NBER WORKING PAPER SERIES

ADJUSTING FOR SCALE-USE HETEROGENEITY IN SELF-REPORTED WELL-BEING

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

Working Paper 31728
http://www.nber.org/papers/w31728

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023, Revised December 2023

We are grateful for NIH/NIA grants R01-AG065364 to the Hebrew~ University of Jerusalem, and R01-AG051903 to the University of California Los Angeles; to Matt Adler, Angus Deaton, Marc Fleurbaey, Arie Kapteyn, Laura Kubzansky, Richard Lucas, Arthur Stone, Louis Tay, conference participants at the Hebrew University Federmann Rationality Center's 31st Annual Retreat, the Dartmouth Workshop on Philosophy and Economics, the Normative Economics and Economic Policy webinar, the Los Angeles Experiments conference, and seminar participants at UCLA Anderson's Behavioral Decision Making Group Lab Meeting, Baylor University, George Mason, the University of Connecticut, the University of Nottingham, the University of Colorado Boulder, the Paris School of Economics, the Southampton Centre for Experimental Social Sciences, and Shandong University for helpful comments and discussion; and to Tal Asif, Yehonatan Caspi, Colby Chambers, 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, Pierre-Luc Vautrey, and Shira Zadik, for outstanding research assistance. This work utilized the Summit supercomputer and the Alpine high performance computing resource at the University of Colorado Boulder. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding bodies. 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 They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Adjusting for Scale-Use Heterogeneity in Self-Reported Well-Being
Daniel J. Benjamin, Kristen Cooper, Ori Heffetz, Miles S. Kimball, and Jiannan Zhou
NBER Working Paper No. 31728
September 2023, Revised December 2023
JEL No. C83,D60,D63,D90,D91,I14,I31

ABSTRACT

Analyses of self-reported-well-being (SWB) survey data may be confounded if people use response scales differently. We use calibration questions, designed to have the same objective answer across respondents, to measure dimensional (i.e., specific to an SWB dimension) and general (i.e., common across questions) scale-use heterogeneity. In a sample of ~3,350 MTurkers, we find substantial such heterogeneity that is correlated with demographics. We develop a theoretical framework and econometric approaches to quantify and adjust for this heterogeneity. We apply our new estimators in several standard SWB applications. Adjusting for general-scale-use heterogeneity changes results in some cases.

Daniel J. Benjamin
University of California Los Angeles
Anderson School of Management
and David Geffen School of Medicine
110 Westwood Plaza
Entrepreneurs Hall Suite C515
Los Angeles, CA 90095
and NBER
daniel.benjamin@anderson.ucla.edu

Kristen Cooper
Gordon College
229 Jenks
255 Grapevine Road
Wenham, MA 01984
kristen.cooper@gordon.edu

Ori Heffetz
S.C. Johnson Graduate School of Management
Cornell University
324 Sage Hall
Ithaca, NY 14853
and The Hebrew University of Jerusalem
and also NBER
oh33@cornell.edu

Miles S. Kimball
Economics Building
University of Colorado Boulder
261 UCB
#212
Boulder, CO 80302
and NBER
miles.kimball@colorado.edu

Jiannan Zhou
Shandong University
27 Shanda South Road
Jinan, Shandong 250100
China
jiannan.zhou@icloud.com

Large and growing literatures in economics and other social sciences use data on self-reported well-being (SWB), such as responses to survey questions about happiness or life satisfaction. Most of this work implicitly assumes that respondents use the survey response scale in the same way. If instead there is scale-use heterogeneity—if, for example, on a 0-100 happiness scale, one person's 70 corresponds to another person's 80—then such heterogeneity can frequently be a confound for conclusions researchers would like to draw. When analyzing panel data on SWB, researchers often estimate regressions with individual fixed effects, which adjust for mean differences in scale use. However, fixed effects do not address other scale-use heterogeneity, such as differences in how much of the scale individuals use or changes in scale use over time. Analyses of cross-sectional data typically make no adjustments for scale use.

Oswald (2008) and Kapteyn, Smith, and van Soest (2007, 2009) called attention to the problem of scale-use heterogeneity in SWB research. Kapteyn et al. proposed using the "anchoring vignette" approach of King, Murray, Salomon, and Tandon (2004) to adjust for the way people use the response scale when answering a particular SWB question. The idea is to have respondents rate the SWB of hypothetical individuals described in vignettes, and use these ratings to translate respondents' ratings of their own SWB onto a common scale. However, the anchoring vignette approach has only been developed for response scales with a small number of options (such as Likert scales). With a small number of response options, conclusions may be sensitive to untestable assumptions about the distribution of latent, continuous SWB (Bond and Lang, 2019). Moreover, Deaton (2011) and others have raised serious conceptual concerns about the assumption that vignette ratings are comparable across respondents, e.g., because respondents may differ in their empathy for the hypothetical individual or may fill in unspecified details about the vignette situation in ways correlated with the respondent's own situation.

A separate literature in psychology, marketing, and survey research—which has made little contact with SWB research[1]—studies people's "response styles" when answering survey questions in general (for reviews, see, e.g., van Vaerenbergh and Thomas, 2013; Weijters, Baumgartner, and Geuens, 2016). The most widely studied are acquiescence (giving high

[1] An exception is Stone, Schneider, Junghaenel, and Broderick (2019). In regressions of SWB on a cubic in age, they find that controlling for measures of response style changes the relationships between age and some SWB measures (pain and fatigue) but has a relatively minor effect on the relationship between age and other SWB measures (life satisfaction and health). 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 SWB responses, presumably because both reflect cultural norms of scale use.

ratings), disacquiescence (giving low ratings), extreme responding, and midpoint responding (the last two are related to an individual's variability in ratings). There are two main approaches to adjusting for response style. First, some methods use the survey questions themselves to assess scale use; for example, the simplest such method is to standardize responses to a set of survey questions at the individual level (for a more sophisticated such method, see Rossi, Gilula, and Allenby, 2001). This approach confounds scale-use heterogeneity with true response heterogeneity. Second, some methods measure scale use with an independent set of "control items," such as questions randomly sampled from different psychological batteries (e.g., Weijters, Geuens, and Schillewaert, 2010). This approach requires assuming that individuals' responses to the survey question of interest (say, SWB) are independent of their responses to the control items conditional on scale use. That assumption will rarely hold for broad SWB dimensions, such as life satisfaction, which are likely correlated with most psychological and economic variables that existing survey questions aim to measure. Thus, neither approach cleanly identifies scale-use heterogeneity.[2]

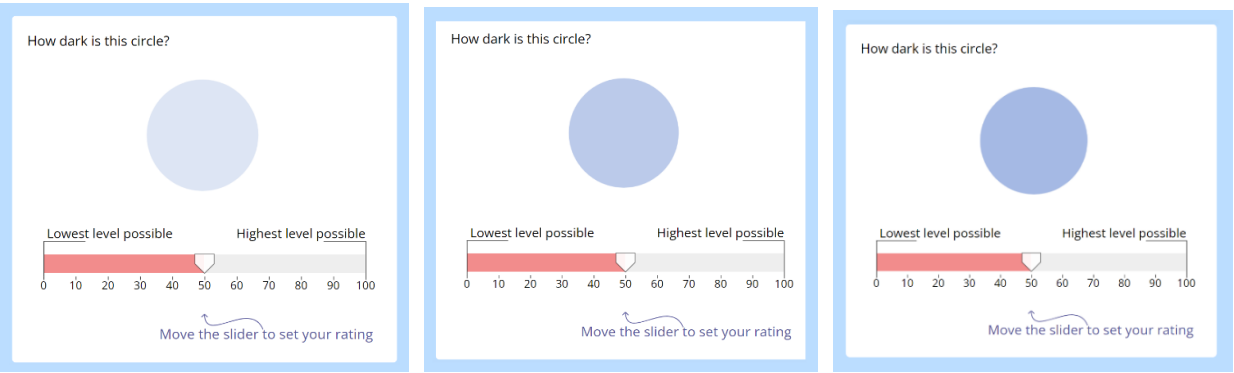
In this paper, we propose an approach to adjusting for scale-use heterogeneity that overcomes concerns with existing methods. Our approach applies when the response scale is continuous (e.g., a slider, which we use). It can be applied to either what we call *dimensional scale use*—the focus of the anchoring-vignette approach—or to what we call *general scale use*—the focus of the response-style literature. As in the anchoring-vignette approach, we identify scale use by what we call *calibration questions* (*CQs*): questions asked on a scale without physical units but designed so that the "true" answer should be the same across respondents. When adjusting for general scale use, the CQs do not need to be vignettes—for example, they can be visual, as in Figure 1 below—and hence can avoid the above concerns about vignettes. We propose several estimators that adjust for scale-use heterogeneity across a range of common analyses of SWB data. To illustrate our approach, we run a proof-of-concept survey with a diverse (but not representative) sample of MTurkers, with 3,358 respondents in our main sample

[2] Three other closely related literatures in psychology and survey research are response shifts, range-frequency theory, and psychophysics. In quality-of-life and self-reported-health research, "response shifts" refer to changes over time in how respondents map their own quality of life to survey responses (e.g., Sprangers and Schwartz, 1999; Rapkin and Schwartz, 2019). We focus on scale-use heterogeneity, which is considered one among several sources of response shifts (called "scale recalibration"). Range-frequency theory specifies how respondents' scale use for rating stimuli (e.g., the size of a box) depends both on the range of the stimuli and on their frequency distribution (e.g., Parducci, 1965). In this paper, we focus on measuring and correcting for scale-use heterogeneity and remain agnostic about its determinants. We discuss the relationship between our work and psychophysics in Section III.A.

after we restrict to high-quality respondents. In our data, we find that (i) there is substantial heterogeneity in scale use, (ii) much of that heterogeneity (roughly 3/5 of the variance) can be attributed to general scale use, and (iii) adjustment for general-scale-use heterogeneity changes the conclusions from some SWB analyses but not others—and our framework helps shed light on when and why the adjustment matters.

To anchor our theory and econometric methods introduced in later sections in a concrete empirical context, we first describe, in Section I, our survey sample and design. We ask a variety of SWB questions, including standard questions about happiness, life satisfaction, and other questions about well-being in particular life domains. To study how vignettes perform, we include them among our CQs, but we also include what we call *visual CQs* that avoid the critiques of vignettes by eliciting perceptual judgments of visual objects. For example, among our visual CQs is the following trio (asked on the same screen, where they are stacked vertically), designed to elicit a range of responses on the response scale:

Figure 1. Example of a Trio of Visual CQs



Under assumptions laid out in later sections, variation across respondents in their responses to CQs that ask about a single dimension (e.g., darkness of a circle) identifies dimensional scale-use heterogeneity. General scale-use heterogeneity is identified by variation in responses to CQs across many dimensions, on average (in a sense that we define precisely in Section III.B).

Section II documents that in our data, a respondent's mean response across multiple CQs is correlated with the respondent's mean response across multiple SWB questions. Similarly, the standard deviation of a respondent's responses to multiple CQs is correlated with the respondent's standard deviation of responses to multiple SWB questions. These findings suggest

that general scale-use heterogeneity accounts for part of the variation in SWB responses: some respondents report generally higher numbers than others, and some use more of the response scale. We also find that the mean and standard deviation of responses to CQs vary systematically across demographic groups, suggesting demographic variation in scale use.

We develop the theoretical framework and core assumptions underlying our approach to measuring and adjusting for scale-use heterogeneity in Section III. A central simplifying assumption for our approach is motivated by an empirical observation from our data: every respondent's answers to CQs are, after accounting for response errors, well approximated as a linear transformation of all other respondents' answers to CQs. This assumption implies that each respondent's (dimensional and) general scale use can be characterized by two parameters (that map onto the main "response styles" studied in prior work): a *shifter* and a *stretcher* relative to the average respondent.[3] Our econometric model flows from our conceptual model by allowing for response errors in answers to each survey question.

A theme of Section III is the presence of an inherent tradeoff between mitigating potential systematic biases in CQs and achieving a more complete adjustment for scale-use heterogeneity. For example, fully adjusting for (dimensional) scale use associated with life satisfaction necessitates relying exclusively on life satisfaction vignettes, but responses to such vignettes may be prone to systematic biases. In contrast, adjusting for general scale use can employ CQs from a diverse variety of dimensions, including non-vignette CQs such as our visual ones. A researcher worried about specific biases could therefore omit suspect CQs, thereby mitigating any specific systematic biases that may arise from any subset of CQ dimensions. However, if there exist scale-use components specific to life satisfaction, adjusting for general scale use will only partially correct for the scale-use heterogeneity relevant for life satisfaction.

Another lesson in Section III is that the amount of scale-use heterogeneity can vary with the average level of a survey question's responses, which we refer to throughout the paper as a question's *height*. If scale-use heterogeneity depends on an SWB question's height, then so does the appropriate scale-use adjustment. Our evidence from CQs suggests that such dependence indeed occurs. For example, we find that for CQs with sample mean responses near 70 on our 0-

[3] The point that the main response styles can be modeled by two parameters corresponding to a shifter and a stretcher has been previously highlighted by, e.g., Greenleaf (1992) and Rossi, Gilula, and Allenby (2001). This previous work (which focused on Likert-type response scales with few response options) did not examine the empirical adequacy of the implied linear relationship between respondents' scale use.

100 scale, mean responses within many demographic groups are near 70, but for CQs with sample mean responses near 55, mean responses within many of these demographic groups diverge from each other. Because scale-use heterogeneity depends on an SWB question's height, so does the appropriate scale-use adjustment.

In Section IV, using responses to all 18 CQs in our main sample, we quantify general-scale-use heterogeneity in our sample. We estimate each respondent's shifter and stretcher. Although noisy, these estimates can be used as dependent variables in regressions to study the demographic variation in the shifter and stretcher. Using maximum likelihood, we estimate the distributions of the shifter and stretcher parameters in our survey sample and find substantial heterogeneity in both. Our results also indicate that general scale-use heterogeneity within demographic groups is large relative to the variation across demographic groups.

In Section V, we describe the econometric methods we use to adjust for scale-use heterogeneity. We show that when the number of CQs is small—the empirically realistic case—an individual-level estimator for scale-use-adjusted SWB is biased and noisy. This motivates our approach of estimating four classes of moments of SWB that collectively cover the typical SWB applications: mean SWB, its covariance with a demographic variable, its variance, and its covariance with another SWB measure. For each of these moments of SWB, we develop several different estimators that have distinct advantages and disadvantages.

In Section VI, we use our survey data to demonstrate our methods of adjusting for general scale use. For each of the four moments of SWB, we derive closed-form expressions for the bias resulting from ignoring scale-use heterogeneity. These expressions show when, why, and how scale-use adjustment will matter. We use these theoretical results to shed light on why, in some applications using SWB data, we find that scale-use adjustment makes little difference, while in others, it matters much more. For example, we find that in a regression of life satisfaction on demographics, coefficients are barely affected, whereas several coefficients change in a regression of (lack of) anxiety on demographics. We show the reason: there is much less general-scale-use heterogeneity for responses near the mean for life satisfaction (i.e., that question's height) than near the mean for anxiety. Our adjustment matters a great deal for estimates of SWB inequality: we find that the cross-sectional variance in SWB is reduced by more than half after accounting for response-error variance and scale-use heterogeneity.

Section VII discusses four sets of additional results. First, we validate our approach to scale-use adjustment by showing that it strengthens the relationship between a respondent's self-reported physical height in objective units and on a 0-100 scale, and similarly for weight and several other objective-unit quantities. Second, we explore the relative importance of general-scale-use heterogeneity. On average across the full set of CQs we study, we estimate that roughly 3/5 of the total heterogeneity in scale use is due to general scale use. This result implies that general-scale-use adjustment addresses much, but not all, of the confounding in SWB analyses due to scale-use heterogeneity. Third, we estimate that roughly 50% of the variance across respondents in the shifter and nearly 90% in the stretcher are persistent when assessed with a median time gap of seven weeks apart. Finally, while our scale-use-adjustment methods are designed for continuous response scales, we explore whether general-scale-use heterogeneity also matters for SWB questions asked on the more common response scales that have just a few response options. We find that relationships between CQ responses and SWB responses elicited on these scales mirror those we find with our 0-100 scale, suggesting that the biases from scale-use heterogeneity that we identify also apply to analyses of SWB elicited using other scales.

We conclude in Section VIII. We briefly discuss issues relating to which CQs SWB researchers should ask when collecting data. We also discuss how SWB scale-use heterogeneity can matter in randomized experiments, as well as dynamically, where changes in scale use could be a confound for apparent evidence of hedonic adaptation (Loewenstein and Ubel, 2008).

We preregistered our sample exclusions and some analyses, using Open Science Framework (https://osf.io/qk5ta/?view_only=c2c05eb8d51d4c088c363db7a26d2f15). Web Appendix A.1 explains our preregistration and when and why we deviated from it.

I. Survey Design and Sample

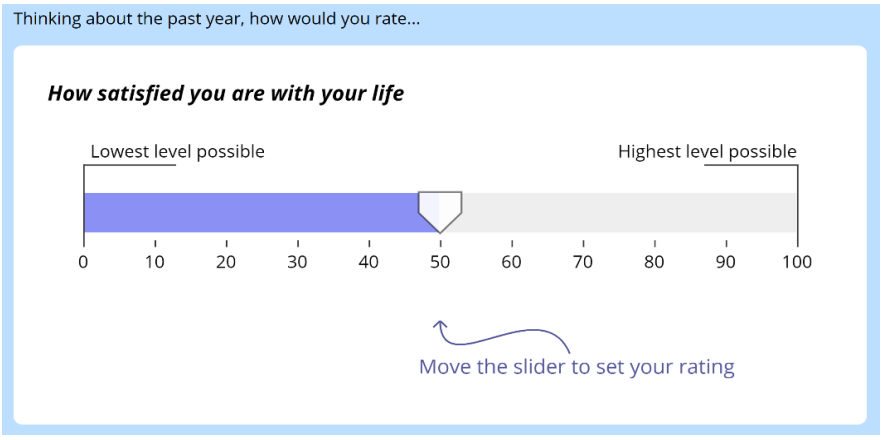
The two key types of questions in our baseline survey (henceforth, *Baseline*) are SWB questions, described in Section I.A, and calibration questions (*CQs*), described in Section I.B. The survey flow is: (i) consent form; (ii) basic demographic questions (age, gender, household income, ZIP code); (iii) instructions; (iv) SWB and stated-preference questions (the latter are not analyzed in this paper) and (v) CQs ((iv) and (v) in randomized order); (vi) additional demographic, behavioral, and psychological questions; and (vii) exit questions about how the respondent approached the survey. Web Appendix A.2 contains details and screenshots.

We also conducted a follow-up survey, which we call *Bottomless* because it contains many more CQs and SWB questions than Baseline. We report a few analyses based on Bottomless, such as when we need the large number of questions or second-time responses to questions asked on Baseline (to estimate measurement error), but we relegate the description of it to Web Appendix A.4; in this section we focus on describing Baseline.

I.A. Self-Reported Well-Being (SWB) Questions

Our SWB questions elicit respondents' ratings of various dimensions of life, over the past year, using a slider. Figure 2 shows an example.

Figure 2. Example of a Self-Reported Well-Being (SWB) Question



Note: Screenshot of SWB question for the dimension of well-being ***How satisfied you are with your life***.

Response options are integers from 0 (labeled "Lowest level possible") to 100 (labeled "Highest level possible"). The default slider position is at 50. To give a rating, the respondent moves the slider, and then clicks "Confirm Rating." To prevent lazy default responses, this button appears only after the slider has been moved.

The quasi-continuous 0-100 integer scale makes conclusions less susceptible to untestable assumptions about a latent variable that are needed when there are only a few response categories (Bond and Lang, 2019).[4] We specify the timeframe of the past year to reduce

[4] 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

heterogeneity in interpretation of the question across respondents (Benjamin, Debnam Guzman, Fleurbaey, Heffetz, and Kimball, 2023). The endpoint labels "Lowest/Highest level possible" result in response options that can be used for both SWB questions and CQs, and are meant to sound extreme in order to reduce potential top- and bottom-coding.

The Baseline survey includes 33 SWB questions in randomized order. This paper mainly focuses on four SWB questions based on the U.K.'s Office for National Statistics Opinions Survey (ONS, 2011). These four questions are described in Table 1. For the full list of questions in Baseline, see Web Appendix A.2.ii.

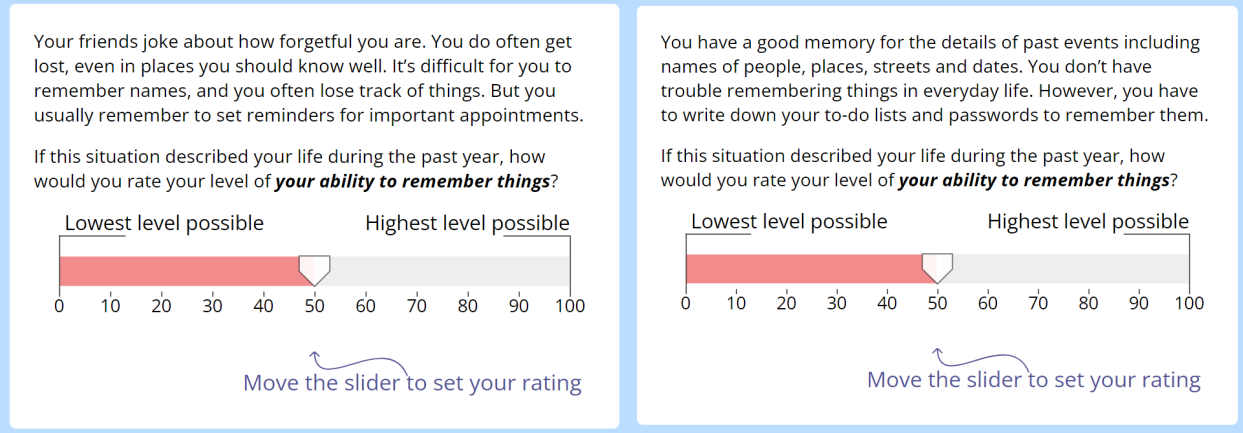
I.B. Calibration Questions (CQs)

Our Baseline survey has 18 CQs—9 visuals and 9 vignettes—which come in trios: groups of three similar CQs, presented together on the same screen. In addition to the trio of CQs shown in the Introduction ("How dark is this circle?"), the other Baseline visual CQ trios ask respondents to rate the curviness of line segments and the size of a region in a fictional continent. In designing the visual CQs, we avoided eliciting ratings for visual properties with a natural physical scale (e.g., the percentage of a circle that is shaded), with a strong good/bad valence, or that could not be represented easily with visual stimuli. We also tried to minimize the difficulty of answering CQs.

In a vignette CQ, the respondent is presented with a short description of a situation in "your" life and asked to rate some dimension of "your" well-being. For example, here are the "low" and "high" vignette CQs in the trio on ability to remember things:

Figure 3: Example of Two Vignette Calibration Questions (in a Trio)

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 latent SWB from discrete response categories. In practice, however, many respondents likely round their answers. Future research should account for such rounding, as in Giustinelli, Manski, and Molinari (2022).



Notes: Screenshots of vignette CQs for the dimension of well-being ***Your ability to remember things***. Shown on the left and right are the low and high CQs in the trio, respectively (a third, middle CQ is omitted from the figure). Respondents were instructed to "Imagine everything in your life is the same as it is now, except for the details described in each situation below."

In the other two vignette trios, we ask the respondents to rate "access to information" and "living environment not being spoiled by crime and violence." Only this last trio matches a dimension of well-being that we ask about in an SWB question in Baseline.

For both visual and vignette CQs, we designed each trio to have a low, medium, and high level in the dimension being rated. We used trios to guarantee power for identifying scale-use parameters and response-error variances across all our planned analyses. A trio of CQs is also cognitively simpler than three distinct CQs, enabling respondents to answer more CQs in any given amount of survey time. To minimize top- or bottom-coding, in addition to using the same extreme endpoint labels as for SWB ratings, we designed the visual and vignette stimuli for CQs deliberately to avoid extreme true levels. In Web Appendix A.2.iv, we detail our algorithms for constructing CQ trios and report some statistics on how successful we were at making top- and bottom-coding of CQs rare.

Baseline has three independent, uniform randomizations of the order of CQs: visuals or vignettes first; order of trios within type; and question order within trio. Also, within the survey, with equal probabilities, the CQs are asked before or after the SWB questions.

I.C. Survey Sample

Using Amazon's Mechanical Turk (MTurk) platform, we collected data between June 13 and December 7, 2022. We restricted eligibility for all surveys to MTurk workers located in the United States (as identified by MTurk), with a HIT approval rating of at least 95%, and at least

100 previously approved HITs. MTurk restricts all participation to workers who are at least 18 years of age. Respondents were recruited to Baseline with a HIT titled "Academic survey about what is important in life," after passing an initial prescreening survey (see Web Appendix A.3.ii). Respondents were compensated $4.50 for completing Baseline. The median completion time for those who passed our stringent quality control checks (see Web Appendix A.3.i) was 36 minutes.

In total, 5,970 respondents completed Baseline. Table 2 summarizes the demographics among those who answered the demographic questions used in our main applications (5,466 respondents) and among those who additionally passed quality control (3,358 respondents). The latter are our main analysis sample. Though similar to the U.S. population (according to the 2020 Census) on most demographics, as in other MTurk studies, our sample differs from the U.S. population in several ways: respondents are more likely to have completed college, be younger than 50, and be unemployed (other than MTurk); and less likely to have annual household income above $120,000 and to be Black or Hispanic/Latino.

II. Evidence of General Scale Use

Figure 4 provides descriptive evidence about general scale use in our main analysis sample. Using the 9 visual CQs and 33 SWB questions in Baseline, Panels A and B show correlations, at the individual level, both between a respondent's mean CQ and mean SWB ratings (corr. = 0.12, SE = 0.02) and between the standard deviations of a respondent's CQ and SWB ratings (0.12, SE = 0.02). These correlations are weak but statistically meaningful, and their magnitude is attenuated by true variation in SWB. Since the visual CQs are unrelated to SWB, we view the correlations as evidence that the SWB responses are influenced by general scale use.

Panels C and D show the analogous correlations using the 9 vignette CQs instead of the visual CQs: 0.39 (SE = 0.02) for the means and 0.46 (SE = 0.02) for the standard deviations. These correlations are much stronger than those for the visual CQs. That may be because scale-use tendencies for the SWB dimensions elicited by the SWB questions are more correlated with scale-use tendencies for the SWB dimensions asked about in the vignette CQs. Alternatively or additionally, the stronger correlations could reflect systematic biases in the vignette CQs. For example, vignettes cannot specify all aspects of a person's situation, and if respondents fill in the blank for unspecified details by projecting their own situation (as we in fact instructed them to

do; see Figure 3's note and the discussion in Web Appendix A.4.i), it would spuriously inflate the relationship between the vignette CQ and SWB responses. We discuss this and other potential biases in Section III.

Pooling the visual and vignette CQs, Table 3 shows that respondents' mean and standard deviation of the 18 CQ responses are predicted by demographics including age, income, employment, marital status, education, and religiousness (at a false discovery rate threshold of 1%). These results suggest that cross-demographic differences in scale use could confound studies of cross-demographic differences in SWB ratings.[5]

Moving beyond descriptive evidence, using CQs to measure, quantify, and correct for scale-use heterogeneity requires additional structure. In the next section, we develop a theoretical framework that clarifies the assumptions under which CQs can be used for these purposes.

III. Theoretical Framework

In this section, we introduce the assumptions underlying our theoretical framework, provide relevant empirical evidence from Baseline, and introduce our econometric model.

III.A. Model of Scale-Use Differences

Our model aims to capture differences in how people use the response scale when answering survey questions, including questions about, e.g., happiness, life satisfaction, or overall health. When researchers study such SWB questions, they assume that the questions ask about a quantity—a person's level on some dimension of well-being—that in principle has an ordinal scale, even if (given current science and technology) it cannot be reported in physical or objective units. Accordingly, our model assumes that people agree on the ordering of the underlying states, even though the self-reports are not interpersonally comparable.

We index individuals by i and survey questions by q. A survey question q may be either a CQ, indexed by c, or an SWB question, indexed by s. Every survey question q is an element of a set of questions, which we call a *dimension* and denote $d(q)$. A dimension is determined by features of the question; for example, one dimension is all questions that ask about darkness of a

[5] Our results are difficult to compare with existing evidence on correlations between response styles and demographics because few consistent patterns emerge in the literature (e.g., Van Vaerenbergh and Thomas, 2013), presumably because of differences in samples and control variables.

circle, and another is those that ask about overall health. All questions in a dimension share a state space $\Omega_{d(q)}$, which is the set of possible physical states corresponding to an answer to $q \in d(q)$. Individual i's state for question q is denoted $\omega_{iq} \in \Omega_{d(q)}$. For example, for the darkness-of-a-circle CQ, each state would be i's perceived shade of gray, whereas for an SWB question about overall health, each state would be a vector of components of i's health.

We assume that for every survey question, the set of response-scale options is $\mathbb{R}$. In our survey, we use a 0–100 integer scale, but we think of it as approximating the continuous interval $[0, 100]$, with responses outside the interval top-coded to 100 or bottom-coded to 0 (we find little top/bottom-coding in our data; see Web Appendix F.3).

Following Oswald (2008), we refer to the mapping from any state to how respondent i would rate it in response to survey question q as i's *reporting function*, $r_{iq}: \mathbb{R} \to \mathbb{R}$. An individual's reporting function may be shaped by factors such as culture, personality, and experiences, and it will depend on the labeling of response-scale options. Psychophysics, a subfield of psychology, is concerned with identifying reporting functions when the survey question asks about the individual's perception of some objectively measurable physical stimulus, e.g., loudness of a sound (for a review, see Birnbaum, 1994). Oswald (2008) highlights that SWB scale-use differences could be corrected-for by inverting $r_{iq}(\cdot)$ if the reporting function were known. However, he also points out that reporting functions for SWB cannot be identified because, unlike in psychophysics contexts, the underlying state ω_{iq} (here, "objective well-being") cannot at present be measured.

Assumption 1 formalizes the statement that individuals order the states in the same way, but individuals may differ in their use of the response scale.

Assumption 1: Common Monotonicity. Given survey question q, $r_{iq}(\cdot)$ is a strictly increasing transformation of $r_{i'q}(\cdot)$ for every pair of individuals i and i'.

Assumption 1 may fail for a broad and highly subjective SWB question, such as life satisfaction, for which respondents may differ in how they weight aspects of well-being when aggregating

them into life satisfaction. It is more plausible for a narrower and more objective SWB dimension, such as your ability to remember things.[6]

Instead of basing an approach to scale-use correction on recovering the unobservable state ω_{iq}, Assumption 1 enables us to base it on translating observable reports across individuals. We pick an arbitrary person, i^*, as the *reference individual* and treat i^*'s reporting function as the measurement scale to be used for interpersonal comparisons. Define the *translation function* $\tau_{i^*\to i,q}\colon \mathbb{R} \to \mathbb{R}$ as the mapping between i^*'s scale use and i's scale use: $\tau_{i^*\to i,q} \equiv r_{iq} \circ r_{i^*q}^{-1}$ (where in case of ties, $r_{i^*q}^{-1}$ can be defined to map to any state that is ranked equally).

The translation function cannot be identified empirically from SWB questions because individuals differ in both their reporting functions and their states. That is the fundamental identification problem of scale use. To solve it, we use *calibration questions (CQs)*, defined as survey questions for which the state is constant across individuals.

Assumption 2: Calibration Questions Identify a State. For each CQ c, the state ω_c about which it elicits a rating is the same across individuals.

Since CQs are designed by researchers, satisfying Assumption 2 should be a primary design consideration. For vignette CQs, one way Assumption 2 may fail is if respondents differ in how they mentally fill in the blanks about unstated aspects of the vignette subject's situation—for example, assuming that these aspects are like their own.[7] Alternatively, as Deaton (2011) argues, survey respondents whose own situation is different from the subject's may be unable to fully imagine the subject's stated situation; since respondents' misperceptions of the subject's situation are likely to differ from each other, the respondents will end up rating different states. Assumption 2 is most compelling for visual CQs, such as the darkness-of-the-circle CQ shown in

[6] Later in this paper we apply our scale-use-correction approach *as if* Assumption 1 (and subsequent assumptions) holds for all the SWB questions we study. We note, however, that the need to satisfy Assumption 1 is one reason to favor measuring narrower SWB dimensions and aggregating over them (as in Benjamin, Heffetz, Kimball and Szembrot, 2014; Benjamin, Cooper, Heffetz, and Kimball, 2017), rather than the standard approach of aiming to measure well-being with a single, broad SWB question.

[7] In our Baseline vignette CQs, we instructed our respondents to do so (see Figure 3's note), but in our Bottomless survey, we asked otherwise similar CQs that did not contain this instruction and instead asked respondents to "rate situations in other people's lives." Responses to these vignette CQs exhibit nearly identical correlations with SWB questions as responses to our Baseline vignette CQs (Web Appendix A.4.i). We interpret this finding as suggesting that the extent of "fill-in-the-blank" violations of Assumption 2 is similar whether or not participants are explicitly instructed to assume that unstated aspects of the vignette subject's situation are like their own.

the Introduction, that ask about an objective quantity that is fully conveyed by the question. Even then, people may differ in visual perception or may view the question on screens that differ in brightness. Once we add response errors below, which allow for some differences across people, our Assumption 4 below will require that these perception differences are unsystematic.

Under Assumptions 1 and 2, we can estimate translation functions by plotting respondents' ratings of CQs against each other. Figure 5 shows estimated average translation functions across groups of participants, defined by demographic categories or median splits on socioeconomic variables (see Web Appendix Figure B.3 for individual-level translation functions). Each point in the figure corresponds to one of the 18 CQs in Baseline and shows the relationship between the mean rating of that CQ across groups. We discuss the linearity apparent in Figure 5 below, but for now, we highlight four more basic observations.

First, a positive relationship is evident in every case. This observation is consistent with Assumptions 1 and 2, which jointly imply that groups of respondents should share a common ranking of the states in CQs.[8] Second, many of the points are not on the identity line. Under Assumptions 1 and 2, deviations from the identity line imply that there are scale-use differences on average across groups. Third, the scale-use differences can depend on the CQ's *height*, i.e., the level of the mean response on the 0-100 scale. For example, on average compared to respondents who are not employed full-time, respondents who are employed full-time give higher responses to CQs near the bottom but not the top. This third observation implies that the right scale-use correction for an SWB question will depend on its height. Finally, the relationship across groups of respondents in their responses to CQs is essentially the same relationship regardless of which CQs are used. This, together with the evidence from Section II that the mean and standard deviations of individuals' responses to CQs are predictive of their respective mean and standard deviation of responses to SWB questions, suggests that general scale-use tendencies may comprise much of the relevant scale-use variation across individuals—an observation that we confirm formally in Section VII.B.

[8] Assumptions 1 and 2 also jointly imply that *individual* respondents should share a common ranking of the states in CQs. However, this prediction cannot be tested empirically if respondents' ratings contain response errors (see Section III.D below) without strong assumptions on the response errors; in our data, respondents' rankings of the CQs in a trio agree with ours in 68.4% of cases. Response errors do not confound the empirical test of the prediction of a positive relationship across groups of respondents if the number of respondents in each group is large.

III.B. Dimensional Scale Use, General Scale Use, and Identification of Scale Use

The translation function $\tau_{i^*\to i,q}$ is defined separately for each survey question q. Using CQs to measure and correct for scale-use heterogeneity requires some assumption about how scale use on SWB questions relates to scale use on CQs. We now lay out such an assumption.

We refer to scale use that applies to all questions in some category $\mathcal{D}$ of dimensions (e.g., questions about feelings) as *categorical scale use*. While intermediate cases may be of interest, we focus here on two extreme special cases: *dimensional scale use* is when $\mathcal{D}$ is a single dimension (e.g., anxiety), whereas *general scale use* is when $\mathcal{D}$ is the entire population of dimensions under consideration (in our context, including visual CQs).

Define $\mathcal{S}$ as the set of SWB questions with dimensions in $\mathcal{D}$. We are interested in measuring scale use for some SWB question $s \in \mathcal{S}$. As the common scale across individuals, we use what i^* would report for any possible state $\omega_{is} \in \Omega_{d(s)}$; we denote i^*'s report of i's state by $h \equiv r_{i^*s}(\omega_{is})$. Individual i's scale use for SWB question s can be defined as the function of h that translates i^*'s report to i's report: $\tau_{i^*\to i,s}(h)$. To identify $\tau_{i^*\to i,s}(h)$ at a particular h, we imagine constructing a large number of CQs similar to the CQs we will study empirically with dimensions in $\mathcal{D}$ that have height h, i.e., the reference individual i^*'s answer to each of the CQs would be h. We refer to this set of CQs as $\mathcal{C}(s,h)$, and we think of the CQs we will study empirically as a random sample from this set. We measure individual i's scale use at height h by the individual's mean response to the CQs in $\mathcal{C}(s,h)$: $E_{c\in\mathcal{C}(s,h)}\left[\tau_{i^*\to i,c}(h)\right]$. This quantity is estimable using the CQs we study empirically. To measure (and correct for) categorical scale use, the identifying assumption is that for any given h, scale use for an SWB question in $\mathcal{S}$ that has height h, which is unobservable, is the same as average scale use for CQs in $\mathcal{C}(s,h)$.

Assumption 3: Generalized Response Consistency. For any $s \in \mathcal{S}$ and any h,

$$\tau_{i^*\to i,s}(h) = E_{c\in\mathcal{C}(s,h)}\left[\tau_{i^*\to i,c}(h)\right].$$

This assumption generalizes the "response consistency" assumption of the anchoring-vignette approach (King et al., 2004; Kapteyn et al., 2009). That approach aims to adjust for dimensional scale use, asking vignette CQs related to the SWB question of interest. Assumption 3 specializes

to response consistency in the case of dimensional scale use with vignette CQs.[9] In that case, Deaton (2011) points out a testable prediction: people should rate their own SWB the same as they rate the SWB of a vignette subject whose situation is the same as their own. Methods of adjusting for "response styles" can be understood as aiming to correct SWB responses for general scale use, using CQs that do not have corresponding SWB questions.

The main advantage of adjusting for general scale use is that it can use CQs for which Assumptions 1 and 2 are more compelling, such as visual CQs and the vignette CQs corresponding to narrower and more objective dimensions of SWB that we use. The disadvantage is that general scale use may not coincide with the relevant scale use for the SWB question(s) of interest—that is, Assumption 3 may fail.

We note that under Assumption 3, scale-use correction does not require assuming away reverse causation from SWB to scale use or a third variable affecting both. For example, suppose being in a good mood causes people to report higher numbers in response to any survey question but also genuinely increases SWB. Under Assumption 3, the individual's increased SWB is still correctly detected by the *difference* between their response to the SWB question and the CQs.

Hereafter, we assume that the dimensions represented by all of our SWB questions and CQs are in $\mathcal{D}$. That is, we focus on general scale use, and our data analysis uses all the CQs we measured; the discussion would be analogous for dimensional scale use, with the CQs restricted to the relevant dimension.

III.C. Linear Approximation to Translation Functions

If individual i's response to an SWB question is $\tau_{i^*\to i,s}(h)$, a researcher could, in principle, translate this response back to h (that is, to i^*'s scale) using only CQs of height h. In practice, however, there may be few or no such CQs, and the CQ data can be used much more efficiently if we make some assumption about how translation at a given height is related to translation at other heights. Although we could proceed with a more flexible function, we

[9] King et al. (2004) state their response consistency assumption as: "each individual uses the response categories for a particular survey question in the same way when providing a self-assessment as when assessing each of the hypothetical people in the vignettes." Our Assumptions 1, 2 and 4, taken together, essentially coincide with King et al.'s (2004) other key assumption, called vignette equivalence: "the level of the variable represented in any one vignette is perceived by all respondents in the same way and on the same unidimensional scale, apart from random measurement error." Our formulation distinguishes and clarifies what is being assumed about the reporting functions, CQs, and response errors (discussed in Section III.D below).

approximate the translation function as linear, with positive slope; that is, we strengthen Assumption 1 to:

Assumption 1′: Common Linearity. Given survey question q, $r_{iq}(\cdot)$ is a positive affine transformation of $r_{i'q}(\cdot)$ for every pair of individuals i and i'.

We assume the translation function is linear for three reasons. First, it is a reasonable approximation empirically. As can be seen in Figure 5, the translation functions across groups are close to linear. Individual-level translation functions calculated from our Baseline data are noisy, but when calculated from our Bottomless data, which contain many more CQs, they are much less noisy and are well approximated as linear (Web Appendix Figure B.2). Second, linearity is a helpful simplifying assumption, making the econometrics less complex and facilitating interpretation of model parameters. Third, linear translation functions are attractive theoretically because they are exactly what is needed for the conclusions of typical SWB applications to be invariant to the choice of reference individual; for example, under this assumption, ratios of coefficients from regressions with SWB as the dependent variable do not depend on the reference individual. Note that linear translation functions (from one respondent's report to another's) do *not* imply linear reporting functions (from true state to one's report).[10]

Assuming linear translation functions, individuals can be characterized by how they shift and stretch their use of the scale relative to others when responding to a given question q. To parameterize the translation functions in a convenient way, let $r_{iq} = r_{iq}(\omega_{iq})$ denote the report individual i would make for state ω_{iq}. Define $w_{iq} \equiv r_{i^*q}(\omega_{iq})$ as what i^* would report for the same state. When q is an SWB question s, we refer to w_{is} as individual i's *common-scale SWB*.[11] Individual i's translation function can then be written as:

[10] Assumption 1′ will imply that scale-use-corrected SWB is measured on an interval scale. It therefore addresses the problems that arise when SWB is treated as merely ordinal, such as the signs of regression coefficients being sensitive to which monotonic transformation is applied to SWB (Schröder and Yitzhaki, 2017). SWB is measured on an ordinal scale under other approaches to scale-use correction, such as the anchoring-vignette approach and (if it were feasible) inverting the reporting function (see Section III.A). Furthermore, as noted previously, our use of a continuous response scale rather than discrete response options eliminates the need for untestable assumptions about the distribution of latent SWB. Sensitivity to such assumptions is another problem that arises in typical SWB applications (Bond and Lang, 2019).

[11] We call w_{is} "common-scale SWB" to emphasize its tight relationship to empirical SWB data. We avoid calling it "true SWB" because that terminology would connote a normative status. Taking a normative stand would require

(1) $$\tau_{i^*\to i,q}(w_{iq}) = r_{iq} = a_{iq} + \beta_{iq} w_{iq},$$

where $a_{iq} \in \mathbb{R}$ and $\beta_{iq} \in \mathbb{R}_{++}$.

Linear translation functions imply that dimensional and general scale use are also linear. Consider general scale use. Define individual i's (general scale use) *gross shifter* and *stretcher*, respectively, by the mean a_{ic} and β_{ic} across the population of CQs: $a_i \equiv E_i(a_{ic})$ and $\beta_i \equiv E_i(\beta_{ic})$. By Assumption 3 above, these same parameters also apply to average scale use for SWB questions, so $a_i = E_i(a_{is})$ and $\beta_i = E_i(\beta_{is})$. For any survey question q, we can therefore decompose equation (1) as follows:

(2) $$r_{iq} = \underbrace{a_i + \beta_i w_{iq}}_{\text{general scale use}} + \underbrace{\left[\left(a_{iq} - a_i\right) + \left(\beta_{iq} - \beta_i\right) w_{iq}\right]}_{\text{question-specific scale use}},$$

where the first term corresponds to general scale use at height w_{iq}, and the second term corresponds to the deviation from general scale use for question q.

III.D. Econometric Model

To facilitate data analysis and interpretation of results, we re-parameterize the model in two ways. First, rather than using a specific individual as the "reference individual," we use the conditional population mean (whose scale use will be estimated more precisely): for any state ω_{iq}, $w_{iq} \equiv E\left[r_{iq}(\omega_{iq}) | \omega_{iq}\right]$, where from now on all expectations are taken over individuals, unless the expectation is conditioned on i. In words, w_{iq} is defined to be what the population mean report would be if everyone experienced i's state ω_{iq}. It follows from this choice of reference individual that $E(a_{iq}) = 0$ and $E(\beta_{iq}) = 1$ for all q, and therefore $E(a_i) = 0$ and $E(\beta_i) = 1$.[12]

additional philosophical assumptions. For some recent evidence on the relationship between SWB and standard normative concepts in economics, see Benjamin, Debnam Guzman, Fleurbaey, Heffetz, and Kimball (2023).

[12] We highlight that the mean of the shifter is zero by normalization only for the population overall. The mean of the shifter can be different between subsamples, for example, for men and women.

Second, we re-center the shifter parameter, both to give it a more natural interpretation and to allow for interpreting the shifter and stretcher as distinct parameters. The gross shifter a_i is likely to be correlated with the stretcher β_i (because individuals who use the top part of the scale will tend to have a smaller stretcher). The gross shifter corresponds to the rating a respondent would give when $w_{iq} = 0$ (the lowest height), an extreme and likely rare situation; we instead measure the shift relative to the *center*, $\gamma \equiv -Cov(a_i, \beta_i)/Var(\beta_i)$, which in our data corresponds to a more typical height for SWB questions. We define the *net shifter* α_i as the deviation of the respondent's rating from γ when $w_{iq} = \gamma$: $\alpha_i \equiv (a_i + \beta_i\gamma) - \gamma$. With this re-parametrization, the shifter and stretcher are uncorrelated: $Cov(\alpha_i, \beta_i) = Cov(a_i, \beta_i) + \gamma\, Var(\beta_i) = 0$. Substituting and rearranging equation (2):

$$r_{iq} - \gamma = \alpha_i + \beta_i(w_{iq} - \gamma) + \left[(a_{iq} - a_i) + (\beta_{iq} - \beta_i)w_{iq}\right]. \tag{3}$$

Hereafter, we refer to the net shifter as just the *shifter* when there is no risk of confusion.

Our econometric model flows directly from equation (3) but allows for response errors in both the stretcher and shifter components of the translation function:

$$r_{iq} - \gamma = \alpha_i + \beta_i(w_{iq} - \gamma + \epsilon_{iq}) + \eta_{iq}, \tag{4}$$

where ϵ_{iq} and η_{iq} are mean-zero error terms.[13] The error ϵ_{iq}, which is in units of individual i^*'s scale use, picks up heterogeneity and noise in how individuals assess a given state ω_{iq}; we refer to it as the *perception error*. We refer to the error η_{iq}, which is in units of individual i's scale use, as the *trembling-hand error* because it includes mechanical survey-response error, but equation (3) makes clear that it also picks up question-specific scale use.

Identifying and correcting for scale-use heterogeneity requires several independence assumptions.

[13] Equation (4) is closely related to the econometric models in the literature on estimating political candidates' positions based on survey respondents' (or interest group's) ratings, which also features shifter and stretcher parameters (e.g., Aldrich and McKelvey, 1977; Groseclose, Levitt, and Snyder, 1999). Because candidates' actual positions are fixed, the ratings of candidates' positions are analogous to CQ responses. The main complications in our analysis arise from scale-use adjustment of the SWB responses (for which the true level differs for each respondent), but in the political context, there is no analog of SWB responses.

Assumption 4: Response-Error Independence. The response errors $(\epsilon_{iq}, \eta_{iq})$ are:
(a) independent across individuals i and survey questions q and independent of each other, and
(b) independent of the stretcher, common-scale SWB, and any covariate: (β_i, w_{iq}, x_i).

Both 4(a) and 4(b) can be weakened to mean independence instead of independence, at the cost of modeling heteroskedasticity in the response errors. Similarly, it would be straightforward to allow for common components of ϵ_{iq} and η_{iq} across questions within a dimension or other question category. Some modifications of Assumption 4 can be viewed alternatively and equivalently as departures from Assumptions 1, 2 or 3. For example, mean-dependence of a CQ's perception error ϵ_{ic} on the underlying common-scale SWB, w_{is}, could be viewed as a failure of Assumption 2. Such modifications of Assumption 4 are likely the most difficult to accommodate.

When we describe an estimator for α_i and estimate its distribution in Section IV, we additionally assume that the response errors $(\epsilon_{iq}, \eta_{iq})$ are independent of the shifter α_i. We omit this statement from Assumption 4 because it is not needed for any of the scale-use-adjustment estimators we propose in Section V, all of which can be implemented after differencing out an estimate of α_i.

Below, we use equation (4) to calculate the asymptotic bias in empirical applications that ignore scale-use heterogeneity. The stretcher contributes to bias in all of the applications we consider. Thus, approaches that correct only for heterogeneity in the shifter—including fixed-effects regressions common when analyzing SWB panel data—are, in general, biased.

IV. Estimating Variation in General Scale Use

Under Assumptions 1′, 2, 3, and 4, we can use the responses to the CQs to estimate each individual's shifter and stretcher parameters. We do so in three steps. First, for each CQ c, we estimate w_c by the mean response across individuals. This estimator is unbiased and consistent as the number of individuals I becomes large:

$$\widehat{w}_c \equiv \frac{1}{I}\sum_i r_{ic} \underset{I\to\infty}{\longrightarrow} E(r_{ic}|w_c) = \gamma + E(\alpha_i + \eta_{ic}|w_c) + E(\beta_i|w_c)(w_c - \gamma) + E(\beta_i\epsilon_{ic}|w_c) = w_c.$$

Second, for each individual i, the ordinary least squares (OLS) regression of r_{ic} on $\widehat{w}_c$, where each observation in the regression corresponds to one of the CQs, gives estimators for the gross shifter a_i and the stretcher β_i, which we call $\hat{a}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$, respectively. Intuitively, this regression estimates the average level and dispersion of individual i's ratings of the CQs relative to the sample mean ratings of the CQs. As long as $\widehat{w}_c$ is estimated with negligible error, the estimators $\hat{a}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$ are unbiased and consistent as the number of CQs becomes large, $C \to \infty$. Finally, to obtain an estimate of the net shifter that is uncorrelated with the estimated stretcher, we set $\hat{\gamma}$ equal to the negative of the coefficient from a regression of $\hat{a}_{i,OLS}$ on $\hat{\beta}_{i,OLS}$ across individuals, and we set $\hat{\alpha}_{i,OLS} \equiv \left(\hat{a}_{i,OLS} + \hat{\beta}_{i,OLS}\hat{\gamma}\right) - \hat{\gamma}$.

While large I is often reasonable, we anticipate that large C will generally be a poor approximation. Indeed, even in our data with 18 calibration questions—substantially larger than we expect will normally be available in practice—the mean standard errors of $\hat{\alpha}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$ (4.10 and 0.27) are large relative to their standard deviations (9.14 and 0.38).

Despite their estimation error (which acts as measurement error in this context), $\hat{\alpha}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$ are useful as dependent variables in regressions to study variation in scale use. To illustrate, columns 3-4 of Table 3 show regressions of $\hat{\alpha}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$ on demographics. Like the descriptive regressions of the mean and standard deviation of CQ responses on demographics in columns 1-2 of Table 3, these regressions point to demographic correlates of scale use. However, unlike the earlier regressions, the regressions in columns 3-4 of Table 3 are informative about variations in the shifter and the stretcher separate from each other and from response error variances.[14]

To measure the overall amount of scale-use heterogeneity in our sample, we estimate the distributions of the shifter and the stretcher. Unfortunately, the estimation error in $\hat{\alpha}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$ inflates their variance relative to α_i and β_i, so the distributions of $\hat{\alpha}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$ are poor estimators of the distributions of α_i and β_i when C is small. However, we can consistently (as $I \to \infty$) estimate the distributions of α_i and β_i by maximum likelihood if we assume

[14] From equation (4), an individual's mean of CQ responses is an affine combination of shifter and stretcher, $E(r_{ic}|i) = \gamma + \alpha_i + \beta_i(E(w_c) - \gamma)$, while an individual's standard deviation of CQ responses depends on not only variation in the stretcher but also the error variances, $\sqrt{Var(r_{ic}|i)} = \sqrt{\beta_i^2 Var(w_c) + \beta_i^2 Var(\epsilon_{ic}|i) + Var(\eta_{ic}|i)}$.

parametric distributions for $(\alpha_i, \beta_i, \epsilon_{ic}, \eta_{ic})$. The distributions of $\hat{\alpha}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$, together with the residuals from the OLS regression from which they are estimated, can then be compared with what would be predicted by the fitted model; in this way, they give us diagnostics for guiding the choice of the parametric distributions, as we illustrate below.

We assume these parametric distributions for the scale-use parameters:

$$\begin{aligned} &(\alpha_i, \beta_i) \text{ jointly normal} \\ &\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2) \\ &\beta_i \sim \mathcal{N}\left(1, \sigma_\beta^2\right), \end{aligned} \tag{5}$$

and the response errors,

$$\begin{aligned} \epsilon_{ic} &\sim \mathcal{N}\left(0, \sigma_\epsilon^2\right) \\ \eta_{ic} &\sim \mathcal{N}\left(0, \sigma_{\eta_i}^2\right) \\ \ln\left(\sigma_{\eta_i}\right) &\sim \mathcal{N}\left(\mu_{\ln\sigma_\eta}, \sigma_{\ln\sigma_\eta}^2\right). \end{aligned} \tag{6}$$

By Assumption 4, $(\beta_i, \epsilon_{ic}, \eta_{ic})$ are mutually independent. Since α_i and β_i are uncorrelated by construction, the assumption of their joint normality implies that they are independent. As mentioned earlier, we also assume that α_i is independent of ϵ_{ic} and η_{ic}.

To assess whether the parametric distributions provide a good fit to the data, we simulated 100 datasets the same size as ours (3,358 individuals and 18 calibration questions), with parameter values assumed equal to their estimates. In each simulated dataset, we estimated $\hat{\alpha}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$ (see Web Appendix Figure J.1 for their distributions). The simulated distributions of $\hat{\alpha}_{i,OLS}$ are very similar to what we observe. The simulated distributions of $\hat{\beta}_{i,OLS}$ are not as good but better than when we instead assume a log-normal distribution for β_i. (Log-normal has the advantage that it rules out negative values of β_i, but this advantage turns out to be minor because, given the estimated normal distribution, a negative value occurs with probability less than 0.03%.) We allow the variance of η_{ic} (but not the variance of ϵ_{ic}) to vary across individuals because we find that the heterogeneity in σ_{η_i} enables the simulated datasets to

provide a reasonable fit to the density of the standard deviation of the residuals from the OLS regression for $\hat{\alpha}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$ (Web Appendix Figure J.1).

We estimate the parameters by maximum likelihood (see Web Appendix F.1 for details) and report their estimates in Table 4. The point estimate for σ_α means that a 1-standard-deviation increase in α_i corresponds to a 7.87-point higher rating on the 0-100 scale. The point estimate for σ_β implies that 95% of β_i's fall in the interval [0.4, 1.6]. The estimated γ of 59.90 is the value that provides the best fit for the independence between α_i and β_i, suggesting that differences in scale use across respondents are centered around roughly 60 on the 0-100 scale. The point estimate for σ_ϵ suggests a 1-standard-deviation increase in ϵ_{ic} increases r_{ic} by about 7.1 points at the mean β_i value of 1. The estimates of $\mu_{\ln\sigma_\eta}$ and $\sigma_{\ln\sigma_\eta}$ suggest that there is considerable trembling-hand error, with an unconditional standard deviation of 16.1, as well as much cross-individual heterogeneity in σ_{η_i}. Overall, the estimates are consistent with substantial general scale-use heterogeneity in our sample.

The fraction of this heterogeneity explained by our measured demographic variables is small (consistent with findings for response styles; Weijters, Geuens, and Schillewaert, 2010), suggesting that most general scale-use heterogeneity is within-group. More precisely, using our results in Tables 3 and 4, we calculate that the variances in α_i and β_i explained by the demographics are only 7% and 14% of the total variances in α_i and β_i, respectively.

In addition to providing estimates of the extent of scale-use heterogeneity, the maximum-likelihood estimates of $\left(\sigma_\beta, \sigma_\epsilon, \mu_{\ln\sigma_\eta}, \sigma_{\ln\sigma_\eta}\right)$ are an input to our semi-parametric estimators for scale-use correction (Section V.B). In our maximum-likelihood estimators for scale-use correction (Section V.D), we extend equation (5) with additional parametric assumptions about the SWB data.

V. Adjusting for Scale-Use Heterogeneity

V.A. Econometric Strategy

Our econometric strategy is motivated by the observation that adjusting for scale use at the individual level is unlikely to be fruitful. Using an individual's SWB-question response r_{is},

our estimates $\left(\hat{\alpha}_{i,OLS}, \hat{\beta}_{i,OLS}, \hat{\gamma}\right)$ from CQ responses, and equation (4), the natural estimator for an individual's common-scale SWB is

$$\widehat{w}_{is} = \frac{r_{is} - \hat{\gamma} - \hat{\alpha}_{i,OLS}}{\hat{\beta}_{i,OLS}} + \hat{\gamma} \xrightarrow[I\to\infty \text{ and } C\to\infty]{} w_{is} + \epsilon_{is} + \frac{\eta_{is}}{\beta_i}. \quad (7)$$

This estimator is unbiased (albeit still affected by response errors) in the large-sample limits of both individuals and CQs. However, the division by $\hat{\beta}_{i,OLS}$, which is noisy and may be close to zero, means that the number of CQs would have to be extremely large to avoid outliers and serious biases. Therefore, rather than trying to estimate w_{is} at the individual level, we instead construct estimators of the SWB moment underlying each of four common types of applications: the SWB's mean, $E(w_{is})$, its covariance with a demographic variable, $Cov(x_i, w_{is})$, its covariance with another SWB measure, $Cov(w_{is}, w_{is'})$, and its variance, $Var(w_{is})$. None of our estimators require dividing by $\hat{\beta}_{i,OLS}$.

For each SWB moment of interest, we develop three classes of estimators: a method-of-moments (MOM) estimator, a semi-parametric estimator, and a maximum-likelihood estimator (MLE). We refer to the latter as the "comprehensive MLE" to distinguish it from Section IV's "CQ-only MLE." As summarized in Table 5, the three classes of estimators rely on different key assumptions (in addition to Assumptions 1′, 2, 3, and 4), and each has advantages and disadvantages. While there are reasons to prefer other estimators in some cases, we have default recommendations regarding which estimator to use: the semi-parametric estimator for $E(w_{is})$, the MOM estimator for $Cov(x_i, w_{is})$, and the comprehensive MLE estimator for $Cov(w_{is}, w_{is'})$ and $Var(w_{is})$. We explain the reasons for these recommendations below. Moreover, for the SWB moments other than $Cov(x_i, w_{is})$, we discourage use of our MOM estimators because the estimators require assuming independence between β_i and w_{is}, and violation of that assumption is likely to lead to substantial bias; for that reason, the corresponding entries in Table 5 are in gray.

All of the estimators involve estimating some parameter(s) and then subsequently treating those parameters as known. To obtain standard errors that correctly account for uncertainty in these parameter estimates, we draw bootstrap samples of individuals and repeat the estimation procedure. In our main specification, we draw 100 bootstrap samples.

For concreteness and brevity, we focus here on describing each of the estimators in the context of the moment(s) of SWB for which it is our default recommendation. We formally derive all three estimators for the four SWB moments in Web Appendices D through F.

V.B. Semi-parametric Estimator

We begin with our recommended estimator for $E(w_{is})$. Using equation (4), the bias from naïvely using the population mean of the reports, $E(r_{is})$, as an estimator of $E(w_{is})$ is:

$$E(r_{is}) - E(w_{is}) = Cov(\beta_i, w_{is}). \tag{8}$$

There is no bias from the shifter because its mean in the population is zero. Bias arises from covariance between the stretcher and common-scale SWB because individuals with larger stretchers effectively get larger weights when the average of raw SWB responses is naïvely used to estimate $E(w_{is})$.

The semi-parametric estimator eliminates the bias from this covariance by estimating a non-parametric model of the relationship between w_{is} and β_i. It takes as inputs estimates of the parameters $\left(\sigma_\beta, \sigma_\epsilon, \mu_{\ln\sigma_\eta}, \sigma_{\ln\sigma_\eta}\right)$ from the CQ-only MLE described in Section IV.

We begin by expressing the estimand as

$$E(w_{is}) = E\left(w_{is} - \tfrac{1}{C}\mathbf{1}'\boldsymbol{w_C}\right) + \tfrac{1}{C}\mathbf{1}'\boldsymbol{w_C}, \tag{9}$$

where $\frac{1}{C}\mathbf{1}'\boldsymbol{w_C}$ is the (non-random) mean of the common-scale heights of the CQs; we do so because $w_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{w_C}$ is closely related to the empirical object we will use for estimation, $r_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{r_{iC}}$. This mean of the common-scale heights, $\frac{1}{C}\mathbf{1}'\boldsymbol{w_C}$, can be estimated using the estimates $\widehat{w}_c = \frac{1}{I}\sum_i r_{ic}$ of the elements of $\boldsymbol{w_C}$ (see Section IV). The first term on the right-hand side of equation (9) implies that the relationship between w_{is} and β_i that we need to model is the dependence of $E\left[\left(w_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{w_C}\right)|\beta_i\right]$ on β_i. The remainder of this derivation obtains an estimator for this function.

We non-parametrically model the conditional expectation:

$$E\left[w_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{w}_C|\beta_i\right] = A_{0s} + A_{1s}\beta_i + A_{2s}\beta_i^2 + \cdots + A_{Ks}\beta_i^K, \quad (10)$$

where the A_{ks}'s are unknown parameters that we will estimate. In principle, this polynomial expansion can be made an arbitrarily good approximation by taking K large. In practice, we use $K = 1$, which works well in our simulations (see Web Appendix G). Once we estimate the A_{ks}'s, we integrate equation (10) with respect to β_i to arrive at the desired unconditional expectation: $E\left[w_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{w}_C\right] = A_{0s} + A_{1s}E(\beta_i) + A_{2s}E(\beta_i^2) + \cdots + A_{Ks}E(\beta_i^K)$. We know $E(\beta_i) = 1$, and for any $k \geq 2$, we estimate $E\left(\beta_i^k\right)$ using the distribution of β_i estimated from the CQ-only MLE. Plugging the resulting expression into equation (9) gives our semi-parametric estimate.

To estimate the A_{ks}'s, we observe that

$$E\left[r_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{r}_{iC}|\beta_i, \hat{\beta}_{i,OLS}\right] = E\left[r_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{r}_{iC}|\beta_i\right] = \beta_i E\left[w_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{w}_C|\beta_i\right], \quad (11)$$

where $\hat{\beta}_{i,OLS}$ is our estimate of i's stretcher based on CQ ratings, defined in Section IV. The first equality is trivially true if $\frac{1}{C}\mathbf{1}'\boldsymbol{r}_{iC}$ and $\hat{\beta}_{i,OLS}$ are calculated from different CQs. We prove in Web Appendix D.4 that it is still true when all the CQs are used for both, as we do in practice in order to use the data efficiently, for large I. The second equality holds because once we subtract out the shifter, i's SWB response is just a stretched version of common-scale SWB.

Our estimating equation comes from substituting equation (10) into (11) and then taking the expectation of the resulting equation with respect to β_i, conditional on $\hat{\beta}_{i,OLS}$:

$$\begin{aligned} E\left[r_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{r}_{iC}|\widehat{\beta}_{i,OLS}\right] = \; & A_{0s}E\left(\beta_i|\widehat{\beta}_{i,OLS}\right) + A_{1s}E\left(\beta_i^2|\widehat{\beta}_{i,OLS}\right) + A_{2s}E\left(\beta_i^3|\widehat{\beta}_{i,OLS}\right) + \cdots \\ & + A_{Ks}E\left(\beta_i^{K+1}|\widehat{\beta}_{i,OLS}\right). \end{aligned} \quad (12)$$

Given the assumptions of the CQ-only MLE from Section IV, we can calculate the values of the regressors by numerically integrating β_i^k with respect to the density of $f\left(\beta_i|\hat{\beta}_{i,OLS}\right)$. The density can be found by applying Bayes' rule to $\hat{\beta}_{i,OLS}|\beta_i$, whose distribution is normal with mean β_i and

variance that depends on the CQ-only MLE parameters. Substituting the CQ-only MLE estimates for the parameter values, we can then calculate the conditional expectations $E(\beta_i|\hat{\beta}_{i,OLS})$, $E(\beta_i^2|\hat{\beta}_{i,OLS})$, etc. Regressing each respondent's $\left(r_{is} - \frac{1}{C}\mathbf{1}'\boldsymbol{r}_{iC}\right)$ on these provides consistent estimates of the A_{ks}'s.

The semi-parametric estimator is our recommended estimator for $E(w_{is})$ because, relative to the comprehensive MLE estimator, it is computationally lighter and makes less restrictive distributional assumptions.

V.C. Method-of-Moments Estimator: Mean-Matched Benchmarking

For estimating $Cov(x_i, w_{is})$, we turn to our MOM estimator. The idea is to use CQs to estimate, for each respondent, what their response to SWB question s would be if they applied their own general scale use to the population mean of common-scale SWB, and then adjust their actual response r_{is} by that amount before calculating the covariance with x_i. Before deriving the estimator and highlighting its key assumptions, we describe the three steps for implementing the estimator:

Step (i): We construct the OLS prediction of what a particular individual would report for something that has a common-scale value of h based on that individual's CQ ratings:

$$\hat{r}_i(h) = \hat{a}_{i,OLS} + h\hat{\beta}_{i,OLS}.$$

(See Section IV for the computation of $\hat{a}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$.) We call $\hat{r}_i(h)$ a *mean-matched benchmark (MMB)* because h here is typically (though it does not have to be) an estimate of a population mean, as in step (ii) below. The MMB is the empirical counterpart to the unobserved general-scale-use translation function. Indeed, when Assumptions 1', 2, 3, and 4 are satisfied and all scale-use heterogeneity is assumed to be general,

$$\tau_{i^*\to i}(h) = a_i + h\beta_i.$$

The MMB is the same formula, except with the true individual-specific parameter values a_i and β_i replaced by the estimates $\hat{a}_{i,OLS}$ and $\hat{\beta}_{i,OLS}$.

Step (ii): For the SWB question s, we estimate its mean in the population, $E(w_{is})$ (preferably using the semi-parametric or comprehensive MLE estimator for mean SWB).

Step (iii): Using the MMB $\hat{r}_i(E(w_{is}))$, we estimate $Cov(x_i, w_{is})$ by the sample analog of $Cov(x_i, r_{is} - \hat{r}_i(E(w_{is})))$, or equivalently, $Cov(x_i, r_{is}) - Cov(x_i, \hat{r}_i(E(w_{is})))$.

Thus, the MOM estimator adjusts the naïve estimator, $Cov(x_i, r_{is})$, for general scale use by subtracting (an estimate of) $Cov(x_i, \hat{r}_i(E(w_{is})))$.[15]

The bias from naïvely using $Cov(x_i, r_{is})$ as an estimator for $Cov(x_i, w_{is})$ clarifies both the value and the limitations of the MOM estimator (see Web Appendix C.2 for a derivation of this bias formula):

$$Cov(x_i, r_{is}) - Cov(x_i, w_{is}) = Cov(x_i, \hat{r}_i(E(w_{is}))) + E\big[\big(\beta_i - E(\beta_i)\big)\big(x_i - E(x_i)\big)\big(w_{is} - E(w_{is})\big)\big]. \quad (13)$$

Since (aside from response errors uncorrelated with x_i) the MMB $\hat{r}_i(E(w_{is}))$ differs across individuals solely due to heterogeneity in general scale use, the first term on the right-hand side of equation (13) captures between-demographic-group differences in general scale use. The second term is the (unstandardized) co-skewness between the stretcher, the common-scale SWB, and the demographic x_i. Intuitively, all individuals are equally weighted in $Cov(x_i, w_{is})$, but individuals with larger stretchers get more weight in $Cov(x_i, r_{is})$.[16] Note that the co-skewness may be non-zero even when x_i is independent of β_i and w_{is} separately. The moment condition that underlies our MOM estimator comes from rearranging equation (13) and assuming the co-skewness term is equal to zero:

[15] We caution that while $r_{is} - \hat{r}_i\big(E(w_{is})\big)$ is being used as if it were an individual-level estimate of scale-use-corrected SWB, it is not; even after ignoring response errors, it is an estimate of $r_{is} - \big(a_i + \beta_i E(w_{is})\big) = \beta_i\big(w_{is} - E(w_{is})\big)$. That is, it corrects for heterogeneity in the shifter but not heterogeneity in the stretcher.

[16] Another intuition is related to the bias from using the naïve estimator for the mean, $E(r_{is}) - E(w_{is}) = Cov(\beta_i, w_{is})$, discussed in Section VI.A. The co-skewness indicates how this bias varies with x_i.

(14) $$Cov(x_i, w_{is}) = Cov(x_i, r_{is} - \hat{r}_i(E(w_{is}))).$$

The MOM estimator's key assumptions are zero co-skewness and the assumptions of whichever method is used to estimate $E(w_{is})$. We recommend the MOM estimator for $Cov(x_i, w_{is})$ because it is the most intuitive; it is computationally light once $E(w_{is})$ has been estimated (the only non-trivial step); and it makes less restrictive distributional assumptions than the comprehensive MLE estimator. However, since our semi-parametric estimator for $Cov(x_i, w_{is})$ (detailed in Web Appendix D) does not assume zero co-skewness, it can be used to assess robustness to violations of the zero-co-skewness assumption. By finding similar results for $Cov(x_i, w_{is})$ from the MOM and semi-parametric estimators in Section VI.B below, we conclude that the zero-co-skewness assumption is a good approximation in our data.

V.D. Comprehensive MLE Estimator

For estimating $Cov(w_{is}, w_{is'}) \equiv \sigma_{w_s,w_{s'}}$ and $Var(w_{is}) \equiv \sigma^2_{w_s}$, we now turn to our comprehensive MLE estimator. In addition to CQ ratings (used in the CQ-only MLE), the comprehensive MLE also uses data on SWB ratings and demographics. The comprehensive MLE estimator is a natural extension of the CQ-only MLE: it jointly estimates the scale-use parameters and response-error variances as before but now, in addition, estimates the variances and covariances of common-scale SWB. Moreover, the comprehensive MLE naturally delivers estimates of the coefficients from a regression of common-scale SWB on the demographics, as well as mean common-scale SWB. We describe its application to multiple SWB questions at once in order to estimate covariances.

Specifically, we estimate the model described by equations (4) and (6) and by the following extension of (5):

(15) $$\begin{bmatrix} a_i \\ \beta_i \\ w_{i1} \\ \vdots \\ w_{iS} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 1 \\ \boldsymbol{x}_i'\boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{x}_i'\boldsymbol{b}_S \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{a,\beta} & \sigma_{a,w_1} & \cdots & \sigma_{a,w_S} \\ \sigma_{a,\beta} & \sigma_\beta^2 & \sigma_{\beta,w_1} & \cdots & \sigma_{\beta,w_S} \\ \sigma_{a,w_1} & \sigma_{\beta,w_1} & \sigma_{w_1}^2 & \cdots & \sigma_{w_1,w_S} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{a,w_S} & \sigma_{\beta,w_S} & \sigma_{w_1,w_S} & \cdots & \sigma_{w_S}^2 \end{bmatrix} \right),$$

where the (α_i, β_i) part of (15) is equivalent to (5) but parameterized with the gross shifter α_i instead of the net shifter (after estimation, we recover estimates of the center $\gamma = -\sigma_{\alpha,\beta}/\sigma_\beta^2$ and the net shifter's variance $\sigma_\alpha^2 = \sigma_\alpha^2 - \gamma^2\sigma_\beta^2$). The rest of (15) uses the SWB data. The coefficients from regressions of common-scale SWB, w_{is}, on a vector of demographics, $\boldsymbol{x}_i$, are the parameter vector $\boldsymbol{b}_s$. The sample average of $\boldsymbol{x}_i'\boldsymbol{b}_s$ estimates mean SWB $E(w_{is})$. The bottom-right $(S \times S)$-dimensional submatrix of the variance-covariance matrix in equation (15) is the variance-covariance matrix of the common-scale SWBs; thus, estimating these parameters yields estimates of the variance of SWB and the pairwise covariances of SWBs.

The comprehensive MLE estimator is the most computationally intensive,[17] and its parametric assumptions about the joint distribution of $(\alpha_i, \beta_i, w_{is})$ are strong. For example, the MOM estimator's assumption of zero co-skewness is implied by the assumption in equation (15) that the covariance between β_i and w_{is} does not vary with x_i. Despite these limitations, the comprehensive MLE estimator has four desirable features. First, it jointly estimates all four SWB moments. Second, it guarantees positive definiteness of the variance-covariance matrix in (15)—which is why it is our default recommendation for applications involving the variance or covariance of SWB. Third, it enables the estimation of covariances between common-scale SWB and scale-use parameters, which is useful for descriptive purposes and for quantifying the biases from not adjusting for scale use. Fourth, it can be extended to allow for top- and bottom-coding, an extension we develop and estimate in Web Appendix F.3 (the estimates do not meaningfully change).

V.E. Finite-Sample Performance, Convergence Speed, and Robustness

Under their respective assumptions, all of our estimators are consistent for fixed C as $I \to \infty$ (and $K \to \infty$ for the semi-parametric estimator) but may be biased in small samples. We conduct simulations to evaluate our estimators' finite-sample performance, convergence speed, and robustness to violations of their assumptions. In brief, for all of the recommended estimators,

[17] The comprehensive MLE estimator may become computationally prohibitive to estimate when the number of SWB questions is large; in that case, we recommend using the semi-parametric estimators of the variance and covariances, and whenever necessary using subsequent rectification procedures (such as replacing negative eigenvalues in the diagonalization of the estimated variance-covariance matrix with tiny positive values) to impose positive definiteness.

convergence speed appears to be $\sqrt{I}$ (for $C \geq 3$), and, with 100 replications, we cannot statistically detect bias. The semi-parametric estimator is the most robust to the various forms of misspecification we examine. We describe the simulations in detail in Web Appendix G.

VI. Applications

We now illustrate our general-scale-use adjustment methods by applying them to four common applications of SWB data, each corresponding to one of the four SWB moments of interest. We highlight that correctly interpreting results of SWB applications generally requires additional assumptions (see, e.g., Benjamin, Cooper, Heffetz, and Kimball, 2023), which we do not discuss here because our focus here is estimation.

VI.A. Mean SWB: $E(w_{is})$

Customer and employee satisfaction surveys aim to estimate mean SWB. Besides interest in mean SWB from these applications, other estimators (including the MOM estimator below) need an estimate of mean SWB to estimate other SWB moments. For each of the four SWB questions asked on the UK's Office of National Statistics (ONS) Opinions Survey, Figure 6 shows the mean r_{is} in our sample as well as our estimates of $E(w_{is})$ from our semi-parametric and comprehensive MLE estimators. The two estimators produce largely similar estimates. In almost all cases, the mean of common-scale SWB is greater than the mean of raw SWB reports (consistent with our estimates of $Cov(\beta_i, w_{is}) < 0$; recall equation (8)), although the magnitude of the bias varies by SWB question.

VI.B. Covariance with a Demographic: $Cov(x_i, w_{is})$

The most common application of SWB data by economists is a regression of an SWB measure on demographic and other variables (for surveys of this literature, see Krueger and Stone, 2014; Helliwell and Barrington-Leigh, 2010). For example, regressions that include gender, religiosity, or having children are the basis of well-known findings. When the regressors include income, coefficient ratios have been used to price (in money) non-market goods, including, for example, unemployment, clean air, and the life of a relative (e.g., Di Tella,

MacCulloch, and Oswald, 2001; Clark, Frijters, and Shields, 2008; Levinson, 2012; Deaton, Fortson, and Tortora, 2010).

Such a regression aims to estimate $Var(\boldsymbol{x}_i)^{-1}Cov(\boldsymbol{x_i}, w_{is})$, where $Cov(\boldsymbol{x_i}, w_{is})$ is the vector of covariances between w_{is} and each regressor in the vector $\boldsymbol{x}_i$. We refer to the regressors as "demographic variables," but the asymptotic bias we discuss applies to any right-hand-side variables that are not themselves affected by scale-use heterogeneity. The asymptotic bias from ignoring scale-use heterogeneity in the coefficients from such a regression is $Var(\boldsymbol{x}_i)^{-1}[Cov(\boldsymbol{x}_i, r_{is}) - Cov(\boldsymbol{x_i}, w_{is})]$, where each element of the vector in square brackets is given by equation (13).

Table 6 reports a life satisfaction regression. Column 1 reports a "standard" regression from the literature, unadjusted for scale-use heterogeneity: the dependent variable is respondents' rating of life satisfaction on a 0-100 scale. We chose 29 regressors common in the literature. The no-scale-use-correction regression in the column replicates some typical results. For example, life satisfaction is positively associated with log income, religiosity, being married and not separated, and with the age-squared term, and is negatively associated with being unemployed.

Columns 2-4 adjust for scale-use heterogeneity using the MOM, semi-parametric, and comprehensive MLE estimators, respectively. To compare the columns statistically, Web Appendix Table J.1 reports the differences in coefficients across columns, with standard errors calculated from our bootstrap samples. The results are largely similar across columns. Under the assumptions of the semi-parametric estimator, the difference between columns 2 and 3 is an estimate of $-Var(\boldsymbol{x}_i)^{-1}E\left[\left(\beta_i - E(\beta_i)\right)\left(w_{is} - E(w_{is})\right)\left(\boldsymbol{x}_i - E(\boldsymbol{x}_i)\right)\right]$. The magnitudes are small relative to the standard errors. Indeed, across all our 33 Baseline SWB ratings as dependent variables (including the life satisfaction regression in Table 6) and all the 29 demographic regressors, only nine coefficients have a difference between columns 2 and 3 that is statistically distinguishable from zero at the 10% false-discovery-rate level. These results suggest that, in our data, the co-skewness term in equation (13) is small relative to other sources of error.

Columns 6-9 are analogous to columns 1-4, but with the dependent variable "no anxiety." Qualitatively, both the unadjusted and adjusted regression results look similar to those when life satisfaction is the dependent variable, but the general-scale-use adjustment often makes a bigger difference for "no anxiety." Analysis of the MMB can help shed light on why general-scale-use adjustment matters more for some SWB questions and demographics than others. Using equation

(13), if the co-skewness is zero, the asymptotic bias in the regression coefficients is

$$Var(\boldsymbol{x}_i)^{-1}[Cov(\boldsymbol{x}_i, r_{is}) - Cov(\boldsymbol{x}_i, w_{is})] = Var(\boldsymbol{x}_i)^{-1}Cov\left(\boldsymbol{x}_i, \hat{r}_i\big(E(w_{is})\big)\right).$$

That is, the differences between the corresponding coefficients in columns 1 and 2 are given by the coefficients in a regression of the MMB on the demographics. Thus, this regression can explain (and statistically test) our scale-use adjustments. Since the MMB depends on $E(w_{is})$, the scale-use adjustment will depend on the SWB question's height $h = E(w_{is})$. Columns 5 and 10 of Table 6 show regressions of the MMB on the demographics for two different target heights. The results indicate that, for most of the demographics in the regressions, adjustment for general-scale-use heterogeneity will make little difference for SWB questions such as life satisfaction whose $E(w_{is})$ near 70 but will matter more for SWB questions such as "no anxiety," that have values of $E(w_{is})$ near 55.

VI.C. Covariance across SWB: $Cov(w_{is}, w_{is'})$

Several common applications, especially in psychology, depend on the covariance between SWB questions, such as factor analysis or calculating the correlation between questions included in an SWB battery. In economics, researchers often study the time-series covariance of a single SWB question, which is a covariance where the question s' is question s asked at a different date with response errors assumed to be entirely transitory.

The asymptotic bias from ignoring scale use, $Cov(r_{is}, r_{is'}) - Cov(w_{is}, w_{is'})$, is:

$$\begin{aligned} &Cov(r_{is}, r_{is'}) - Cov(w_{is}, w_{is'}) \\ &= \sigma_\alpha^2 + Var(\beta_i)Cov(w_{is}, w_{is'}) + Var(\beta_i)[E(w_{is}) - \gamma][E(w_{is'}) - \gamma] \\ &+E\left[\left(\beta_i^2 - E(\beta_i^2)\right)\left(w_{is} - E(w_{is})\right)\left(w_{is'} - E(w_{is'})\right)\right] - Cov(\beta_i, w_{is})Cov(\beta_i, w_{is'}) \\ &+[E(w_{is'}) - \gamma][Cov(\beta_i^2, w_{is}) - Cov(\beta_i, w_{is})] + [E(w_{is}) - \gamma][Cov(\beta_i^2, w_{is'}) - Cov(\beta_i, w_{is'})]. \end{aligned} \tag{16}$$

There are two sources of bias, each of which contributes multiple terms to the formula. The first is due to variance across individuals in the shifter and stretcher parameters, which inflate $Cov(r_{is}, r_{is'})$ relative to $Cov(w_{is}, w_{is'})$. The second is that individuals with larger stretchers get more weight in the covariance of SWB.

The covariances between the four ONS (2011) SWB questions with and without adjustment for general scale use are shown in Web Appendix Table J.2. In our data, the comprehensive-MLE adjusted covariances differ a great deal from the unadjusted covariances.

To illustrate how ignoring scale use can lead to misleading conclusions in a specific application, we use the estimated covariances (and variances, discussed in Section VI.D below) to conduct a factor analysis of the 33 SWB questions. For the implied unadjusted and adjusted correlation matrix of the 33 SWB questions, respectively, Panels A and B of Table 7 show loadings of the first two factors estimated by unrotated orthogonal factor analysis. For reference, the table also shows the mean SWB response, $E(r_{is})$, and the semi-parametric estimate of $E(w_{is})$ for each SWB question. As general observations, note that (i) the variances explained (and factor loadings) are larger in Panel B than in Panel A because the adjustment removes the variation that is due to general-scale-use heterogeneity and response errors, and (ii) the standard errors are larger in Panel B because they account for estimation error in the adjustment.

The specific biases that will arise in factor analysis depend on how the various terms in equation (16) play out. For example, the first bias term in equation (16) generates an artifactual positive covariance between all SWB questions, because respondents with larger shifters give higher responses. However, the variance explained by the first factor in Panel B is *larger* than in Panel A because, in our data, this bias is more than offset by removing spurious variance. As another example, the third bias term in equation (16) says the covariance between two SWB questions is biased to a greater extent when both questions have $E(w_{is})$ further from γ; this is because covariance due to stretcher variation matters more relative to common-scale SWB covariance for such SWB questions. In the factor analysis, this bias can show up as a factor that loads more strongly on SWB questions with $E(w_{is})$ further away from γ. In our data, we think that the second factor in Panel A is an artifact of heterogeneity in the stretcher β_i. Consistent with this interpretation, the correlation between that factor and our estimates of $E(w_{is})$ is quite large: 0.67 (SE = 0.05).[18]

VI.D. Variance of SWB: $Var(w_{is})$

[18] Since $E(r_{is})$ will generally be highly correlated with $E(w_{is})$, our analysis implies that a factor having loadings highly correlated with the (non-scale-use-adjusted) means of the variables can be a diagnostic for possible scale-use confounding that does not require CQs.

A small but growing literature examines inequality in SWB, generally finding that inequality in SWB has declined in Western countries in the past few decades (e.g., Veenhoven, 2005; Stevenson and Wolfers, 2008; Easterlin, 2012; Clark, Flèche, and Senik, 2014), an interesting contrast with the increase in income inequality over the same period. Typically, inequality in SWB is measured as the variance (or standard deviation) in SWB responses. If the biases from ignoring scale use and response error are substantial, then it raises the concern that conclusions about differences over time or between groups in the variance of SWB may be sensitive to an implicit assumption that the biases are constant over time or between groups.

The formula for the asymptotic bias from ignoring scale use, $Var(r_{is}) - Var(w_{is})$, is the same as equation (16), except with $w_{is'}$ replaced by w_{is} and with one additional term: $\sigma_{\varepsilon}^2(Var(\beta_i) + 1) + E(\sigma_{\eta_i}^2)$. Thus, the asymptotic bias has the same two sources as in $Cov(w_{is}, w_{is'})$—variance in the shifter and stretcher across individuals, and individuals with larger stretchers getting more weight in the variance of SWB responses—plus an additional contribution to bias from the response errors in SWB responses. This additional term biases the variance of SWB responses upward relative to the variance of common-scale SWB.

All of our estimators for $Var(w_{is})$ require subtracting out an estimate of $\sigma_{\varepsilon}^2(Var(\beta_i) + 1) + E(\sigma_{\eta_i}^2)$, but the response-error variances for the SWB questions cannot be separately identified from $Var(w_{is})$. As a rough proxy, we obtain estimates of these response-error variances by assuming that these are equal to the response-error variances for the CQs (reported in Table 4). Since we expect the perception-error variance to be larger for SWB questions than CQs, we view this approach as likely underestimating this bias term and therefore likely providing an upper-bound estimate of $Var(w_{is})$.

For the four U.K. ONS SWB questions, Table 8 reports estimates of their variances using our data, with and without adjusting for scale use. The columns are analogous to the corresponding columns in Table 6: Column 1 shows the unadjusted estimates, most similar to what is typically reported, while columns 2 and 3 show adjusted estimates using the semi-parametric and comprehensive MLE estimators, respectively. The two estimators give somewhat different estimates. The standard errors on the MLE estimates are much smaller, presumably because of the structure of the parametric assumptions.

Despite the quantitative differences, the estimators tell a similar story qualitatively. For all four SWB questions, the adjusted variances are less than half as large as the unadjusted

variances. (Web Appendix Table J.3 shows similar results for the full set of 33 SWB questions we study.) Web Appendix Table J.4 indicates that roughly 4/5 of the adjustment is due to response-error variance, while the remaining 1/5 is due to scale-use heterogeneity. Using our data from a follow-up survey that contains repeated measures, we find that adjusting the variance of SWB responses for transitory measurement error—which comprises part of the response-error variance—reduces the variance by only about half of the amount that adjusting using the comprehensive MLE estimator does (see Column 2 of Web Appendix Table J.3).

These results indicate that caution is warranted in drawing conclusions from estimates of SWB inequality based on unadjusted SWB. For example, Clark, Flèche, and Senik (2014) find that "variation in happiness within countries is typically twice as high as that across countries," but the variation across countries is based on mean happiness within each country; taking the mean eliminates much of the variation due to scale-use heterogeneity and to response errors that affect raw SWB variation within countries.

Confounded estimates of the variance of SWB also matter in other applications. For example, if effect sizes are expressed in units of (unadjusted) standard deviations of SWB—e.g., the effect of income on SWB—these effect sizes will be too small by roughly a factor of 1.5.

VII. Additional Results

VII.A. Validation of Scale-Use Adjustments

As a validation test (inspired by Oswald, 2008), we examine whether self-reported height on our 0-100 scale becomes more strongly related to height measured in objective units after scale-use adjustment. To conduct this test, we asked Baseline respondents to report their height both on the 0-100 scale—which we call "subjective height"—and, in a different part of the survey, in feet and inches—which we call "objective height" (although it is self-reported). If scale-use heterogeneity and the response errors we model were the only source of discrepancy and if there were no measurement error in objective height, then the scale-use-adjusted coefficient from a regression of subjective height on objective height, after standardizing both (demeaning and dividing by their standard deviations), would be one. In practice, because both premises are imperfect approximations (for example, subjective height may be judged in comparison to a reference group), we expect the coefficient to be attenuated, but we expect less

attenuation after scale-use adjustment. When we regress standardized, unadjusted subjective height on objective height, the coefficient is 0.49 (SE = 0.04). To run the scale-use-adjusted analog of this regression, we standardize subjective height by subtracting the MMB for subjective height and dividing by the square root of the variance of common-scale subjective height as estimated by our comprehensive MLE. When we regress this variable on standardized objective height, the coefficient is much closer to one: 0.85 (SE = 0.09). We obtain similar results when we conduct the analysis for weight: the coefficient from the unadjusted regression is 0.40 (SE = 0.02), compared with 0.91 (SE = 0.04) from the scale-use-adjusted regression. Analyses using four additional subjective-objective pairs concerning air quality, crime rate, and family financial support confirm that adjusting for scale use aligns subjective measures more closely with objective measures. See Web Appendix I for full details.

VII.B. Relative Importance of General-Scale-Use Heterogeneity

How much scale-use heterogeneity is *general*-scale-use heterogeneity? To address this question, our analysis strategy needs to overcome several challenges. First, the variance in responses to SWB questions partly reflects heterogeneity in common-scale SWB. To focus only on scale-use heterogeneity, we therefore restrict our analysis to CQs. Second, the fraction of variance in CQ responses that is explained by general scale use will depend on the CQ dimension. We estimate the fraction of variance for each dimension we study, but focus on the average across dimensions. Third, the fraction of variance explained will depend on the height of the CQ. Thus, rather than directly studying CQ responses, we study MMBs constructed from the CQ ratings within each dimension, with the MMB target height held fixed across all the dimensions. Fourth, the MMBs inherit from the CQ responses perception and trembling-hand errors, which are not part of scale-use heterogeneity. To isolate the variance due to scale use, we estimate dimension-specific response-error variances (using a dimension-specific version of the CQ-only MLE described in Section IV) and correct for them.

To include a larger number of dimensions than we have in Baseline, we analyze data on 388 CQs from 60 dimensions collected in Bottomless (with 701 respondents). To assess the robustness of our conclusions to the height used when constructing the MMB, we use two empirically relevant heights: the highest of the four SWB questions asked on the UK's ONS (70, the mean for worthwhileness) and the lowest (54, the mean for "no anxiety").

If the MMBs did not contain response errors, our estimator for the fraction of scale-use variance in CQ dimension *d* explained by general scale use would be the R^2 from a regression of the mean MMB calculated from dimension-*d* CQs on the mean MMB calculated from all other CQs, excluding dimension *d*. We calculate the R^2's from these hypothetical regressions analytically, using our MLE estimates of the dimension-specific response-error variances to subtract out their effects. We then measure the overall fraction of scale-use heterogeneity explained by general scale use as the mean of the dimension-specific R^2's. Here we briefly summarize the results; for full details of the analysis and results, see Web Appendix H.

For the MMBs at means 70 and 54, the mean R^2's are 56.3% and 60.0%, respectively. We conclude that, in our data, general scale use comprises roughly 3/5 of the scale-use heterogeneity on average across all the CQs we study.

When we exclude visual dimensions (e.g., "How dark is this circle?") and what we call non-local public goods (e.g., vignettes asking about the level of corruption in government), restricting the data to 316 vignette CQs over 42 dimensions, the mean R^2's are 75.6% and 75.7% for MMBs at means of 70 and 54, respectively. One possible explanation for why the percentage is higher when we restrict to these vignette CQs is that the dimensional scale use is more similar within this category of dimensions. Another possible explanation is that the vignette CQs share a common confound, e.g., respondents projecting their own situation into the vignette.

VII.C. Persistence of General Scale Use

How persistent is general scale use? To obtain preliminary evidence, we analyzed data from a subset of 2,472 of our respondents who responded to our 18 Baseline CQs a second time in our Bottomless survey, with a median time gap of seven weeks apart. We extended our CQ-only MLE from Section IV to estimate persistent and transitory components of the model parameters (for details, see Web Appendix F.4). We estimate that 55.4% (SE = 2.2%) of the variance in the shifter and 88.5% (SE = 2.7%) of the variance in the stretcher are persistent; respondents are more persistent in how much they spread out their responses than in how high or low their ratings are on average. These results suggest that SWB panel regressions with respondent fixed effects only partially correct for shifter heterogeneity. We estimate that 99.7% (SE = 0.2%) and 3.2% (SE = 2.0%) of the variances in the perception error and trembling-hand error, respectively, are persistent. The negligible persistence of trembling-hand error suggests

that *question*-specific scale use (as defined in Section III.D) is minimal (in contrast, our analysis in Section VII.B suggests that *dimension*-specific scale use is non-trivial).

VII.D. General Scale Use with Alternative Scales

In this paper, we analyzed (CQs and) SWB questions that have a 0-100 scale and labeling only at the extremes of 0 and 100. Is general-scale-use heterogeneity also relevant for SWB questions asked with more commonly used scales? We address that question using a broad set of frequently employed SWB questions that we asked on Bottomless. Figure 7 shows the mean CQ rating (on our 0-100 scale) for each respondent (x-axis) and their mean SWB rating elicited on alternative response scales (y-axis). We find positive correlations (0.16–0.54) with 95% confidence intervals that exclude zero for all eight response scales. Figure 8 shows the analogous figure for standard deviations, limited to the four response scales for which we asked multiple SWB questions and had at least four response options. We again find positive correlations (0.10–0.53) with 95% confidence intervals that exclude zero in all cases. These results suggest that the scale-use tendencies captured using our 0-100 scale CQs apply more generally to SWB questions with other response scales.

VIII. Concluding Remarks

In this paper, we proposed a framework for measuring and adjusting for heterogeneity in scale use. Our framework overcomes concerns with existing approaches: it applies to continuous response scales, it does not rely solely (or at all) on vignettes when used for general-scale-use adjustment, and it identifies scale-use heterogeneity with calibration questions (CQs), which can be designed with the assumptions of our framework in mind. At least two CQs are needed to identify the shifter and stretcher parameters, and our simulations reported in Section V.E and Web Appendix G provide guidance on the value of additional CQs, depending on the number of survey respondents. In a proof-of-concept survey, we found evidence of substantial heterogeneity in scale use, and we discussed when, why, and how this heterogeneity can matter for conclusions of SWB research.

Many issues remain; we list six. First, what specifically should SWB researchers do to adjust for scale-use heterogeneity? Of course, survey designers can incorporate CQs such as those we developed into their surveys and implement the estimators we proposed. To help

researchers who analyze existing datasets that lack CQs, future work should incorporate CQs into a nationally representative survey. Then, in that data, researchers could regress a MMB on demographics to obtain the coefficients needed to implement our MOM adjustment to happiness regressions estimated in the dataset that lacks CQs.

Second, should researchers adjust for dimensional or general scale use? Our framework clarifies the tradeoff. In principle, adjusting for dimensional scale use would be a full adjustment for the scale-use heterogeneity relevant to the SWB dimension of interest, thereby satisfying Assumption 3 exactly. However, satisfying Assumptions 1 and 2 is challenging, especially for broad, highly subjective SWB dimensions such as life satisfaction. Those assumptions are more plausibly satisfied when adjusting for general scale use, especially with non-vignette CQs such as our visual CQs or vignette CQs for narrow, relatively objective dimensions. A potentially promising direction for future research would be to try to identify a category of dimensions within which scale use is similar to that of the SWB dimension of interest (approximately satisfying Assumption 3) *and* find corresponding CQs that are relatively immune to biases (approximately satisfying Assumptions 1 and 2).

Third, in our regression application, we examined regressions of SWB on objectively measured independent variables, such as age and income. However, sometimes researchers run regressions where both the dependent and independent variables are measured in non-physical units (for example, psychological measures of personality) and are thus influenced by scale use. In such a regression, scale-use heterogeneity induces correlation between the independent variable and the error term, leading to additional bias. The common practice in psychological batteries of reverse-coding half the questions could neutralize the effect of the shifter, but only if the shifter for reverse-coded items is the same as the shifter for non-reverse-coded items (so that these identical shifters cancel out when the reverse-coded items are subtracted from the non-reverse-coded items in the overall index)—an assumption that, as far as we are aware, remains untested in the literature. And even with reverse-coding, the stretcher can be a confound. Studying these issues and other applications of our scale-use adjustment methods to non-SWB measures are additional directions for future work.

Fourth, the applications of SWB data that we have focused on in this paper are correlational analyses—which yield causal conclusions only under additional assumptions that can be difficult to satisfy. Can scale-use heterogeneity be ignored when SWB measures are

dependent variables in randomized experiments, since it is balanced in expectation by randomization? Only under the assumption that the treatment does not affect scale use. Even then, if the stretcher is correlated with the treatment effect, individuals who have larger stretchers will be weighted more heavily in the estimated average treatment effect. Measuring and adjusting for scale-use heterogeneity in randomized experiments would avoid these issues.

Fifth, while our focus has been on cross-sectional analysis of SWB data, the important issue of hedonic adaptation is about the *dynamics* of SWB. The observation that SWB 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 as Lacey et al. (2008) point out, it could instead reflect a change over time in scale use. An important direction for future work is to study hedonic adaptation after correcting for scale use.

Finally, while our approach aims to eliminate (or mitigate) biases due to scale-use heterogeneity, whether or not common-scale SWB is on the appropriate scale for applications, or whether it needs to be further transformed, depends on additional assumptions. For example, economists typically wish to use SWB as a utility proxy to make group welfare comparisons. This use of the data requires three major assumptions: (i) survey respondents' interpretation of the SWB question matches the researcher's intended utility notion (which is often not the case; see Benjamin, Debnam Guzman, Fleurbaey, Heffetz, and Kimball, 2023) so that common-scale SWB *is* utility; (ii) a group's well-being is the mean of some monotonic transformation of the group members' utilities (which requires certain normative assumptions; see, e.g., Fleurbaey and Maniquet, 2011; Adler and Norheim, 2022); and (iii) common-scale SWB is the "right" monotonic transformation of utility. Assumption (iii) makes clear that putting the SWB survey responses on an interpersonally comparable scale is necessary but not sufficient for SWB group means (or regression coefficients) to have meaningful welfare interpretations (see Benjamin, Cooper, Heffetz, and Kimball, 2023).

References

Adler, Matthew D., and Ole F. Norheim. 2022. *Prioritarianism in Practice*. Cambridge, UK: Cambridge University Press.

Aldrich, John H., and Richard D. McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections*." The American Political Science Review* 71 (1): 111–130.

Benjamin, Daniel J., Kristen B. Cooper, Ori Heffetz, and Miles Kimball. 2023. "From Happiness Data to Economic Conclusions." Working Paper.

Benjamin, Daniel J., Kristen Cooper, Ori Heffetz, and Miles S. Kimball. 2017. "Challenges in Constructing a Survey-Based Well-Being Index." *American Economic Review* 107 (5): 81–85.

Benjamin, Daniel J., Jakina Debnam Guzman, Marc Fleurbaey, Ori Heffetz, and Miles Kimball. 2023. "What Do Happiness Data Mean? Theory and Survey Evidence." *Journal of the European Economic Association,* Published online, May 26.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot. 2014. "Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference." *American Economic Review* 104 (9): 2698–2735.

Birnbaum, M. H. 1994. "Psychophysics." In *Encyclopedia of Human Behavior*, Volume 3, Academic Press: 641– 650.

Bond, Timothy N., and Kevin Lang. 2019. "The Sad Truth About Happiness Scales." *Journal of Political Economy* 127 (4): 1629–1640.

Clark, Andrew E., Sarah Flèche, and Claudia Senik. 2014. "The Great Happiness Moderation." In *Happiness and Economic Growth: Lessons from Developing Countries*, edited by A.E. Clark and C. Senik: 32–139. Oxford, UK: Oxford University Press.

Clark, Andrew E., Paul Frijters, and Michael A. Shields. 2008. "Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles." *Journal of Economic Literature* 46 (1): 95–144.

Deaton, Angus. 2011. Comment on "Work Disability, Work, and Justification Bias in Europe and the U.S." In *Explorations in the Economics of Aging*, edited by D. A. Wise: 312–314. Chicago, IL: University of Chicago Press.

Deaton, Angus, Jane Fortson, and Robert Tortora. 2010. "Chapter 5: Life (Evaluation), HIV/AIDS, and Death in Africa." In *International Differences in Well-Being*: 105–136. Oxford, UK: Oxford University Press.

Di Tella, Rafael, Robert J. MacCulloch, and Andrew J. Oswald. 2001. "Preferences Over Inflation and Unemployment: Evidence from Surveys of Happiness." *American Economic Review* 91 (1): 335–341.

Easterlin, Richard A. 2012. "Life Satisfaction of Rich and Poor Under Socialism and Capitalism." *International Journal of Happiness and Development* 1 (1): 112–126.

Fleurbaey, Marc, and François Maniquet. 2011. *A Theory of Fairness and Social Welfare*. Cambridge, UK: Cambridge University Press.

Giustinelli, Pamela, Charles F. Manski, and Francesca Molinari. 2022. "Tail and Center Rounding of Probabilistic Expectations in the Health and Retirement Study." *Journal of Econometrics* 231 (1): 265–281.

Greenleaf, Eric A. 1992. "Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles." *Journal of Marketing Research* 29 (2): 176–188.

Groseclose, Tim, Steven D. Levitt, and James M. Snyder. 1999. "Comparing Interest Group Scores Across Time and Chambers: Adjusted ADA Scores for the U.S. Congress." *American Political Science Review* 93 (1): 33–50.

Helliwell, J. F. and Barrington-Leigh, C. P. 2010. "Viewpoint: Measuring and Understanding Subjective Well-being." *The Canadian Journal of Economics / Revue Canadienne d'Economique* 43 (3): 729–753.

Kapteyn, Arie, James P. Smith, and Arthur van Soest. 2007. "Vignettes and Self-reports of Work Disability in the U.S. and the Netherlands." *American Economic Review* 97 (1): 461–473.

Kapteyn, Arie, James P. Smith, and Arthur van Soest. 2009. "Life Satisfaction." *No. 4015.*

King, Gary, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *The American Political Science Review* 98 (1): 191–207.

Krueger, Alan B. and Arthur A. Stone. 2014. "Progress in Measuring Subjective Well-Being." *Science* 346 (6205): 42–44.

Lacey, Heather P., Angela Fagerlin, George Loewenstein, Dylan M. Smith, Jason Riis, and Peter A. Ubel. 2008. "Are They Really That Happy? Exploring Scale Recalibration in Estimates of Well-Being." *Health Psychology* 27 (6): 669–675.

Levinson, Arik. 2012. "Valuing Public Goods Using Happiness Data: The Case of Air Quality." *Journal of Public Economics* 96 (9): 869–880.

Loewenstein, George and Peter A. Ubel. 2008. "Hedonic adaptation and the role of decision and experience utility in public policy." Journal of Public Economics, 92 (8-9): pp. 1795–1810.

Luhmann, Maike, Wilhelm Hofmann, Michael Eid, and Richard E. Lucas. 2012. "Subjective Well-being and Adaptation to Life Events: A Meta-analysis." *Journal of Personality and Social Psychology* 102 (3): 592–615.

Márquez-Padilla, Fernanda and Jorge Álvarez. 2018. "Grading Happiness: What Grading Systems Tell Us About Cross-country Wellbeing Comparisons." *Economics Bulletin* 38 (2): 1138–1155.

Office for National Statistics (ONS). 2011. *Initial Investigation into Subjective Well-being from the Opinions Survey.* Newport, South Wales: ONS.

Oswald, Andrew J. 2008. "On the Curvature of the Reporting Function from Objective Reality to Subjective Feelings." *Economics Letters* 100 (3): 369–372.

Parducci, Allen. 1965. "Category Judgment: A Range-Frequency Model." *Psychological Review* 72 (6): 407–418.

Preston, Carolyn C. and Andrew M. Colman. 2000. "Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences." *Acta Psychologica* 104 (1): 1–15.

Rapkin, Bruce D., and Carolyn E. Schwartz. 2019. "Advancing Quality-of-Life Research by Deepening Our Understanding of Response Shift: A Unifying Theory of Appraisal." *Quality of Life Research* 28: 2623–2630.

Rossi, Peter E., Zvi Gilula, and Greg M. Allenby. 2001. "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach." *Journal of the American Statistical Association* 96: 20–31.

Schröder, Carsten, and Shlomo Yitzhaki. 2017. "Revisiting the Evidence for Cardinal Treatment of Ordinal Variables." *European Economic Review* 92: 337–358.

Sprangers, Mirjam A.G., and Carolyn E. Schwartz. 1999. "Integrating Response Shift Into Health-Related Quality of Life Research: A Theoretical Model." *Social Science and Medicine* 48 (11): 1507–1515.

Stevenson, Bestey and Justin Wolfers. 2008. "Happiness Inequality in the United States." *The Journal of Legal Studies* 37 (S2): S33–S79.

Stone, Arthur A., Stefan Schneider, Doerte U. Junghaenel, and Joan E. Broderick. 2019. "Response Styles Confound the Age Gradient of Four Health and Well-being Outcomes." *Social Science Research* 78: 215–225.

van Vaerenbergh, Yves, and Troy D. Thomas. 2013. "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies." *International Journal of Public Opinion Research* 25 (2): 195–217.

Veenhoven, Ruut. 2005. "Return of Inequality in Modern Society? Test by Dispersion of Life-Satisfaction Across Time and Nations." *Journal of Happiness Studies* 6 (4): 457–487.

Weijters, Bert, Hans Baumgartner, and Maggie Geuens. 2016. "The Calibrated Sigma Method: An Efficient Remedy for Between-group Differences in Response Category Use on Likert Scales." *International Journal of Research in Marketing* 33 (4): 944–960.

Weijters, Bert, Maggie Geuens, and Niels Schillewaert. 2010. "The Stability of Individual Response Styles." *Psychological Methods* 15 (1): 96–110.

Table 1. Four SWB Questions Based on the U.K.'s Office of National Statistics (2011)

SWB Question	Text of SWB Question
	Thinking about the past year, how would you rate…
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*

Table 2. Respondent Demographics

Demographic	Category	Baseline	Baseline, Passed Quality Control	Census
Gender	Male	49.6%	43.5%	49.0%
	Non-Male	50.4	56.5	51.0
Marital status	Married, not separated	65.7	50.4	51.1
	Never married	24.6	35.4	33.5
	Other	9.6	14.3	15.4
Highest education level completed	High school grad	7.9	10.4	28.3
	Some college	16.6	25.7	27.1
	Bachelor's degree	52.6	46.2	22.2
	Graduate degree	22.6	17.2	12.8
Age	18–29	18.7	14.9	20.6
	30–39	35.3	33.6	17.5
	40–49	21.6	23.4	16.0
	50–64	19.1	20.7	24.9
	65 and older	5.2	7.4	21.1
Household income	Less than $30,000	14.2	17.3	22.1
	$30,000–$49,999	21.7	21.1	15.7
	$50,000–$69,999	22.2	20.1	13.4
	$70,000–$89,999	18.4	15.4	10.8
	$90,000–$119,999	13.5	12.8	12.0
	$120,000 and above	10.0	13.3	26.0
Region	Midwest	22.8	22.9	20.8
	Northeast	16.6	18.1	17.4
	South	38.0	38.5	38.1
	West	22.6	20.5	23.7
Race	White, and other	79.1	78.6	63.3
	Black	6.8	8.1	12.1
	Hispanic, Latino, or Spanish	9.8	7.3	18.7
	Asian	4.3	6.0	5.9
Household size	1	15.6	20.7	28.5
	2	27.9	30.3	35.0
	3	19.0	19.5	15.0
	4 and above	37.6	29.5	21.5
Employment status	Employed	85.7	78.4	57.6
	Unemployed	6.3	9.9	3.8
	Not in labor force	8.0	11.8	38.5
Obs.		5466	3358	

Note: All numbers are percentages. *Baseline Survey Demographics:* "Non-Male" includes Female and Non-Binary options. 41 Baseline survey respondents reported Non-Binary, including 37 in the "Passed Quality Control" subsample. "Other" marital status includes people who are divorced, widowed, separated, or chose the "Other" option. The percentage of respondents who did not complete high school is omitted. "Region" is based on state of residence, using U.S. Census Bureau definition. "Not in labor force" includes Homemaker, Student, Disabled, and "Other" categories. Race options on the survey are non-exclusive; mutually exclusive categories (shown) are defined as follows: "Black" excludes respondents who also reported "Hispanic, Latino, or Spanish"; "Asian" excludes "Hispanic, Latino, or Spanish" and "Black"; "White, and other" includes those who reported "White," "American Indian or Alaska Native," "Middle Eastern or North African," "Native Hawaiian or Other Pacific Islander," and "Other" categories, but excludes "Hispanic, Latino, or Spanish," "Black," and "Asian." See Appendix K for screenshots. *Sources*: Authors' survey, 2020 Census.

Table 3. Regression of Mean and Standard Deviation of Baseline CQs, and $\widehat{\alpha}_i$ and $\widehat{\beta}_i$, on Demographics

	(1)	(2)	(3)	(4)
Demographics	Mean of CQs	Std. Dev. of CQs	$\hat{\alpha}_i$	$\hat{\beta}_i$
Demeaned age/10	-0.7†††	0.2††	-0.51††	0.01†
	(0.2)	(0.1)	(0.16)	(0.01)
(Demeaned age)^2/100	0.2††	-0.1	0.13	-0.01††
	(0.1)	(0.1)	(0.08)	(0.004)
Log(HH income)	-0.7†††	0.4†††	-0.23	0.05†††
	(0.2)	(0.1)	(0.24)	(0.01)
Unemployed	-1.3†††	1.0†††	0.05	0.13†††
	(0.5)	(0.3)	(0.51)	(0.02)
Employed part-time	-1.1†	0.7††	-0.50	0.06†††
	(0.5)	(0.3)	(0.53)	(0.02)
Out of labor force/other	-2.1	0.7	-1.09	0.12
	(0.5)	(0.3)	(0.54)	(0.02)
Married, not separated	1.9†††	-0.7†††	1.24††	-0.06†††
	(0.4)	(0.2)	(0.39)	(0.02)
Ever divorced	-0.4	0.4	-0.07	0.03
	(0.5)	(0.3)	(0.56)	(0.02)
Have ≥1 child	-0.5	0.2	-0.45	0.00
	(0.4)	(0.3)	(0.48)	(0.02)
Log(HH size)	1.4†††	-0.8†††	0.89	-0.05††
	(0.4)	(0.2)	(0.43)	(0.02)
College grad	1.2†††	-0.9†††	0.94†	-0.02
	(0.4)	(0.2)	(0.37)	(0.01)
Male	0.3	-0.1	0.54	0.02
	(0.3)	(0.2)	(0.34)	(0.01)
Religious attendance (0 to 5, 'Never' to 'More than once a week')	1.1†††	-0.4†††	0.66†††	-0.04†††
	(0.1)	(0.1)	(0.10)	(0.004)
Asian	-1.5†	0.2	-1.53	0.00
	(0.7)	(0.4)	(0.73)	(0.02)
Obs.	3,358	3,358	3,358	3,358

Notes: The dependent variables are constructed from each individual's responses to the 18 Baseline CQs. $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the intercept and slope, respectively, from the regression of respondent i's 18 Baseline CQ ratings onto the sample means of those 18 ratings. The sample is 3,358 Baseline respondents who passed quality control. Daggers signal false-discovery-rate (FDR) significance levels using the Benjamini-Hochberg procedure applied to the 29 *p*-values in each column separately (variables included in FDR correction also include additional race and employment status indicators, as well as indicators for region, day of week, political party, obesity, and population density; "Other" categories in race and employment status are excluded—we do not pose hypothesis tests for them); †††, ††, and † indicate significance at the 1-percent, 5-percent, and 10-percent levels, respectively. Indicators for political party, obesity, Black/African American race, residing in the South, and taking the survey on a Saturday are also significant. See the full set of results in Web Appendix Table J.5.

Table 4. MLE Estimates

σ_α	7.87 (0.14)
γ	59.90 (1.13)
σ_β	0.29 (0.01)
σ_ϵ	7.07 (0.63)
$\mu_{\ln\sigma_\eta}$	2.68 (0.03)
$\sigma_{\ln\sigma_\eta}$	0.31 (0.02)

Notes: The sample is 3,358 Baseline respondents who passed quality control. Standard errors in parentheses. See Web Appendix F.1 for the likelihood function and additional details for this MLE.

Table 5. Key Assumptions of Estimators

Estimand	Method of Moments	Semi-parametric	Comprehensive Maximum Likelihood
All moments		Distribution assumed for β_i	(a_i, β_i, w_{is}) jointly normal
$E(w_{is})$	Independence of β_i and w_{is}	$E(w_{is}\|\beta_i)$ polynomial in β_i	No additional assumptions needed
$Cov(x_i, w_{is})$	Assumptions of semi-parametric or MLE estimator for $E(w_{is})$ Zero co-skewness	$E(w_{is}x_i\|\beta_i)$ polynomial in β_i	$Cov(\beta_i, w_{is}\|x_i) = Cov(\beta_i, w_{is})$ (Implies zero co-skewness)
$Cov(w_{is}, w_{is'})$	Independence of β_i and $(w_{is}, w_{is'})$ Distribution assumed for β_i	$E(w_{is}\|\beta_i)$, $E(w_{is'}\|\beta_i)$, and $E(w_{is}w_{is'}\|\beta_i)$ polynomial in β_i	No additional assumptions needed
$Var(w_{is})$	Independence of β_i and w_{is} Distribution assumed for β_i SWB response error variances equal to those for calibration-questions	$E(w_{is}\|\beta_i)$, $E(w_{is}^2\|\beta_i)$ polynomial in β_i SWB response error variances equal to those of CQs	SWB response error variances equal to those of CQs

Notes: All estimators assume Assumptions 1′, 2, 3, 4, and equation (6). Greyed text indicates that use of the estimator for that purpose is discouraged.

Table 6. Life Satisfaction and "No Anxiety" Regression and General Scale-Use Adjustment

	Life Satisfaction					No Anxiety				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Demographics	No scale-use correction	MOM	Semi-parametric	Compre-hensive MLE	MMB (66.90)	No scale-use correction	MOM	Semi-parametric	Compre-hensive MLE	MMB (54.04)
Demeaned age/10	1.3††† (0.4)	1.7††† (0.4)	1.6††† (0.5)	1.6††† (0.4)	-0.4 (0.2)	3.6††† (0.5)	4.2††† (0.4)	3.9††† (0.5)	3.9††† (0.4)	-0.6††† (0.2)
Demeaned age^2/100	1.7††† (0.2)	1.7††† (0.2)	1.7††† (0.3)	1.6††† (0.2)	0.1 (0.1)	1.1††† (0.3)	0.9††† (0.3)	1.0††† (0.3)	0.9††† (0.3)	0.2† (0.1)
Log(HH income)	5.1††† (0.7)	5.1††† (0.6)	5.4††† (0.8)	5.3††† (0.6)	0.1 (0.3)	2.0†† (0.8)	2.4†† (0.8)	2.9††† (0.9)	2.7††† (0.8)	-0.5† (0.2)
Unemployed	-8.3††† (1.6)	-9.2††† (1.5)	-7.5††† (1.5)	-7.0††† (1.3)	1.0 (0.6)	-7.7††† (2.0)	-7.0††† (1.8)	-6.0††† (2.0)	-5.6††† (1.7)	-0.7 (0.5)
Employed part-time	-2.9† (1.3)	-2.8† (1.4)	-3.6† (1.6)	-2.3 (1.2)	-0.1 (0.6)	-5.9††† (1.5)	-5.1††† (1.5)	-6.2††† (1.5)	-4.6††† (1.4)	-0.8 (0.5)
Out of labor force/other	-6.0††† (1.4)	-5.6††† (1.5)	-5.4††† (1.7)	-4.5††† (1.3)	-0.3 (0.6)	-6.0††† (1.7)	-4.4†† (1.7)	-3.7 (1.9)	-4.3†† (1.6)	-1.6 (0.5)
Married, not separated	9.7††† (1.0)	8.9††† (1.0)	8.5††† (1.3)	8.6††† (1.0)	0.8 (0.4)	4.7††† (1.4)	3.1†† (1.3)	3.1† (1.5)	3.1†† (1.2)	1.6††† (0.4)
Ever divorced	2.3 (1.3)	2.2 (1.3)	1.4 (1.6)	2.1 (1.2)	0.1 (0.6)	-0.1 (1.7)	0.2 (1.7)	-0.9 (1.9)	0.5 (1.7)	-0.2 (0.5)
Have ≥1 child	4.1††† (1.0)	4.5††† (1.0)	4.6††† (1.3)	3.8††† (0.9)	-0.4 (0.6)	1.1 (1.2)	1.6 (1.2)	1.4 (1.4)	1.7 (1.1)	-0.5 (0.4)
Log(HH size)	-2.1† (1.1)	-2.7†† (1.1)	-1.3 (1.3)	-2.4††† (1.0)	0.5 (0.5)	-1.2 (1.2)	-2.4† (1.1)	-1.5 (1.4)	-2.2 (1.1)	1.2†† (0.4)
College grad	2.1† (1.0)	1.3 (0.9)	1.6 (1.2)	1.5 (0.9)	0.8 (0.4)	3.3††† (1.0)	2.2† (1.0)	2.5† (1.2)	2.3† (1.0)	1.1†† (0.4)
Male	0.1 (0.9)	-0.6 (0.9)	-0.3 (1.1)	0.1 (0.9)	0.7 (0.4)	5.7††† (1.0)	5.2††† (1.0)	4.8††† (1.2)	5.7††† (1.0)	0.4 (0.3)
Religious attendance (0 to 5, 'Never' to 'More than once a week')	1.9††† (0.2)	1.6††† (0.2)	1.5††† (0.3)	1.5††† (0.2)	0.4†† (0.1)	1.8††† (0.3)	0.9††† (0.3)	1.0†† (0.4)	1.0††† (0.3)	0.9††† (0.1)
Asian	0.0 (1.9)	1.5 (1.8)	1.7 (2.1)	0.4 (1.8)	-1.5 (0.8)	2.5 (2.0)	4.0† (1.9)	4.2 (2.2)	3.2 (1.9)	-1.5† (0.7)
Obs.	3,358	3,358	3,358	3,358	3,358	3,358	3,358	3,358	3,358	3,358

Notes: The sample is 3,358 Baseline respondents who passed quality control. Dependent variables for columns (5) and (10) are MMBs, matched to the semi-parametric estimates of Life Satisfaction (66.90) and No Anxiety (54.04) means, respectively. Daggers signal false-discovery-rate significance levels using the Benjamini-Hochberg procedure applied to the 29 *p*-values in each column separately. (See Table 3 notes for description of the FDR correction procedure and significance levels. Variables included in the FDR correction also include additional race and employment status indicators, as well as indicators for region, day of week, political party, obesity, and population density.) Indicators for political party, obesity, Black/African American race, residing in the West, high population density, and taking the survey on a Monday are also significant. See the full set of results in Web Appendix Table J.6.

Table 7. Factor Loadings

	(A) No scale-use correction			(B) After scale-use correction		
SWB	Factor 1	Factor 2	$E(r_{is})$	Factor 1	Factor 2	$E(w_{is})$
Satisfaction	0.88 (0.00)	-0.15 (0.02)	65.91 (0.38)	0.95 (0.00)	-0.25 (0.02)	66.90 (0.43)
Happiness	0.88 (0.00)	-0.14 (0.02)	66.51 (0.40)	0.96 (0.00)	-0.24 (0.01)	67.33 (0.45)
Worthwhileness	0.80 (0.01)	-0.10 (0.03)	68.87 (0.39)	0.91 (0.01)	-0.24 (0.03)	69.90 (0.43)
No Anxiety	0.68 (0.01)	-0.15 (0.03)	54.04 (0.51)	0.76 (0.02)	0.01 (0.08)	54.04 (0.60)
Ladder	0.80 (0.01)	-0.05 (0.02)	63.42 (0.37)	0.96 (0.01)	-0.11 (0.03)	64.06 (0.42)
Well-being of Your Family	0.80 (0.01)	0.14 (0.02)	74.14 (0.33)	0.96 (0.00)	-0.02 (0.03)	74.80 (0.37)
Family Happiness	0.79 (0.01)	0.04 (0.02)	72.62 (0.36)	0.96 (0.00)	-0.10 (0.03)	73.66 (0.41)
Physical Health	0.65 (0.01)	0.05 (0.02)	70.45 (0.34)	0.89 (0.01)	-0.02 (0.04)	71.21 (0.40)
Mental Health	0.80 (0.01)	-0.11 (0.02)	67.15 (0.45)	0.94 (0.01)	-0.17 (0.03)	67.88 (0.49)
Sense Of Purpose	0.77 (0.01)	-0.17 (0.02)	65.72 (0.41)	0.89 (0.01)	-0.25 (0.03)	66.86 (0.47)
Sense Of Control	0.83 (0.01)	-0.07 (0.01)	64.89 (0.41)	0.96 (0.00)	-0.13 (0.02)	65.64 (0.48)
Having People	0.66 (0.01)	0.10 (0.02)	70.51 (0.47)	0.79 (0.02)	-0.04 (0.04)	70.74 (0.52)
Not Lonely	0.67 (0.01)	-0.03 (0.02)	64.47 (0.55)	0.76 (0.02)	-0.08 (0.04)	64.21 (0.63)
No Anger	0.58 (0.01)	0.11 (0.03)	63.78 (0.44)	0.68 (0.02)	0.23 (0.06)	62.70 (0.51)
No Sadness	0.74 (0.01)	-0.09 (0.02)	58.20 (0.47)	0.87 (0.01)	0.00 (0.06)	57.75 (0.54)
No Stress	0.74 (0.01)	-0.15 (0.04)	52.03 (0.51)	0.82 (0.01)	0.03 (0.07)	52.26 (0.58)
No Worry	0.74 (0.01)	-0.20 (0.04)	51.23 (0.46)	0.82 (0.01)	-0.01 (0.08)	51.25 (0.52)
Good Person	0.57 (0.02)	0.13 (0.03)	79.84 (0.29)	0.86 (0.01)	0.00 (0.06)	81.30 (0.33)
Possibilities	0.75 (0.01)	-0.04 (0.02)	65.68 (0.44)	0.92 (0.01)	-0.14 (0.03)	66.60 (0.48)
Time	0.63 (0.01)	0.08 (0.02)	66.65 (0.40)	0.80 (0.02)	0.06 (0.04)	66.72 (0.45)
Social Status	0.83 (0.01)	-0.20 (0.02)	58.49 (0.46)	0.94 (0.01)	-0.23 (0.03)	59.55 (0.51)
Safety	0.56 (0.01)	0.50 (0.02)	79.43 (0.29)	0.84 (0.01)	0.45 (0.04)	79.82 (0.36)
Financial Support	0.72 (0.01)	-0.03 (0.02)	63.79 (0.45)	0.79 (0.01)	0.03 (0.03)	65.08 (0.55)
Not Unemployed	0.54 (0.01)	0.12 (0.02)	63.10 (0.52)	0.54 (0.02)	0.28 (0.04)	62.97 (0.61)
Eat	0.43 (0.02)	0.47 (0.02)	86.01 (0.31)	0.76 (0.02)	0.37 (0.06)	86.66 (0.39)
Housing Comfort	0.66 (0.01)	0.22 (0.02)	76.78 (0.38)	0.91 (0.01)	0.09 (0.05)	77.59 (0.41)
Enjoyment	0.88 (0.00)	-0.13 (0.02)	68.41 (0.38)	0.95 (0.00)	-0.25 (0.01)	69.29 (0.43)
Knowledge Skills	0.62 (0.01)	0.07 (0.02)	73.47 (0.29)	0.88 (0.01)	-0.05 (0.05)	74.83 (0.32)
Local Safety	0.36 (0.02)	0.58 (0.03)	74.91 (0.43)	0.52 (0.03)	0.79 (0.02)	74.38 (0.52)
Local Air	0.35 (0.02)	0.38 (0.03)	68.17 (0.40)	0.44 (0.03)	0.66 (0.04)	67.57 (0.51)
Citizen Influence	0.42 (0.01)	-0.20 (0.03)	45.66 (0.42)	0.43 (0.02)	0.09 (0.06)	46.66 (0.48)
Citizen Trust	0.53 (0.01)	-0.06 (0.03)	51.28 (0.35)	0.57 (0.02)	0.25 (0.06)	51.12 (0.43)
Culture Honor	0.46 (0.02)	0.12 (0.03)	65.46 (0.35)	0.53 (0.02)	0.23 (0.05)	65.59 (0.45)
Variance Explained (Proportion)	47.17% (0.55%)	4.24% (0.15%)		67.36% (0.79%)	6.39% (0.27%)	

Notes: The sample is 3,358 Baseline respondents who passed quality control. Standard errors in parentheses. 10 factors were extracted in the factor analyses, the first two of which are shown above. Panel A is based on the correlation matrix of the raw ratings of 33 Baseline SWBs, Panel B is based on the correlation matrix estimated by the comprehensive MLE using 33 Baseline SWBs. Raw SWB means and semi-parametric estimates of means of the common-scale SWBs are also shown. Standard errors in parentheses. The full text of each aspect of well-being can be found in Appendix A.

Table 8. SWB Question Variance Estimates

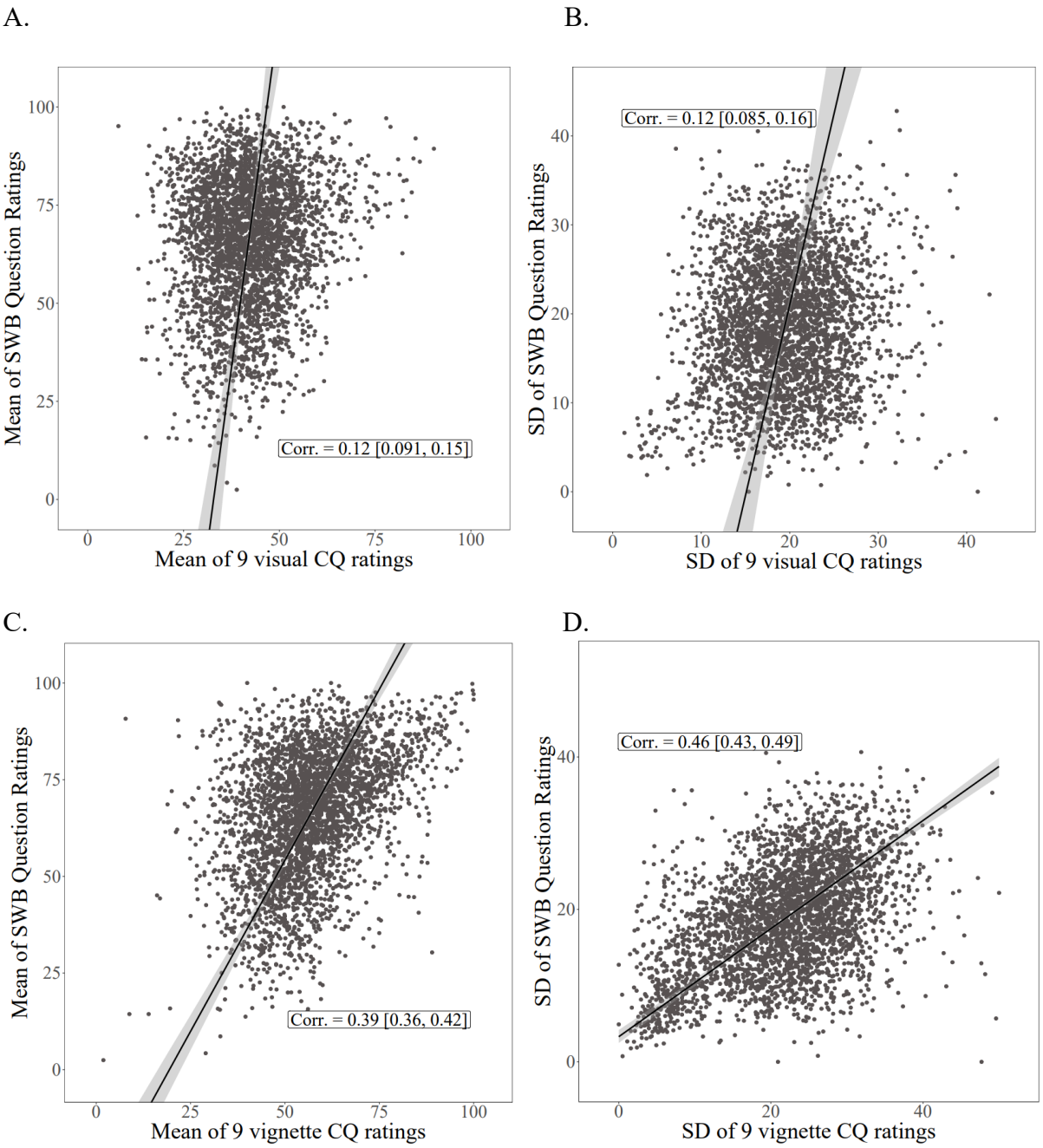
	(1)	(2)	(3)
SWB Question	No scale-use correction	Semi-parametric	Comprehensive MLE
Life Satisfaction	661.8 (14.5)	318.8 (26.6)	273.6 (11.1)
Happiness	605.6 (14.0)	259.9 (25.2)	263.6 (9.5)
Worthwhileness	554.6 (14.7)	199.8 (27.2)	222.9 (9.9)
No Anxiety	830.0 (13.0)	446.4 (26.0)	370.4 (12.4)

Notes: The sample is 3,358 Baseline respondents who passed quality control. Standard errors in parentheses.

Figures

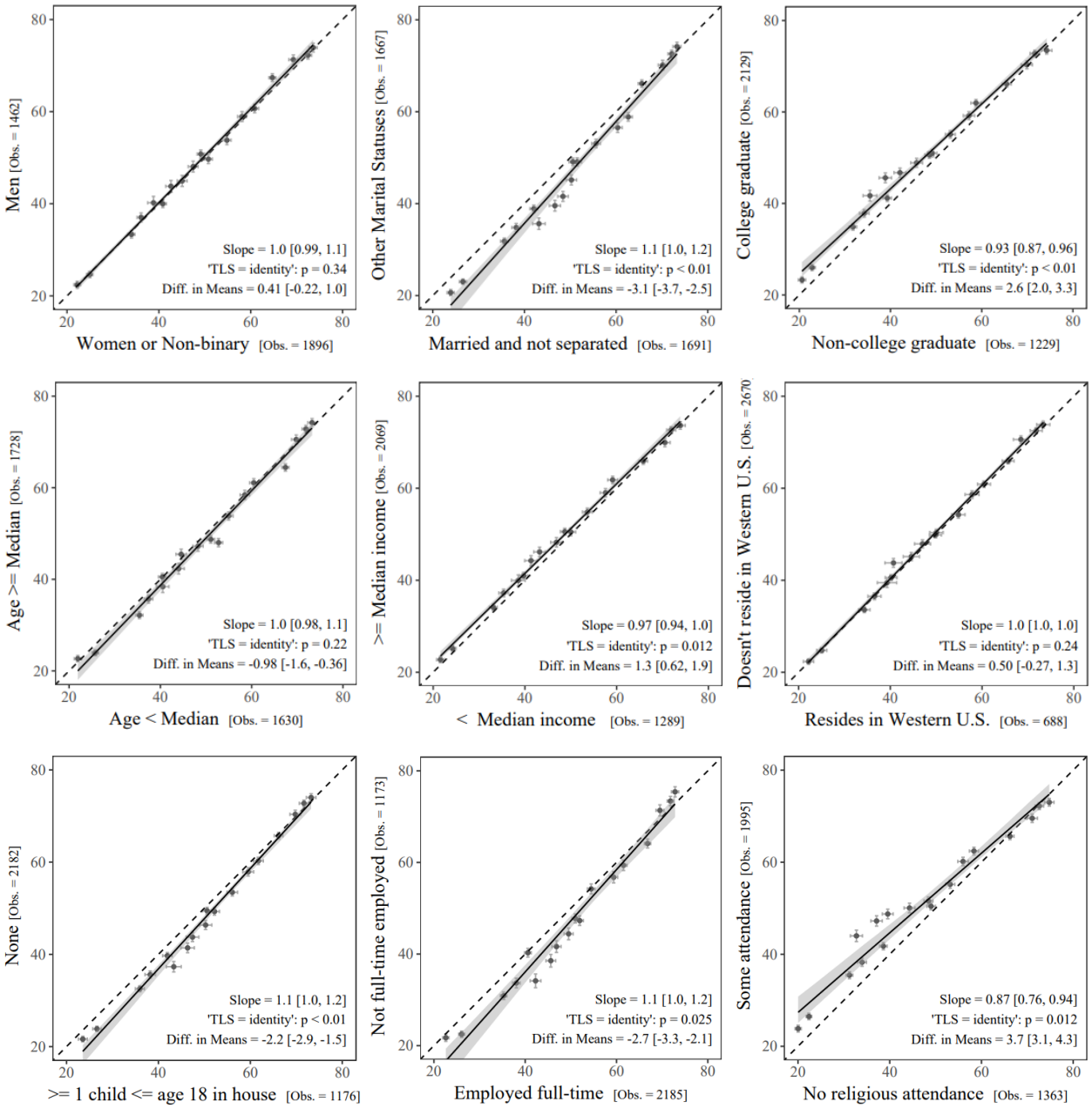
Figures 1-3 are screenshots in the paper's text.

Figure 4. Relationships Between CQ- and SWB-Rating Means and Standard Deviations



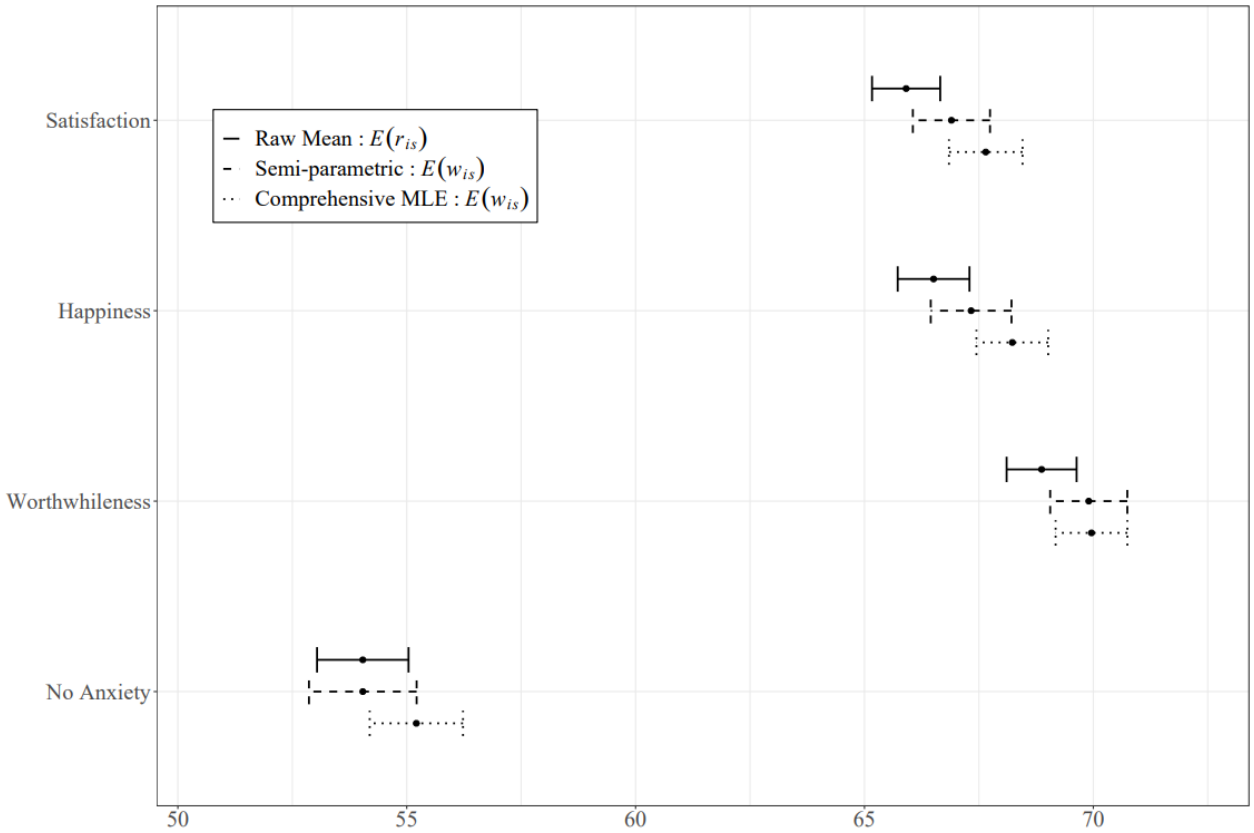
Notes: For panels A and B, each point corresponds to a respondent and reflects her mean or standard deviation on the 9 Baseline survey visual CQ ratings (*x*-axis) and her mean or standard deviation on the 33 Baseline survey SWB question ratings (*y*-axis). The sample is 3,358 respondents who passed quality control and completed our main demographic questions. The black line is the total least squares regression line with 95% confidence region in gray. Panels C and D are the same except they use the 9 Baseline vignette CQ ratings (*x*-axis).

Figure 5. Translation Functions Across Demographic Groups



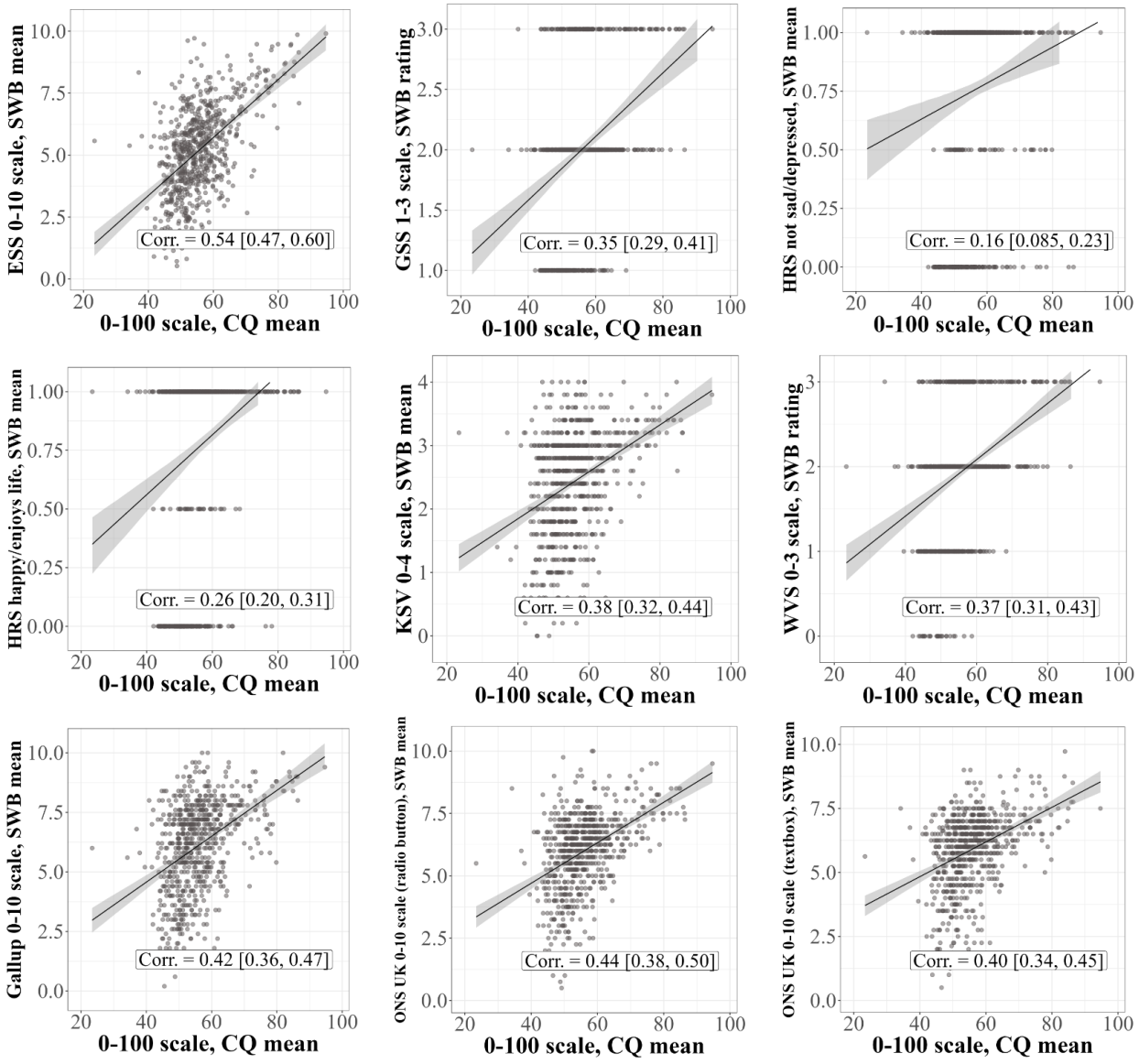
Notes: Each point plots the mean for a single CQ for both groups, with 95% CIs (capped bars), for 18 Baseline CQs. The sample is 3,358 respondents who passed quality control. Dashed line is the 45-degree line. The black line is the total least squares (TLS) regression line with 95% confidence region in gray. 'TLS = identity' *p*-value comes from the F-statistic: (Sum of Squared Orthogonal Errors, TLS line) / (Sum of Squared Orthogonal Errors, identity line). Intuitively, this F-test tests the joint difference of the TLS (intercept, slope) from (0,1). Difference in means calculated as mean of *y*-axis CQ ratings minus mean of *x*-axis CQ ratings. Correlations (not reported in the plots) range from 0.95 to 1.00. Square brackets contain 95% confidence intervals.

Figure 6. Mean SWB Before and After General-Scale-Use Correction



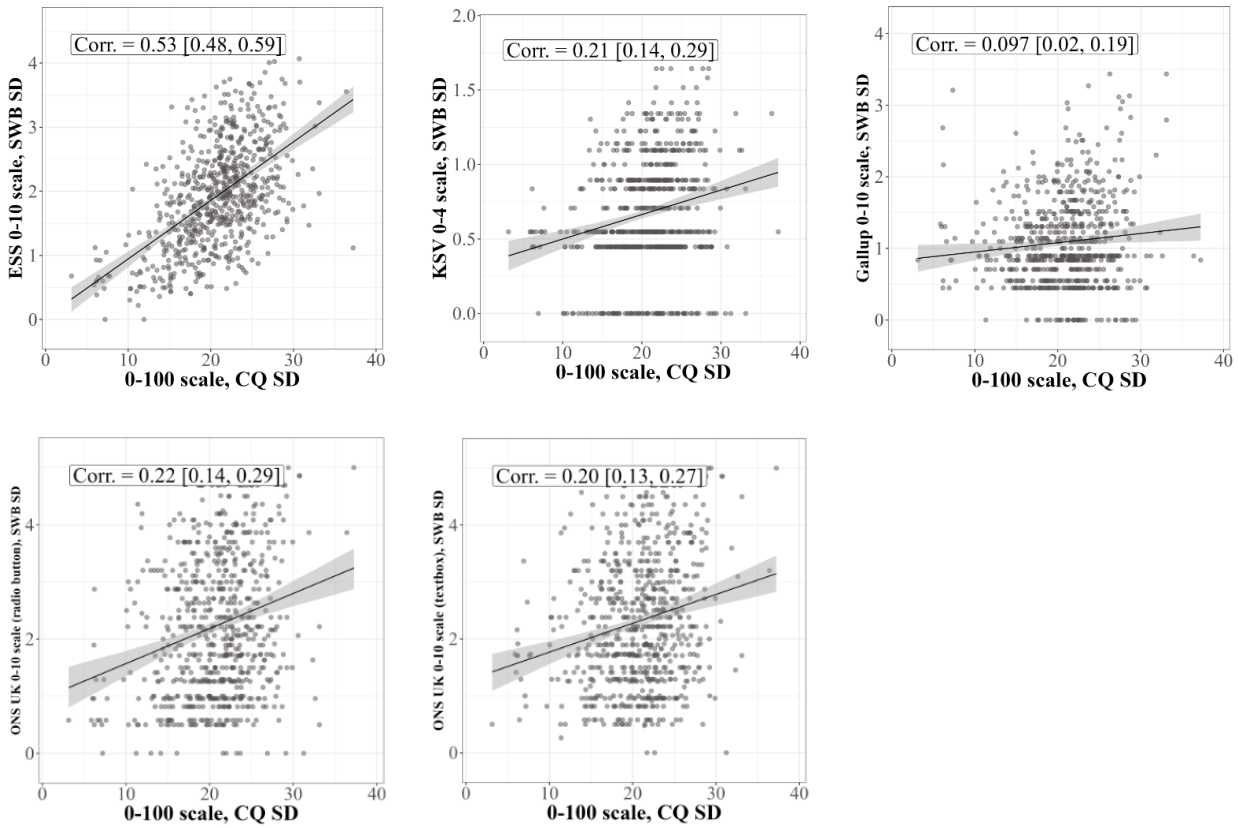
Notes: The intervals are 95% confidence intervals. For satisfaction, happiness, worthwhileness, and "no anxiety," respectively, $Cov(w_{is}, \beta_i)$ as estimated by the comprehensive MLE is: -0.86, -0.84, -0.66, and -0.57. We can also estimate $Cov(w_{is}, \beta_i)$ indirectly as $E(r_{is}) - E(w_{is})$, with $E(r_{is})$ estimated by the sample mean and $E(w_{is})$ by the semi-parametric estimator; these estimates are, respectively: -0.98, -0.82, -1.03, and 0.00. The sample is 3,358 Baseline respondents who passed quality control.

Figure 7. CQs (0-100 Scale) vs. SWB Ratings on Alternative Scales



Notes: Black curve: Fitted line from total least squares (TLS) regression of dependent variable on the individual-level average of 388 CQ responses asked on 0-100 scale. Each observation is an individual (Sample consists of 701 respondents who completed all relevant CQs). Each dependent variable is the respondent's SWB rating, or mean rating across SWB questions, elicited on a scale used in an existing dataset (not 0-100). This includes the ESS 0-10 scale (21 questions), GSS 1-3 scale (1 question), HRS yes/no scale (split out by 2 positively coded and 2 reverse-coded questions), KSV 0-4 scale (5 questions), WVS 0-3 scale (1 question), Gallup 0-10 scale (5 questions), and ONS UK 0-10 scale (4 questions, in radio button or textbox format). The full list of alternative-scale SWB questions is included in Web Appendix K4. CQs used are 388 Bottomless CQs; see Web Appendix A.4.iii for details.

Figure 8. CQ SDs (0-100 Scale) vs. SWB SDs on Alternative Scales



Notes: Black curve: Fitted line from total least squares (TLS) regression of dependent variable on the individual-level standard deviation of CQ responses asked on 0-100 scale. Each observation is an individual (of 701 respondents who completed all relevant CQs). Each dependent variable is the respondent's standard deviation of SWB questions, elicited on a scale used in an existing dataset (not 0-100). Alternative scales are limited to those for which we have multiple questions, and multiple response options for each question: the ESS 0-10 scale (21 questions), KSV 0-4 scale (5 questions), Gallup 0-10 scale (5 questions), and ONS UK 0-10 scale (4 questions, in radio button or textbox format). The full list of alternative-scale SWB questions is included in Web Appendix K4. CQs used are 388 Bottomless CQs; see Web Appendix A.4.iii for details.