (i.e., $\text{Var}(\eta_{ic}) = 0$ in equation 2).

Because response error is pervasive in CQ ratings (Web Appendix C), a credible test of Common Ranking of CQs must allow for it. We therefore propose a test based on comparing mean CQ ratings across demographic subsamples: by the law of large numbers, group means are estimated precisely even when individual ratings are noisy, so the test is robust to response error with arbitrary heteroskedasticity. Figure 4 showed that, for every CQ dimension, sociodemographic subsamples share the same ranking of mean CQ ratings, but only for univariate splits. Columns 3–5 of Table 4 extend this analysis to multivariate splits: the groups are the cells formed by crossing five binary demographics (see the table notes). Column 3 reports the fraction of groups whose ranking of mean Low, Medium, and High CQ ratings matches the full-sample ranking, and Column 4 reports the number of groups. The full-sample ranking is replicated in every group of every CQ set except one

Table 4: Tests of Common Ranking

		Respondents		Demographic Groups		
Dimension	Kind	Fraction consistent with full-sample ranking (1)	Number of respondents (2)	Fraction consistent with full-sample ranking (3)	Number of groups (4)	p-value for Common Ranking (5)
Life Satisfaction	Vignette	0.37	10,100	1.00	32	1.00
Happiness	Vignette	0.48	5,063	1.00	30	1.00
Worthwhileness	Vignette	0.41	5,037	1.00	29	1.00
No Anxiety	Vignette	0.49	5,062	1.00	29	1.00
Anxiety	Vignette	0.51	5,038	1.00	30	1.00
Breathe	Vignette	0.49	10,100	1.00	32	1.00
Darkness	Visual	0.86	10,100	1.00	32	1.00
Shape	Visual	0.61	10,100	1.00	32	1.00
Confidence	Visual	0.62	10,100	1.00	32	1.00
Pain	Visual	0.77	10,100	1.00	32	1.00
Remember[a]	Vignette	0.79[a]	10,100	1.00[a]	32	1.00[a]
Life Satisfaction (S2)[b]	Vignette	0.39	2,503	1.00	20	1.00
Life Satisfaction 2 (S2)	Vignette	0.45	2,503	1.00	20	1.00
Height (S2)	Visual	0.75	2,503	1.00	20	1.00
Weight (S2)	Visual	0.60	2,503	0.95	20	0.84

[a] Because the Remember entries in Columns 1, 3, and 5 use only two CQs, they are not directly comparable to the three-CQ entries in other rows.
[b] S2 denotes Survey 2.
Notes: Demographic Groups (Columns 3–5): Among respondents who answered all CQs in a trio, five dummies are crossed (above median age, above med. income, gender, marital status, and college completion), resulting in $2^5 = 32$ cells. Cells < 50 respondents are dropped. Column 5: p-value for Column 3's ordering restriction, computed as a generalized likelihood ratio test (see Web Appendix C.5 for details).

Weight group out of twenty. Column 5 reports the p-value from a test of whether such failures could arise from sampling variation alone (Web Appendix C.5); no CQ set rejects at conventional levels.[20]

VII.D The Common Same-Rating Map and SR-CQ Concordance

To understand the remaining assumptions, it is helpful to define the *same-rating map* between SRs and CQs. Fixing a respondent i, it is a pairing between a CQ c and a situation ω such that the situation and CQ are rated the same when in that situation: $\rho_i^\star(\omega) = \rho_i(c|\omega)$. Assumption 2 states that the same-rating map is the same for every respondent.

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

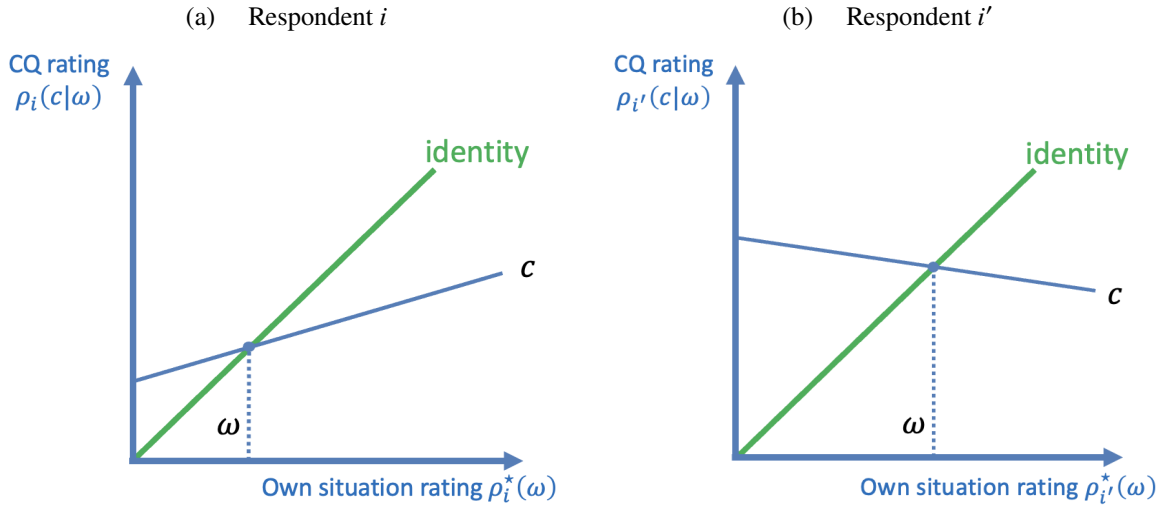
$$\rho_i^\star(\omega) = \rho_i(c|\omega) \iff \rho_{i'}^\star(\omega) = \rho_{i'}(c|\omega).$$

Graphically, as illustrated in Figure 7, Assumption 2 means that for each CQ, every respondent's CQ rating curve intersects the identity line at the point corresponding to the same ω. Given this assumption, we can write $\omega \leftrightarrow c$ for a same-rating pair without indexing by i. We discuss the

[20] The result is robust to a finer split. Crossing nine binary characteristics rather than five—adding nonwhite, children in the household, employment, and residence in the South—divides each CQ set's respondents into as many as 99 non-overlapping groups, each containing at least 30 respondents. Common Ranking is not rejected for any CQ set ($p \geq 0.71$).

interpretation and plausibility of Assumption 2 first for same-dimension CQs and then for other CQs.

Figure 7: Rating Curves Illustrating Assumption 2



For same-dimension CQs, one scenario in which Assumption 2 would be satisfied is what we call *self-CQ consistency*: every respondent would rate a CQ depicting their own situation in $\mathcal{S}$ the same as they rate themselves. This scenario may be plausible for dimensions where respondents can judge situations accurately, are not narcissistic, and do not have social desirability concerns that cause them to rate themselves (but not the CQ) higher. It would also be satisfied if respondents have rose-colored (or doom-colored) glasses, as long as their own-situation ratings and CQ ratings are equally distorted. Assumption 2, however, is more general than self-CQ consistency because it allows respondents to rate the CQ the same as they would rate themselves in a *different* situation, as long as it is a situation on the same iso-rating surface for all respondents—a condition we call *homogeneous self-CQ discrepancy*. For example, Assumption 2 would be satisfied if respondents rate the heights of the men in our Height CQs impartially but short people exaggerate when rating their own heights—as long as equally short people exaggerate equally.[21] Likewise, context effects on CQ ratings relative to SRs (e.g., from preceding questions) do not violate Assumption 2 as long as they are common across respondents; heterogeneous context effects would violate it. In principle, for CQs that are fully detailed (e.g., vignettes with no details left unspecified), Assumption 2 is testable by comparing respondents' SRs to their ratings of CQs that depict the respondents' own situations (Kapteyn et al., 2011, conduct a test in this spirit but with vignettes that are not fully

[21] Reference-group effects that are shared by respondents when rating CQs—for example, respondents rating males' and females' heights relative to average male and female height, respectively—would not cause Assumption 2 to be violated. In contrast, Assumption 2 would be violated by such reference-group effects for SRs (unless the *definition* of the dimension is "height relative to gender," as in our height SR; see Section I.D).

detailed).

For other-dimension CQs, *prima facie*, Assumption 2 may seem absurd, since the SR situation rated on one dimension and the CQ situation rated on another dimension—say, life satisfaction and darkness of a circle—are not comparable. However, since the assumption is about how respondents' *ratings* compare, it could be a reasonable approximation, depending on the psychological process underlying the reporting functions. For example, if every respondent's SR and CQ ratings are the same strictly monotonically increasing transformations of the percentile ranking of the respective situations, and if respondents share a common prior over the distributions, then Assumption 2 will hold. Data on the psychological process could provide indirect evidence on Assumption 2. In principle, there is a direct test, but it is survey-intensive and survey-design-intensive: design fully detailed same-dimension vignettes that satisfy the assumption, then test whether respondents' ratings of those vignettes and of the other-dimension CQs share a same-rating map.

Considering two same-rating pairs, $\omega \leftrightarrow c$ and $\omega' \leftrightarrow c'$, Assumption 3 states that the CQ reporting function ranks c' above c (in both situations ω and ω') exactly when the SR reporting function ranks ω' above ω.

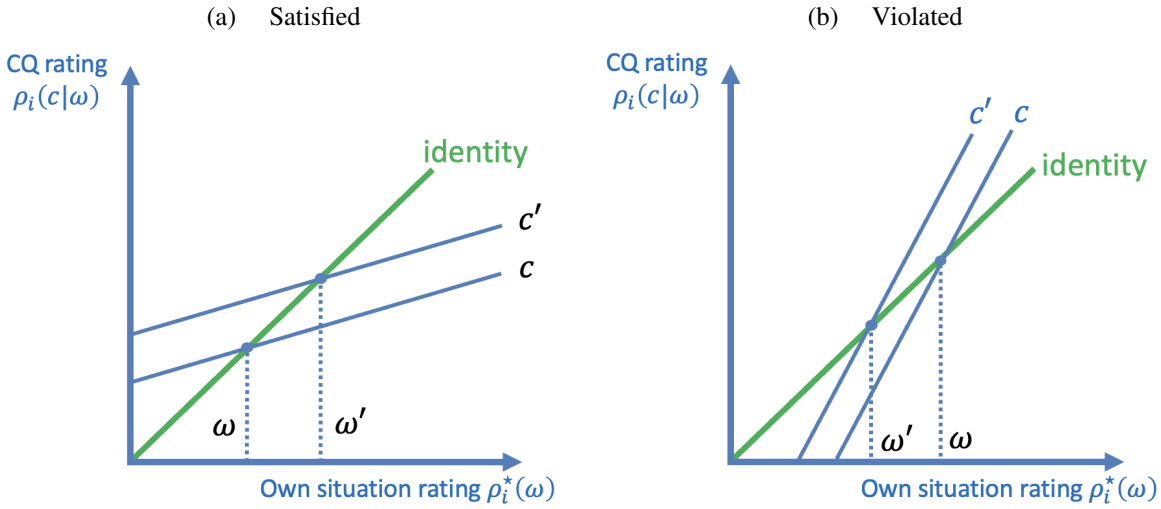
Assumption 3. ***(SR-CQ Concordance)*** *For all* $i \in \mathcal{I}$, $\omega, \omega' \in \mathcal{S}$, *and* $c, c' \in \mathcal{C}$, *with* $\rho_i^\star(\omega) = \rho_i(c|\omega)$ *and* $\rho_i^\star(\omega') = \rho_i(c'|\omega')$,

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

Panel (a) of Figure 8 depicts a case in which Assumption 3 is satisfied. The CQ curve for c' is shown as everywhere above c (because Assumption 1 implies that distinct CQ curves cannot touch). At the intersection between a CQ curve and the identity line, the point on the x-axis corresponds to the set of situations paired with the CQ by the same-rating map. In Panel (a), the situations mapped to c' rank higher than the situations mapped to c. Panel (b) shows how the assumption is violated by rating curves that cross the identity line from below. We view Assumption 3 as a mild regularity condition: if it were violated, then holding fixed the original same-rated CQ, a change in the respondent's situation would change the CQ rating as much as or more than the rating of their own situation.

Given priors on the signs of regression coefficients, Assumption 3 is testable for a group of respondents as long as Assumptions 1 and 2 hold. If it were violated by a slope-one rating curve, then subtracting the benchmark would remove all dependence of the SR on the situation, and our MM estimator (Section IV) would attenuate *all* coefficients to close to zero. If the slope exceeded one, the benchmark's dependence on the situation would overwhelm the SR's, and the MM estimator would in expectation reverse the signs of *all* coefficients relative to the naïve regression. In our application in Section V, the scale-use-corrected coefficients generally have the expected signs, so

Figure 8: Rating Curves Illustrating Assumption 3



violations of Assumption 3 do not drive our findings.

VII.E Correction for Scale-Use Heterogeneity Using CQs

Our central theoretical result states that Assumptions 1–3 are sufficient for respondents' CQ ratings, compared with their SRs, to reveal the common ranking of their SRs:

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* $\rho_i^\star(\omega) = \rho_i(c|\omega)$ *and* $\rho_{i'}^\star(\omega') = \rho_{i'}(c'|\omega')$,

$$\rho_i(c'|\omega) \gtreqless \rho_i(c|\omega) \iff \rho_i^\star(\omega') \gtreqless \rho_i^\star(\omega) \text{ and } \rho_{i'}^\star(\omega') \gtreqless \rho_{i'}^\star(\omega). \tag{P}$$

Proposition 1 says that when i's rating of c equals i's SR, and i''s rating of c' equals i''s SR, the within-respondent ordering of the CQs c and c' reveals the common ranking of the two respondents' situations. All proofs are in Web Appendix M, but the logic is as follows. We use Assumption 2's common-across-respondents same-rating map to map each respondent's situation to a rating-equivalent CQ. By Assumption 1, respondents agree on the ranking of these rating-equivalent CQs. Finally, by Assumption 3, the ranking of these CQs reveals the ranking of respondents' situations.

For the conclusion (P) of Proposition 1 to hold, Assumption 1 only needs to hold along the identity line, i.e., it only matters how a respondent rates a CQ when in the situation that same-rating-maps to it. In Web Appendix M, we prove that with this weakening, the assumptions are necessary and sufficient for the conclusion.

While (P) underlies using CQs to rank respondents' situations, in practice we do not need to find an SR-rating-equivalent CQ for each respondent. Instead, as in Section III.C, we use the CQ translation function to locate the SR on the CQ common scale. The SR common scale is not identified, because we never observe the reference individual's SR at another respondent i's situation, $\rho_{\bar{\imath}}^{\star}(\omega_i)$. Proposition 1, however, justifies using the CQ common scale in its place.

We *define* the CQ-common-scale rating of i's situation as the reference individual's rating of a CQ that same-rating-maps to it: $w_i^{\star} := \rho_{\bar{\imath}}(c|\bar{\omega})$ for any c with $\omega_i \leftrightarrow c$ (well-defined by our assumption on $\mathcal{C}$ and Assumptions 1 and 2). Applying Assumption 1 with $\bar{\imath}$ and $\bar{\omega}$ as the second respondent and situation, the left-hand side of (P) holds if and only if $w_{i'}^{\star} \gtreqless w_i^{\star}$. Thus, $w_i^{\star}$ correctly ranks situations. Moreover, applying i's CQ translation function to it returns i's SR:

$$r_i^{\star} = \tau_i(w_i^{\star} \mid \omega_i) \tag{5}$$

because $\tau_i(\cdot \mid \omega_i)$ maps each CQ's common-scale rating to i's rating of it, and i rates c the same as ω_i by construction. Equation (3) is the special case with a linear translation function and response error added. A different set of CQs defines a different CQ common scale (and hence different values of $w_i^{\star}$) but the same ranking of situations.

Since the SR common scale $\rho_{\bar{\imath}}^{\star}(\omega_i)$ was arbitrary (depending on the choice of reference individual), adopting the CQ common scale $w_i^{\star}$ for SRs is best understood as *choosing* an empirically implementable scale. Any strictly increasing transformation of the SR common scale is an equally valid common scale, and $w_i^{\star}$ is such a transformation. The two coincide if the reference individual's CQ ratings do not depend on their situation (the "slope = 0" case in Figure 6).

The Interpersonal Comparability Presupposition is an immediate consequence of Proposition 1.

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

The logic is: for any two situations, ω and ω', we identify CQs, c and c', that form same-rating pairs, $\omega \leftrightarrow c$ and $\omega' \leftrightarrow c'$. By Proposition 1, the common ranking of c and c' is also the common ranking of ω and ω'. Since the proof of Corollary 1 invokes only Assumption 2 and (P), its conclusion holds under the weakened version of Assumption 1. The contrapositive of Corollary 1 is that if the Interpersonal Comparability Presupposition does not hold, then for *any* set of CQs $\mathcal{C}$, at least one of the assumptions must be violated.

VII.F The Special Case of Linear Translation Functions

To directly justify our econometric model, we end this section by examining the special case in which the CQ translation function is linear. This special case arises from strengthening Assumption 1 (Common Ranking of CQs) to:

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* $\varpi \in \mathbb{R}$,

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

Assumption 1′ says that within any trio of CQs, the intervals between any respondent's ratings are proportional to the intervals between any other respondent's ratings. The presence of response error complicates testing this property at the individual level, so, as for Common Ranking, we compare means across groups, which by the law of large numbers is robust to arbitrary heteroskedasticity. For a trio of CQs, the test has two steps. First, for each respondent, subtract from the Medium CQ rating the value implied by the respondent's Low and High ratings under linearity, with interpolation weights computed from sample-mean ratings. Second, regress this residual, which is mean zero by construction, on demographics. If CQ translation functions are linear, demographics should have no predictive power.

Web Appendix C.4 reports p-values from this test and illustrates how the regression results can be used to assess the robustness of the MM scale-use correction to violation of Assumption 1′. The test rejects at the 5% level for the Breathe, Confidence, Pain, Life Satisfaction 2, and Weight CQs but not for the other dimensions. Such rejections could reflect a failure of Common Ranking or of linearity; the test does not distinguish them. But if the violations mattered substantively, results would be sensitive to dropping one of the three CQs, and for the CQ sets that reject, we find that doing so changes the scale-use correction to a regression coefficient by less than one standard error in almost all cases.

Proposition 2 says that Assumption 1′ implies Assumption 1 and is equivalent to a linear CQ translation function.

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

$$\rho_i(c|\omega) = a_i(\omega) + b_i(\omega)\rho_{\bar{\imath}}(c|\bar{\omega}). \tag{6}$$

Because a_i and b_i are evaluated at i's current situation ω_i, and $\rho_{\bar{\imath}}$ at $\bar{\omega}$, both a_i and b_i are constants and $\rho_{\bar{\imath}}$ is a function of c alone, so equation (6) is the linear CQ translation function of equation (2). In this case, two CQs suffice to identify the translation function and hence, by interpolation or extrapolation, every respondent's CQ-common-scale SR.[22]

[22]Moreover, Assumption 1′ implies that any linear combination of CQs asked on the survey can be treated as another CQ. Thus, under Assumption 1′, there is no need to introduce hypothetical CQs in the theoretical framework. Instead, Assumptions 2 and 3 could be re-stated as holding for CQs in the survey and any linear combination of those CQs.

If the union of two CQ sets satisfies Assumption 1′, then the two share the same CQ translation function and define the same CQ common scale. If they each satisfy Assumption 1′ but their union does not, the common scales they define will be different, related by a strictly increasing function.

VIII Practical Recommendations

Drawing on the theory of Section VII, the econometrics of Section IV, our empirical results, and our own experience, we offer practical recommendations. The first three concern the analysis of existing data; the rest concern the design of new surveys.

Choice of Dataset. When researchers have a choice between datasets, all else equal, we recommend choosing those that include CQs. Web Appendix Table L.7 lists datasets we are aware of that contain vignettes, along with details on vignette dimension, response scale, and sample size. We are not aware of existing visual CQ data in large-scale surveys other than ours, although some have been studied in online convenience samples (e.g., Benjamin et al., 2023; Hu et al., 2022). We hope that over time, researcher awareness (and demand) will result in more datasets containing CQs.

Choice of Dependent Variable. When several alternative survey measures could serve equally well as measures of the outcome (individually or jointly as an index, e.g., Benjamin et al. 2014), and all have same-dimension CQs, we recommend the narrower dimensions. Both the Interpersonal Comparability Presupposition and Common Ranking of CQs are more likely to fail for broad dimensions (Sections VII.B and VII.C).

Choice of CQ Dimension. When same-dimension CQs are available and there is no compelling reason to doubt that they satisfy Common Ranking of CQs (Assumption 1), we recommend using them. Same-dimension vignettes are most likely to fail Common Ranking when the SR dimension is broad because respondents may aggregate its subdimensions differently (Section VII.C). In that case, or when same-dimension CQs are unavailable, CQs in a closely related dimension are the best alternative: Section V finds that they yield nearly proportional coefficients, whereas CQs in distant dimensions do not.[23] Cross-dimension correction can also serve as a robustness check, with the caveat that the more distant the dimension, the more the correction will be off. We now detail the two sets of considerations that underlie these recommendations.

[23]In principle, a rejection of proportional coefficients could reflect a failure of the MM moment conditions or a nonlinear relationship between the common scales defined by the two CQ sets, even if Assumptions 1′, 2, and 3 hold for both. In practice, we believe substantively strong rejections most likely reflect a failure of Assumption 1′ or 2 for one or both CQ sets: our tests indicate that the moment conditions are good approximations; a first-order approximation to the relationship between two valid common scales should generally be accurate; and a violation of Assumption 3 alone would appear as a negative constant of proportionality rather than as non-proportionality. Regardless of the cause, if proportionality holds the two CQ sets yield equivalent coefficient ratios, and if it is rejected they do not.

Visual vs. Vignette CQ Considerations. The choice between same- and cross-dimension CQs is entangled with the choice between vignette and visual CQs because for most SR dimensions same-dimension CQs must be vignettes, while visual CQs are typically cross-dimension. Intuitively, vignettes are more subject to biases such as self-norming, vignette completion, and the empathy gap. An insight from our framework, however, is that such biases *per se* pose no problem for scale-use correction. What matters is the plausibility of Assumptions 1–3, and particularly of the first two, since the third seems benign.

The primary attraction of visual CQs, and hence of cross-dimension correction using them, is that Assumption 1 (Common Ranking of CQs) is especially well justified. For the Darkness-of-a-Circle CQs, for example, it is hard to imagine a violation once response error is accounted for. Consistent with this, a higher fraction of respondents share the full-sample ranking for our six visual trios than for our vignette trios (Table 4, Column 1). At the group level, however, every CQ dimension we examine replicates the full-sample ranking in nearly every group (Column 3), so we find no evidence that vignettes violate Common Ranking more than visual CQs do.

Same- vs. Cross-Dimension Considerations. The primary attraction of same-dimension CQs is that respondents' scale use for them is likely to be maximally similar to their scale use for the SR, so Assumption 2 (Common Same-Rating Map) is likely a better approximation than for other-dimension CQs. For a given individual, we expect scale use for a life satisfaction SR to be more similar to that for Life Satisfaction vignettes than for, say, Happiness vignettes, and less similar still to that for a visual CQ such as Darkness of a Circle. Our evidence bears this out. In Section V, replacing Life Satisfaction vignettes with Happiness vignettes leaves coefficient ratios nearly unchanged, while replacing them with visual CQs changes the ratios substantially. Moreover, in Section VI, we find that correcting the Height and Weight SRs with their own same-dimension CQs yields higher correlations with the objective-scale measures than did nearly every other-dimension alternative.

Together, these considerations favor same-dimension over cross-dimension correction. Best of all may be same-dimension *visual* CQs—e.g., faces for rating pain (Tomlinson et al., 2010)—when feasible, since then both Assumptions 1 and 2 are especially plausible.

Number of CQs and Sampling Design. The MM estimator is identified with as few as two CQs, but for researchers designing their own surveys or experiments, we recommend three, which permits testing Assumption 1′ (Linear Common Ranking of CQs) and estimating response-error variance.[24] Simulations in Web Appendix K, across a range of sample sizes, show that going from two to nine CQs reduces standard errors only slightly, so more than three may not be worth the survey time.

If asking CQs on the entire main survey sample is not feasible, researchers may consider asking

[24] In pilot data, a duo of CQs took about 3/4 of the survey time a trio takes, not 2/3, because there is a fixed cost to understanding and answering each ensemble of CQs.

them only on a subsample or even on an entirely separate sample. Precision will be reduced, but the MM estimator remains consistent if the auxiliary sample has the same joint distribution of scale use and regressors.

Survey Order of SRs and CQs. The order of survey questions can affect responses (e.g., Deaton and Stone, 2016; Hopkins and King, 2010). We recommend asking SRs before CQs, so that SR responses are more comparable to SRs from surveys without CQs.[25] To make Assumption 2 (Common Same-Rating Map) more likely to hold, we recommend asking CQs as soon as possible after SRs.

Design of CQs. When designing CQs, the key goals are (i) a large variance of mean ratings across the CQs in the set, to maximize signal relative to response error, while minimizing top- and bottom-coding; (ii) detailing as many subdimensions as is feasible; (iii) describing the same subdimensions across the CQs in a dimension; and (iv) encouraging all respondents to fill in missing details in the same way.[26]

Although we find similar results across vignette framings (see Web Appendix Table L.8), we recommend a redesigned version of the ***You (stranger)*** framing, spelled out in Web Appendix Table B.3. Among other refinements, it encourages respondents to fill in missing details with their best guess based on the details that *are* mentioned. This instruction is likely to yield better consistency across respondents than alternatives because many respondents will impute their best guess regardless of instructions.

CQ Diagnostics. Web Appendix C proposes diagnostics for assessing how well a set of CQs meets some of these goals and illustrates them on our data.

[25] Our recommendation is opposite to Hopkins and King's (2010). They find that SR responses have lower variance when asked after vignettes, which they interpret as reduced measurement error. However, the lower variance could instead be due to mechanisms that alter SR responses in ways researchers may wish to avoid. Our main worry is that the vignettes may change how respondents weight the subdimensions when answering the SR. This concern motivated us to ask SRs before vignette CQs, but because we had an incomplete understanding of this issue when we designed Survey 1, we asked visual CQs before SRs (see Section I.C). However, we suspect that preceding SRs by visual CQs (as opposed to vignette CQs) is generally less problematic because, in most cases, visual CQs' subdimensions are unrelated to the SR subdimensions.

[26] Here we illustrate how vignette completion can violate the assumptions when vignettes are not fully detailed. Suppose there are two subdimensions but vignette CQs detail only one, and respondents fill in the other with their own situation. Denoting all variables in common-scale units, suppose Respondent i's situation is $(70, 60)$ while i''s is $(60, 70)$, and both rate any situation (w_1, w_2) by $0.4w_1 + 0.6w_2$. Assumption 3 is violated: writing $(70, \cdot)$ for a vignette that specifies only the first subdimension, a respondent in situation $(70, 60)$ and one in $(70, 70)$ both rate $(70, \cdot)$ the same as their own situation, although the latter situation is rated higher. Assumption 1 is also violated: i ranks vignettes $(\cdot, 70) \succ (70, \cdot) \sim (\cdot, 60) \succ (60, \cdot)$, whereas i' ranks them $(70, \cdot) \succ (\cdot, 70) \sim (60, \cdot) \succ (\cdot, 60)$. If the survey includes only $(70, \cdot)$ and $(60, \cdot)$, however, this violation goes undetected, and the conclusion of Proposition 1 fails: i same-rates $(70, \cdot)$, i' same-rates $(60, \cdot)$, and both rank $(70, \cdot) \succ (60, \cdot)$, so we would conclude that i's situation is higher, although both respondents rate i''s situation higher (66 versus 64). Assumption 2 is not violated here, because all respondents fill in details the same way; it would be if, say, some respondents filled in their own situation while others filled in a typical one. Given data on a wide variety of CQs, the violation of Assumption 1 will generally be the easiest to detect.

Testing the Theoretical Assumptions for a CQ Type. Although as few as two CQs are sufficient for scale-use correction under Assumptions 1′, 2, and 3, *testing* the assumptions calls for a wide variety of CQs—CQs with independently varying subdimensions to test Assumptions 1 and 1′ and CQs describing a sample of respondent situations to test Assumption 2. (As noted above, Assumption 3 is likely to hold and relatively easy to assess if the other assumptions hold.)

IX Concluding Remarks

In this paper, we propose an econometric method and a theoretical framework for measuring and correcting for heterogeneity in the use of subjective scales. The method is simple and transparent. Although designed for continuous response scales, it performs similarly on scales with only a few categories (Web Appendices I and K.3). The framework generalizes and unifies existing approaches to scale-use correction, clarifies what assumptions are needed, and shows that scale-use correction is robust to many biases in how respondents answer CQs, such as self-norming, vignette completion, and the empathy gap. Because we translate SRs to a common scale—rather than trying to recover a respondent's unobserved situation ω_i by inverting their unobserved reporting function—we can rely on assumptions that are largely testable.[27] Empirically, scale-use correction can matter. In our life satisfaction regression, it leaves the most robust findings in the literature intact but triples the estimated cost of unemployment relative to income, and in our validation exercise it brings self-rated height and weight closer to their objective-scale counterparts.

The applications we have considered in this paper, such as life satisfaction regressions, are correlational. Can scale-use heterogeneity be ignored when the SR is the dependent variable in a causal study design, since it is balanced in expectation by randomization? In general, no, because the treatment may affect scale use (e.g., by affecting reference points). Moreover, estimates of treatment-effect heterogeneity may be biased if scale use differs across groups.

Our MM correction—subtracting the scale-use-only benchmark from the dependent variable—extends to any estimator in which the dependent variable enters only through its covariances with other variables, such as instruments. This includes many causal-effect estimators: instrumental variables, difference-in-differences, and regression discontinuity, though not their quantile-regression versions. One caveat applies even when treatment does not affect scale use: if the stretcher is correlated with the individual treatment effect, the corrected estimate weights individuals with larger stretchers more heavily in the average treatment effect.

We conclude with two further directions for future work. First, whereas our applications regress

[27] One might worry that scale-use correction does not achieve interpersonal comparability of SRs because respondents could differ in how they experience the same situation ω. However, this is what philosophers call the problem of qualia, and it applies equally to consumption experiences with an objective scale. For example, people might differ in their experience of consuming a banana, yet it remains useful to quantify consumption in terms of number of bananas.

a subjective-scale measure on objectively measured variables, researchers sometimes regress one subjective-scale measure on another (for example, on psychological measures of personality), so that scale use affects both sides. Scale-use heterogeneity then induces correlation between the regressor and the error term, an additional source of bias. Handling it would require modifying our methods and, if scale use differs across the two variables, CQs for each.

Second, whereas the applications we focus on are cross-sectional regressions, researchers may be interested in changes in scale use over time. For example, the observation that life satisfaction partially mean-reverts after major life events (e.g., Luhmann et al., 2012) has usually been interpreted as reflecting the actual dynamics of SWB, but it could instead, or in part, reflect a change over time in scale use (Sprangers and Schwartz, 1999; Rapkin and Schwartz, 2019; but see Lacey et al., 2008 for contrary evidence). An important direction for future work is to study hedonic adaptation after correcting for scale use.

References

Benjamin, Daniel J., Jakina Debnam Guzman, Marc Fleurbaey, Ori Heffetz, and Miles Kimball, "What do happiness data mean? Theory and survey evidence," *Journal of the European Economic Association*, 2023, *21* (6), 2377–2412.

_ , Kristen Cooper, Ori Heffetz, and Miles Kimball, "From happiness data to economic conclusions," *Annual Review of Economics*, 2024, *16* (1), 359–391.

_ , Ori Heffetz, Miles S. Kimball, and Nichole Szembrot, "Beyond happiness and satisfaction: Toward well-being indices based on stated preference," *American Economic Review*, 2014, *104* (9), 2698–2735.

Blanchflower, David G., Alex Bryson, and Xiaowei Xu, "The declining mental health of the young and the global disappearance of the unhappiness hump shape in age," *PLOS ONE*, 2025, *20* (8), e0327858.

Bond, Timothy N. and Kevin Lang, "The sad truth about happiness scales," *Journal of Political Economy*, 2019, *127* (4), 1629–1640.

Clark, Andrew E., Paul Frijters, and Michael A. Shields, "Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles," *Journal of Economic Literature*, 2008, *46* (1), 95–144.

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.

_ and Arthur A Stone, "Understanding context effects for a measure of life evaluation: How responses matter," *Oxford Economic Papers*, 2016, *68* (4), 861–870.

_ , Jane Fortson, and Robert Tortora, "Chapter 5: Life (evaluation), HIV/AIDS, and death in Africa," in Ed Diener, John F. Helliwell, and Daniel Kahneman, eds., *International Differences in Well-Being*, Oxford, UK: Oxford University Press, 2010, pp. 105–136.

Decancq, Koen and Mateo Seré, "Room for happiness? Cross-country differences in response scale use," 2025.

Di Tella, Rafael, Robert J. MacCulloch, and Andrew J. Oswald, "Preferences over inflation and unemployment: Evidence from surveys of happiness," *American Economic Review*, 2001, *91* (1), 335–341.

Frijters, Paul, Andrew E. Clark, Christian Krekel, and Richard Layard, "A happy choice: Wellbeing as the goal of government," *Behavioural Public Policy*, 2020, *4* (2), 126–165.

Giustinelli, Pamela, Charles F. Manski, and Francesca Molinari, "Tail and center rounding of probabilistic expectations in the Health and Retirement Study," *Journal of Econometrics*, 2022, *231* (1), 265–281.

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.

Helliwell, John F. and Christopher P. Barrington-Leigh, "Viewpoint: Measuring and understanding subjective well-being," *The Canadian Journal of Economics*, 2010, *43* (3), 729–753.

Hopkins, Daniel J and Gary King, "Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability," *Public opinion quarterly*, 2010, *74* (2), 201–222.

Hu, Mengyao, Sunghee Lee, Hongwei Xu, Roberto Melipillán, Jacqui Smith, and Arie Kapteyn, "Improving anchoring vignette methodology in health surveys with image vignettes," *Methoden, Daten, Analysen*, 2022, *16* (2), 273–314.

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 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.

Krueger, Alan B. and Arthur A. Stone, "Progress in measuring subjective well-being," *Science*, 2014, *346* (6205), 42–44.

Lacey, Heather P., Angela Fagerlin, George Loewenstein, Dylan M. Smith, Jason Riis, and Peter A. Ubel, "Are they really that happy? Exploring scale recalibration in estimates of well-being," *Health Psychology*, 2008, *27* (6), 669–675.

Layard, Richard and Jan-Emmanuel De Neve, *Wellbeing: Science and Policy*, Cambridge: Cambridge University Press, 2023.

Levinson, Arik, "Valuing public goods using happiness data: The case of air quality," *Journal of Public Economics*, 2012, *96* (9), 869–880.

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.

Luhmann, Maike, Wilhelm Hofmann, Michael Eid, and Richard E. Lucas, "Subjective well-being and adaptation to life events: A meta-analysis," *Journal of Personality and Social Psychology*, 2012, *102* (3), 592–615.

Márquez-Padilla, Fernanda and Jorge Álvarez, "Grading happiness: What grading systems tell us about cross-country wellbeing comparisons," *Economics Bulletin*, 2018, *38* (2), 1138–1155.

Oswald, Andrew J., "On the curvature of the reporting function from objective reality to subjective feelings," *Economics Letters*, 2008, *100* (3), 369–372.

Preston, Carolyn C. and Andrew M. Colman, "Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica*, 2000, *104* (1), 1–15.

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.

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.

Tomlinson, Deborah, Carl L. von Baeyer, Jennifer N. Stinson, and Lillian Sung, "A systematic review of faces scales for the self-report of pain intensity in children," *Pediatrics*, 2010, *126* (5), e1168–e1198.

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.