The Challenge of Measuring National Well-Being

By DANIEL J. BENJAMIN, KRISTEN B. COOPER, ORI HEFFETZ, AND MILES KIMBALL

* Benjamin: Center for Economic and Social Research, University of Southern California, 635 Downey Way, Suite 312, Dauterive Hall, Los Angeles, CA 90089 (email: djbenjam@usc.edu). Cooper: Department of Economics and Business, Gordon College, 255 Grapevine Road, Wenham, MA 01984 (email: kristen.cooper@gordon.edu). Heffetz: Samuel Curtis Johnson Graduate School of Management, Cornell University, 324 Sage Hall, Ithaca, NY 14853 and Department of Economics, Hebrew University of Jerusalem (email: oh33@cornell.edu). Kimball: Department of Economics, University of Colorado Boulder, 256 UCB, Boulder, CO, 80309-0256 (email: miles.kimball@colorado.edu). We are grateful for NIH/NIA grants R01-AG040787 to the University of Michigan, and R01-AG051903 to the University of Southern California, and to the Michigan Institute for Teaching and Research in Economics for financial support, and to Tuan Anh Viet Nguyen, Rebecca Royer, and Robbie Strom for outstanding research assistance.

Many in both government and academia are showing renewed interest in developing new measures of national well-being (NWB) for guiding policy and for moving "beyond GDP" and its focus on market goods when tracking welfare. But how should NWB be conceptualized in theory, and how could it be measured in practice? These questions should be approached by economists with the same level of care that has been taken in the theoretical and practical development of GDP.

In this short paper, we focus on one conceptual framework (Benjamin, Heffetz, Kimball, and Szembrot, 2014; hereafter BHKS), which uses self-reported responses to subjective well-being (SWB) and stated preference (SP) survey questions to construct an index of well-being. We briefly review the framework and highlight challenges in the first two steps a government agency would need to take before conducting the SWB and SP surveys: (1) formulating a list of aspects of well-being that is theoretically valid and can be measured accurately via surveys; and (2) choosing and interpreting the surveys' response scales.

The BHKS framework focuses on an individual-level, personal well-being (PWB) index. For aggregating PWB indices into a measure of NWB, we are skeptical of common approaches such as averaging responses to survey questions across individuals. In addition to the usual doubts about interpersonal comparability of utility levels, different individuals may use the response scales to SWB questions differently. We believe methods for interpersonal aggregation of ordinal utilities (for example, building on money-metric utilities as in Fleurbaey and Blanchet, 2013) are most promising. The present paper is focused on PWB and not on the problem of interpersonal aggregation.

I. Theoretical Framework

A consensus is emerging that well-being is multi-dimensional (Stiglitz, Sen, and Fitoussi, 2009), and evidence suggests that it is unlikely to be fully captured with a single happiness or life-satisfaction question (e.g., Benjamin, Heffetz, Kimball, and Rees-Jones, 2012). Hence the idea to separately measure all dimensions (or aspects) of well-being and combine them, with appropriate weights, into an index. (The idea of being content with a "dashboard" of different indicators ultimately founders on the need to make overall evaluations and decisions in cases where one indicator goes up while another goes down.)

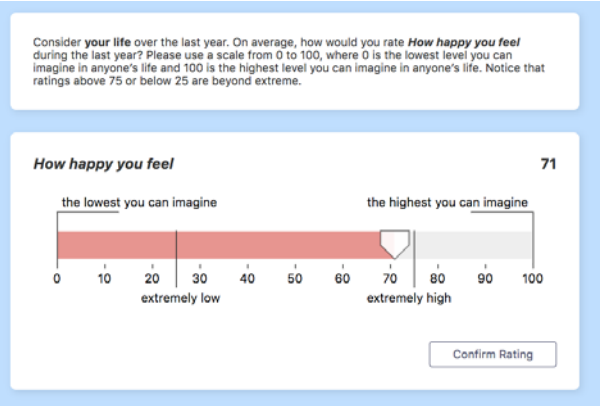
To this end, in BHKS we proposed a simple framework that is analogous to the theory behind the measurement of aggregate consumption. We proposed replacing the typical consumption vector $\boldsymbol{c}$ in individuals' utility $u(\boldsymbol{c})$ with a vector of "fundamental aspects of well-being" $\boldsymbol{w}$, where fundamental aspects include anything about the state of the world that matters to an individual's choices. While objective measures might eventually be available for many of these aspects (e.g., biomarkers for certain health aspects and even emotional states), we assumed that the levels of $\boldsymbol{w}$ would at present be measured with SWB surveys.

A traditional aggregate consumption index, $\sum_{m=1}^{M} \overline{p_m}\, c_m$, weights each good's consumption, c_m, by its price held fixed at a baseline level, $\overline{p_m}$. Assuming consumption is chosen optimally, small changes in the index approximate changes in utility (up to a multiplicative constant): $\sum_{m=1}^{M} \overline{p_m}\, \Delta c_m \propto \sum_{m=1}^{M} \frac{\partial u(c)}{\partial c_m} \Delta c_m \approx \Delta u$.

In our framework, the M consumption goods are replaced by J fundamental aspects of well-being. Because the aspects are not traded in markets, prices are unavailable and different individuals may have different marginal rates of substitution (MRSs) across the aspects. Nevertheless, an analogous PWB index can be constructed using the aspects' MRSs (relative to an arbitrary numeraire aspect) for the individual as weights: $\sum_{j=1}^{J} \overline{\frac{\partial u(\boldsymbol{w})}{\partial w_j}} w_j$. Small changes in this PWB index track changes in an individual's utility. Although utility is ordinal, we hereafter use "marginal utility"—MU—to mean "MRS relative to a numeraire aspect."

Figures 1a and 1b show examples of potential web-survey questions for measuring the levels of two aspects of well-being: happiness and meaningfulness. Figure 2 shows an example survey question for

measuring the MRS between these aspects. Both are taken from an ongoing project of ours that attempts a large-scale implementation of the BHKS framework.



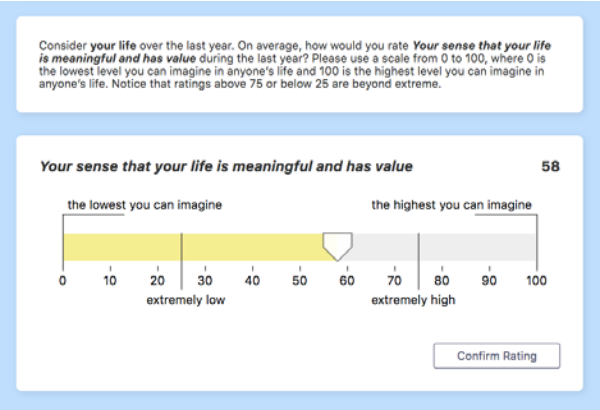


Figure 1. Two sample SWB survey questions.

Figure 2. A sample SP survey question.

II. Challenge #1: Formulating the List of Aspects

Within our theoretical framework, any list of J aspects of well-being can be used for the index as long as it satisfies two properties: *comprehensiveness* and *non-overlappingness*. From a practical perspective, the aspect list must also satisfy the requirements of *explicability* and *accessibility*.

Comprehensiveness means that the list includes all aspects of well-being that matter to the individual. If the well-being index is used only to evaluate specific policies (rather than to track overall well-being), then the list should include all aspects that could be affected by those policies.

Non-overlappingness means that the aspects of well-being enter preferences as conceptually distinct. For GDP, if the value of restaurant meals and the value of alcoholic beverages were independently added up, then the value of alcohol consumed in restaurants would be double-counted. Similarly, a PWB index will suffer from a double-counting problem if an ultimate object of desire that enters preferences once is counted more than once. For example, if "how much you like your life" and "how much you enjoy your life" appear as two different aspects on our list

and get equally high marginal utilities, but they mean essentially the same thing to people and enter their preferences once—then they should be counted only once to avoid double-counting. The solution in the case of GDP is to define the expenditure categories so that they are conceptually distinct. In BHKS we proposed, but did not test, a method of detecting overlap between aspects that could, in principle, be applied to prune the list. Another approach worth exploring is to measure the extent of overlap and then adjust for it.

Explicability requires that the wording of the aspects is understood by respondents as intended, and in a reasonably uniform way across respondents. Making sure the aspects are explicable is partly a matter of using understandable vocabulary. Explicability also requires avoiding ambiguous concepts and specifying things that are potentially ambiguous. In ongoing work with Jakina Debnam and Marc Fleurbaey, when we ask survey respondents to answer standard SWB questions and then ask what they thought about when answering them, we find substantial heterogeneity. We believe such research is needed, and that survey questions should be refined until they are understood appropriately by respondents.

Accessibility means that respondents can accurately introspect about their own level of an aspect of well-being and about how much they care about it. This is analogous to the implicit assumption for constructing GDP of access to accurate production and price data. If respondents are systematically biased in these assessments, then a survey-based approach to measuring the levels and the MUs of aspects of well-being will generate biased conclusions.

Part of accessibility is that an aspect of well-being must (as nearly as possible) either (a) be an ultimate object of desire that would be desired even if it did not produce any other good result outside of itself, or (b) have a crystal-clear necessary and sufficient causal relationship with ultimate objects of desire.

Social desirability biases are closely related to the issue of accessibility: it is hard to think clearly about a desire that is seen by society as reprehensible. However, we suspect diligent search will usually reveal *some* way of talking about an individual's *ultimate* objects of desire that is socially acceptable enough that an appropriately worded survey question can measure the desire reasonably well. (For example, one can ask about "you feeling powerful" rather than "you having power over other people.") If not, perhaps the desire is so genuinely reprehensible from a fairly

objective moral standpoint that it is just as well if respondents self-launder that desire from their reports.

Of course, even when aspects are accessible, aspects' levels and MRSs may not be reported accurately. Survey-based measures of well-being are known to be sensitive to contextual details such as question order, to be overly sensitive to how a respondent feels at a given moment, and to depend heavily on recent changes in an individual's life in comparison to permanent, important changes that have become old news.

Another open question is how to decide at what level of generality to specify the aspects, e.g., "your health" vs. components of health. We conjecture that it matters because of a focusing bias: the sum of the MUs for the aspect's components may exceed the MUs for the aspect considered holistically. We suspect that a reasonable rule of thumb is to try to specify aspects such that they have similarly sized MUs, but this issue and potential solutions requires study.

The challenge of formulating the aspect list poses several trade-offs. Our strategy in BHKS was to try to construct a fairly comprehensive list of aspects by scouring the economics, psychology, and philosophy literatures for lists of what matters to people, and we studied 113 aspects that resulted from this effort. Since BHKS, we have further expanded our list to over 2000 (!) potential aspects of well-being. Our strategy is to begin with an exhaustive list and then learn through empirical testing which potential aspects have low enough MUs, or are duplicative enough that little is lost by omitting them. (Such a long, detailed list of aspects likely dramatically increases the extent of overlap. Moreover, one would wish the aspect list could be short enough that each survey respondent is willing to answer questions about all aspects in the list. A list longer than that necessitates pooling data across respondents and making the auxiliary assumptions that identifiable groups of respondents have homogeneous preferences and are affected identically by events.

To date, we have relied on introspection to search for accessible attributes. But our work to date has not tested the important assumptions of explicability and accessibility, and we would like to see a more systematic examination of these necessary properties.

III. Challenge #2: Choosing the Response Scales

Perhaps surprisingly, from the perspective of the theory, it does not matter what response scale is assigned to an aspect, as long as it is the same response scale in the SWB survey

and the SP survey (as it is in Figures 1 and 2) and the respondent uses the response scale the same way in both surveys. In the SWB survey the level could be elicited on a 0-10 scale, a 0-100 scale, an amount of smiley-ness scale, whatever—as long as the SP survey accurately elicits preferences about changes on that scale. In terms of the model, that is because in the index, $\sum_{j=1}^{J} \overline{\frac{\partial u(\boldsymbol{w})}{\partial w_j}} w_j$, the units of w_j relative to any other aspect w_k cancel out when each is multiplied by its MU.

A key psychological assumption of our framework, however, is that respondents can give accurate self-reports on the SWB and SP surveys. As noted above, choosing aspects that satisfy accessibility is one component of making sure this assumption holds. The other key component is choosing a response scale that does not hinder respondents from accurate self-reports. For example, if respondents are uncomfortable with numbers, then eliciting a numerical (say, 0-100) response may not only lead to noisier responses but also lead to exaggeration of marginal rates of substitution: if, for example, many respondents treat an 8 to 1 tradeoff as if it were a 3 to 1 tradeoff (because of only partial comprehension of the numbers), for those respondents it would make what is really an MRS of 3 look to the econometrician as if it were an MRS of 8.

Although (as noted above) the scale a respondent uses does not matter for the validity of a PWB index, *shifts* in the scale over time are a serious concern. If the observed Δw_j reflects a shift in scale use rather than an actual change in the aspect level, then the resulting change in the index will not accurately reflect a change in well-being.

Systematically studying such possible shifts in scale use and developing ways of correcting for them is a high priority. One approach would be to find aspects of well-being that with evolving technology can be measured biometrically and compare them to survey measures over time. Another under-tested approach that merits further exploration would be to have individuals rate aspects of well-being not only in their own lives but for a set of "vignettes" describing a hypothetical person's life (e.g., Kapteyn, Smith, and van Soest, 2012). If the ratings for an unchanging set of vignettes shift systematically from one survey wave to another, a shift in scale use may be one possible explanation to further investigate. Such a study would be a major effort, but, we think, a worthwhile one.

One reason a respondent might shift scale use is to deal with the ceiling on a scale. This issue has received too little attention. Scales can be designed to reduce (though not

eliminate) top-coding issues by, say, labeling the top of the scale "extremely happy" rather than "very happy." Figure 1 shows our approach of labeling the top of the scale "the highest you can imagine in anyone's life." It remains to be investigated how effective this labeling scheme is in reducing top-coding.

IV. Discussion

Given space constraints, we have focused on the issues that seem most pressing for governments to address before they can begin collecting data that will eventually be used for constructing well-being indices. Yet we are concerned about many other issues. We briefly mention two that seem especially important but that we are not as far along in thinking through.

First, in the preamble of the SP survey questions we have explored (see Figure 2), we ask respondents to imagine that a few aspects of well-being change while all others are held constant. A potential problem is what we call *irrepressible imputation*: when one aspect is varied, respondents may impute variation in a related aspect, in spite of explicit instructions not to do so. For example, when asked to imagine "life is meaningful and has value" increases, respondents might think that "how happy you feel" also increases. Such imputation might occur either because the respondent believes that one causes the other or because they are highly correlated in everyday experience. If such imputation occurs, then we, the econometricians, will obtain a biased estimate of the aspect's MU. A first step toward understanding how widespread this might be is to study correlations and "production-function" relationships.

Second, if policy-makers desire to assess both objective and subjective dimensions of well-being (as in Stiglitz, Sen and Fitoussi, 2009, or anything borrowing from the capabilities approach), how can *objective measures* of aspects of well-being be best incorporated into a PWB index? We have already discussed some of the problems with using subjective measures, which could make objective measures attractive. But bringing in aspects that have an objective scale introduces a thorny issue: respondents must correctly understand what the objective units mean for what the individual cares about when evaluating tradeoffs involving an objective aspect. We hope to begin investigating MRS measurement between subjectively-scaled and objectively-scaled aspects by looking at tradeoffs of SWB with money and with time.

While we share the enthusiasm of many in government and academia for NWB measurement, and think there is a promising

roadmap, we agree with the conclusion of recent reports such as Stone and Mackie (2013) that not all obstacles have yet been overcome. Finding ways around the remaining obstacles seems to us an exciting and important research agenda.

REFERENCES

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2012. "What Do You Think Would Make You Happier? What Do You Think You Would Choose?" In American Economic Review, 102(5): 2083–2110.

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

Fleurbaey, Marc, and Didier Blanchet. 2013. *Beyond GDP: Measuring Welfare and Assessing Sustainability*. Oxford University Press.

Kapteyn, Arie, James P. Smith and Arthur van Soest. 2012. "Anchoring Vignettes and Response Consistency" In International Comparisons of Well-Being, eds. E. Diener, D. Kahneman, and J. Helliwell. Oxford University Press.

Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Commission on the Measurement of Economic Performance and Social Progress. www.stiglitz-sen- toussi.fr.

Stone, Arthur A., and Christopher Mackie, eds. 2014. *Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience*. National Academies Press.