How should Scale-Norming inform the Measurement of Subjective Well-Being for Policy?

Subjective well-being (SWB) is increasingly engaging the interest and enthusiasm of policymakers who see measures of life-satisfaction and happiness as a more direct and relevant guide for policy than standard economic metrics[1]. This has spurred lines of inquiry and criticism ranging from political philosophy to experimental psychology. How should distinct SWB constructs be combined or weighed?[2] How should SWB relate to well-being more broadly?[3] What cognitive mechanisms underly self-reports of life-satisfaction and affect?[4] Lastly, how should measures of SWB inform policy?[5] The present discussion addresses the underappreciated phenomenon of *scale-norming*—where the scale used for self-reported SWB changes between measurements. Despite its critical and far-reaching significance, scale-norming has gone almost entirely unexplored in the existing literature[6].

Results derived from existing measures of SWB often confound common sense and suggest radical departures in policy from the status quo. Notably, dramatic positive or negative changes to a person's objective circumstances often appear to have surprisingly little lasting effect on a person's SWB. This is normally explained in terms of hedonic 'adaptation'—where individuals become habituated to new circumstances over time and return to a 'baseline' level of happiness or life-satisfaction by updating their ambitions and expectations. To the extent adaptation occurs, unexpectedly small changes in measured SWB following changes in objective circumstances reflect real properties of, and changes in, the SWB constructs themselves. By contrast, scale-norming explains these measured effects as artefacts of the evaluation and measurement process, potentially *concealing* real changes in SWB.

In the general case, scale-norming occurs when a quantitative measure yields different absolute values relative to its target construct between or across measurements. In the present context: *scale-norming occurs just in case the criteria[7] for the same reported*

---

[1] Layard (2020)
[2] Haybron (2007), Busseri et al. (2010)
[3] Parfit (1986) (Appendix I)
[4] Perez-Truglia (2012)
[5] Dolan and Metcalfe (2012), Plant (2019)
[6] Exceptions include Fabian (2019), Lacey et al. (2008), and McClimans et al. (2012).
[7] Where 'criteria' encompasses both psychological and measurement standards.

SWB change relative to the actual or underlying SWB, across time or between groups. For those people who undergo dramatic shifts in objective circumstances, such as economics migrants or accident victims, a self-reported life-satisfaction of 7/10 in one year can indicate a non-negligibly higher or lower underlying life-satisfaction compared to the same 7/10 score some years later. Though scale-norming threatens to undermine the validity of existing self-report SWB measures by obscuring changes in the underlying SWB constructs, it should not be viewed as an *intrinsic* defect—as I shall describe, a person might scale-norm for a range of entirely legitimate reasons. Scale-norming is a only a *problem* within certain measurement contexts.

In addition to being both conceptually coherent and empirically plausible, scale-norming *matters*—both for policy and the social science that informs it. Failing to account for scale-norming may undermine comparisons across populations, individuals, and times; both cardinal and ordinal. I will argue that scale-norming is properly distinct from other phenomena such as adaptation; its distinctiveness manifesting in a diversity of causal mechanisms and characteristic features. In light of these features, I will sketch methods for identifying and accounting for scale-norming in a way that preserves the ability of suitably adjusted SWB measures to guide policy. Although I will not seek to estimate the overall extent and severity of scale-norming in measures of SWB, I hope this discussion can fill in some of the theoretical groundwork for further empirical investigation.

Existing Measures

To appreciate when scale-norming arises, it is important to have a clear impression of the measures used. Varieties of SWB are distinguished between the cognitive and evaluative aspect of life-satisfaction on one hand, and affective or emotional components (positive and negative) on the other. Life-satisfaction and affect are distinct constructs[8], and must be treated as such. While both aspects are extensively measured in survey questionnaires, large-scale policy proposals predominantly rely on measures of life-satisfaction—precisely the measure where scale-norming is plausibly the most extensive and severe. I will primarily focus on life-satisfaction measures for this reason.

Ed Diener's popular 'Satisfaction with Life Scale' (SWLS) asks participants to respond to a series of statements on a numerical scale from '1 – Strongly Disagree' to '7 – Strongly agree'. The statements include "In most ways my life is close to my ideal.", and, "If I could live my life over, I would change almost nothing.". An overall score is generated which ranges from 0-7.

[8] E.g. Kahneman and Deaton (2010).

Similarly, the World Happiness Report[9] references data from the following question in the World Gallup Poll:

> *"Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"*

This is known as the 'Cantril Self Anchoring Scale', or 'Cantril Ladder'. The derived measure for a population simply takes an unweighted numerical average of all respondents.

Though the measures yield a single number, the wording of the questions is often imprecise and open to a wide range of interpretations. For instance, the intended referents of 'best possible life' and 'worst possible life' are left ambiguous. Defenders of existing unadjusted life-satisfaction measures such as Alexandrova (2005) and Raibley (2010) respond that such openness and ambiguity is a desirable feature, because it avoids prescribing in advance what counts as a legitimate conception of the 'good life'. This, they argue, is what makes subjective well-being *subjective*. However, this same ambiguity allows for a range of consistent scales—leading to the possibility of scale-norming.

This hints at the conditions that make scale-norming a likelihood in the general case, beyond SWB. I suggest two necessary (but not sufficient) features. Firstly, ambiguity or flexibility about what constitutes a legitimate or correct response given some state of the target construct. For instance, a self-report of the average *units* of alcohol I drink per week is not vulnerable to scale-norming because the scale used is practically unambiguous; admitting little to no ambiguity in what constitutes a 'correct' response. For scale-norming to enter as a possibility, quantitative answers might be replaced with phrases like 'a moderate amount', 'a significant amount' etc. On the other hand, the measure must also have *some* plausible external accuracy conditions, because scale-norming is only coherent when scales can be compared according to some shared yardstick.

Scale-Norming versus Adaptation

In order to contrast scale-norming with hedonic adaptation, I will first describe a simple illustrative thought experiment. Consider two islands, X and Y. They are totally isolated but nearby populations of a similar size and demographic. The inhabitants of both

[9] Heliwell et al. (2020)

islands live relatively comfortable lives and are content with their circumstances. Neither must worry about ill health or being unable to afford necessities. However, Island Y is materially better-off than Island X: its inhabitants enjoy a more diverse array of better-quality luxuries; they have a higher GDP per capita, lower unemployment, and so on. A joint commission of scientists from X and Y decide to conduct a subjective well-being survey. Having distributed surveys to both islands, they find that islands X and Y both enjoy reasonably high levels of life-satisfaction. In fact, the average SWB scores for Island X are marginally higher than Island Y. Nonetheless, after word of the research gets out, a representative group of islanders from X decides to emigrate to Y—perhaps spurred by the economic prospects. The group from X decide to stay, preferring life on Y to life on X. A year later, the scientists ask that group to report on their subjective well-being using the same survey questions they originally distributed. To their surprise, the SWB scores for the migrants from X to Y have not changed. The scientists conclude that the islanders are no happier on Y than X. They elaborate on their finding in research journals, to the press, and to policymakers: they argue that material wealth is less important for SWB than we thought; and people adapt to their improved living conditions rapidly and comprehensively. On these grounds, they continue, we have reason to deprioritise redistributing money and goods to the materially worse-off relative to initiatives which deliver more tangible improvements to measured SWB. And from a global perspective, perhaps migration improves the lots of migrants by less than we imagined.

If the scientists are correct in assuming their measure of SWB accurately tracked real differences and similarities in SWB between islands and across time, then their claims about adaptation would be justified. The measurements would warrant the conclusions that the inhabitants of X are roughly as happy (even a little happier) than the inhabitants of Y, and that migrants from X to Y do not in fact achieve the lasting improvement in SWB than their strong preferences for life in Y over X suggested. If preferences for migrating to Y from X are interpreted as *predictions* of future self-reports of SWB, then they are grounded on false hopes. 100% of the difference between expected improvements in SWB and the measured changes are attributed on this view to hedonic *adaptation*. More precisely, I define adaptation as (potentially unexpected or surprising) *reversion to some stable level of life satisfaction or affect following significant positive or negative change in objective circumstances.*

However, others disagree with the assumption that their unadjusted measures of SWB accurately tracked real differences in SWB. They raise the alternative explanation of scale-norming. They conjecture that inhabitants of Y are in fact happier than the inhabitants of X, and that migrants from X to Y do in fact enjoy an improvement in subjective well-being in moving from X to Y; but that these differences are confounded by corresponding changes in the *criteria* for the same numerical rating of SWB relative to actual or underlying SWB. Suppose the migrants initially gave an average score of 8/10. Although the migrants *in fact* became happier from migrating to Y, their understanding of what kind of subjective states qualify for a score of 8.0 is *also* increased. So, though they might score a 9.0/10 on their original scale for SWB, their

new score is modified downward to 8.0, resulting in the false appearance of no net change in SWB. To the extent that scale-norming occurs, these unexpected results indicate changes in evaluation and measurement processes rather than the SWB constructs themselves. Yet, the two explanations are not mutually exclusive: scale-norming does not 'crowd out' adaptation, nor vice-versa. Instead, scale-norming and adaptation might account for varying proportions of the difference between measured and expected SWB, and the relative magnitude of either mechanism may vary between contexts.[10]

This is the nature of the difference between scale-norming and adaptation, expressed in a simple, didactic example. But neither would matter as *mere* possibilities, so I will now briefly consider empirical evidence that both adaptation *and* scale-norming do occur across a range of domains.

In a classic study, Brickman et al. (1978) compared a sample of winners of major lotteries to a sample of paralysed accident victims. The experimenters asked participants to rate how happy they were with life at the time of the interview, and before either winning the lottery or having the accident. Lottery winners reported a negligibly small improvement in present 'general happiness' relative to the control group (4.00 versus 3.82), while paraplegic and quadriplegic respondents reported being only marginally less happy than the controls. Further, all three groups were asked to rate the happiness they derive from a variety of mundane tasks, and the lottery winners reported a *lower* happiness relative to the other two groups. These surprising findings suggest that both the lottery winners and accident victims adapted to their new circumstances: their actual SWB rose or fell to its baseline after an initial period of shock or novelty. This is often invoked as a paradigm example of adaptation. Subsequent investigations by Oswald and Winkelmann (2019)[11] and Boyce and Wood (2011)[12] have revised the strong conclusions drawn by Brickman et al, while Diener et al. (2006) offers more equivocal counterpoints to the naïve 'hedonic treadmill model'. Nonetheless, adaptation remains a robust and widely recognised phenomenon.

A convincing demonstration of scale-norming from the real world mirrors the islands example. In a study of successful Tongan applicants to the New Zealand visa lottery,

---

[10] Suppose I predict that doubling my income will cause a 2-point increase in life-satisfaction or happiness on a 10-point scale. Suppose my income is in fact doubled, and my new reported SWB is only one point greater than previously. If I adjusted my scale for SWB 'downwards' by 0.5 points, then I in fact enjoyed a 1.5-point improvement in SWB; still less than my expected 2. In this example, scale-norming and adaptation both account for 50% of the difference between actual and predicted SWB.

[11] Panel data from a larger than original sample of German lottery winners showed some evidence for increased life satisfaction.

[12] A larger than original sample and more detailed survey methods suggested an interaction between personality traits and life satisfaction in the aftermath of disabling accidents.

Stillman et al. (2015) conducted subjective well-being surveys at annual intervals before and after migration. Although moving from Tonga to New Zealand involves an improvement in objective economic circumstances including job prospects, weekly wage, household income, and GDP per capita; some self-reported measures of SWB actually *decreased* throughout the process. More precisely, measures of affect taken from a five-question assessment of overall mental health (the MHI-5 scale) showed a small decrease, while questions about respect and self-assessed objective welfare showed a negligible change. On the other hand, *retrospective* estimates for pre-migration affect, perceived respect, and perceived welfare decreased by an even greater amount. If these results can be fully accounted for by adaptation, then an unequivocal improvement in economic circumstances failed to deliver an improvement in actual SWB, and the Tongan migrants inaccurately idealised their past as they adapted. Fabian (2019) suggests a more prima facie plausible explanation: that the migrants became happier after moving to New Zealand while adapting their criteria for the same (numerical) self-reported ratings of SWB.

Pervasive scale-norming effects might also explain why China's economic 'miracle' during the 1990s and 2000s appeared to result in a small net *decline* in SWB (Easterlin et al. 2012). How could such an impressive period of sustained economic growth fail to deliver any measured improvement in SWB? The scale-norming explanation underlines the word 'measured'; suggesting that China may have enjoyed large increases in subjective well-being in conjunction with corresponding increases in the standards or criteria for the same responses to questions about SWB. Although the rural poor in China may have become happier, they nonetheless saw an emerging urban middle class become *even* happier. Since *attainable* or *realistic* levels of happiness increased *more* rapidly than this purported actual improvement in happiness, the criteria for what constitutes a given number on the SWB (e.g. Cantrill Ladder) scale also plausibly increased at a faster rate than actual happiness. While Easterlin et al. attributed their results to the countervailing effects of unhappiness caused by income inequality, scale-norming suggests an additional role for income inequality in renorming the measure itself.

In sum, the adaptation explanation might undermine intuitions about SWB, but it does not pose a threat to the validity of existing measures of SWB. This is because adaptation, being a claim about actual psychological processes, is revealed by existing measures of SWB to the extent that they *successfully* track the underlying construct. By contrast, scale-norming entails that existing measures of SWB may systematically fail to track some changes or differences in the SWB constructs. The question at hand is not which explanation is best, but the *extent* to which scale-norming can in fact account for results which have hitherto been fully attributed to adaptation.

Causes

Neither adaptation nor scale-norming would be credible explanations if they lacked independently plausible causes. While there is some recognition in the literature that

scale-norming and adaptation are distinct[13], the differentiating *causes* of scale-norming have gone effectively unexplored.

There are at least two broad mechanisms for adaptation: habituation and contrast. The migrants from the previous examples may have habituated to their new circumstances such that they experienced no change in SWB once the novelty of increased material wealth wore off—another step on the 'hedonic treadmill'. Alternatively, migration may also have revealed contrastive possibilities the migrants hadn't before conceived of, influencing their actual affective experience. Now their neighbors have such extravagant goods and lifestyles, they feel a new and uncomfortable pressure to 'keep up'. The ubiquity of adaptation can be explained in terms of evolutionary efficiency: natural selection "favors a happiness function that measures the individual's success in relative terms"[14]. The continual promise of 'greener grass' incentivises sexual competition, while excessively intense and prolonged positive or negative sensations pose an obvious burden to fitness.

If scale-norming is real, then what causes, mechanisms or explanations might be given for inconsistent scale usage over time and between populations? No single process (psychological or otherwise) is responsible for scale-norming. Instead, I suggest five overlapping mechanisms, ranging from the mundane to the profound.

Firstly, scale-norming can reflect differences in the interpretation of the wording or language of SWB survey questionnaires. Tyler Cowen (2019) writes, "[t]he observation of a nearly flat happiness-wealth relationship says more about the nature of language than it does about the nature of happiness… Kenyans have recalibrated their use of language to reflect what they can reasonably expect from their daily experiences." Cowen argues it would not be surprising if those in "less happy societies often attach less ambitious meanings to the claim that they are happy." If this is to be usefully distinguished from other explanations, 'linguistic differences' cannot be treated as synonymous with 'scale-norming'. I suggest a *purely* linguistic difference is to be demarcated by a difference in *working definitions*. Consider adjectives whose referents are determined relative to a context or interest. While 'big' and 'tall' pick out different sizes and heights across contexts, these different uses do not entail change in *definition*, or a strictly *linguistic difference*, since the same general definition will apply across contexts. Plausibly, adjectives like 'happy' and 'satisfied' belong to this same category. I suggest this understanding of linguistic differences therefore encourages us to look elsewhere for the main causes of scale-norming.

Alternatively, an individual might be inclined to readjust their scale usage if their previous SWB rating is close to either bound of the closed numerical scales used in

[13] Fabian (2019)
[14] Rayo and Becker (2007). See also Perez-Truglia (2010) for elaboration.

SWB surveys; more often the upper bound—the so-called 'ceiling effect'[15]. Suppose a respondent describes their life-satisfaction as a 9/10 in one year, and subsequently gets married, finds a dream job, or otherwise enjoys an improvement in both objective circumstances and corresponding SWB. When answering the same survey two years later, they may realise their previous answer left too little scope to express this large change in SWB. In response, they update to a new scale where the upper bound represents a greater actual or underlying SWB than previously—analogous to adjusting the sensitivity of a microphone to avoid 'clipping'.

Thirdly, the most salient reference points, or counterfactual circumstances, might change relative to the individuals' actual objective circumstances and SWB. Consider a rural household in a developing country witnessing the emergence of an urban middle class; or learning about the lives of affluent foreigners through hearsay, television, or the internet. While previously a 10/10 score may have implicitly represented the best life possible for some narrow community, now it might come to represent this new and improved (perceived) quality of life. Thus, their scales will normalise upwards even absent an actual change in satisfaction or happiness. In *Transformative Experience* (2014), L. A. Paul describes how major life changes such as becoming a parent can give rise to radically new kinds of experience which cannot be comprehended or assessed in advance. This explains how shifts in one's conception of the range of possible experiences are often impossible to anticipate and thus factor into existing scales in advance.

Fourthly, 'expectancy' or 'observer' effects may generate different scales. This might be caused by cultural differences in conversational norms: a culture which is more conversationally agreeable (more inclined to answer questions in the affirmative or to outward displays of positivity like unprovoked smiling) might also be disposed to give more positive answers in SWB questionnaires for the same actual SWB. This could instead arise from differences in the perceived purpose of a survey or the interests of the experimenters—sometimes called the 'Hawthorne effect'. McClimans et al. (2012) call this the 'communication perspective', suggesting "respondents can communicate a message via response shift". Focusing on the medical context, they note that patients might interpret questions about their SWB as intended to evoke a "weak evaluation" of their contingent circumstances: am I being adequately cared for, and are my immediate needs being met? Healthy controls, on the other hand, might reasonably interpret the same questions as intended to evoke a "strong evaluation" of their life construed broadly; drawing on their personal 'vision of the good'.

Finally, either the wording of survey questions or the quantitative scales they generate may vary between successive survey rounds. This is perhaps the most obvious and yet

[15] Fabian (2019).

overlooked cause of scale-norming; maybe because of the difference in origin with the previous four *psychological* causes. Nevertheless, endogenous changes to the survey questions and scales themselves fits my original definition of scale-norming and therefore ought to be considered a legitimate cause: scale-norming is not a strictly behavioural or psychological phenomenon. In his original presentation of the 'paradox' which later bore his name, Easterlin (1974) observed that economic growth appears to be uncorrelated with improvements in life-satisfaction over time for a number of developed countries. Yet, Stevenson and Wolfers (2008) note that some of the life-satisfaction questions Easterlin relied on for his conclusion changed over time, leaving a positive correlation between SWB and GDP *within* comparable survey periods, as illustrated in the figure below.

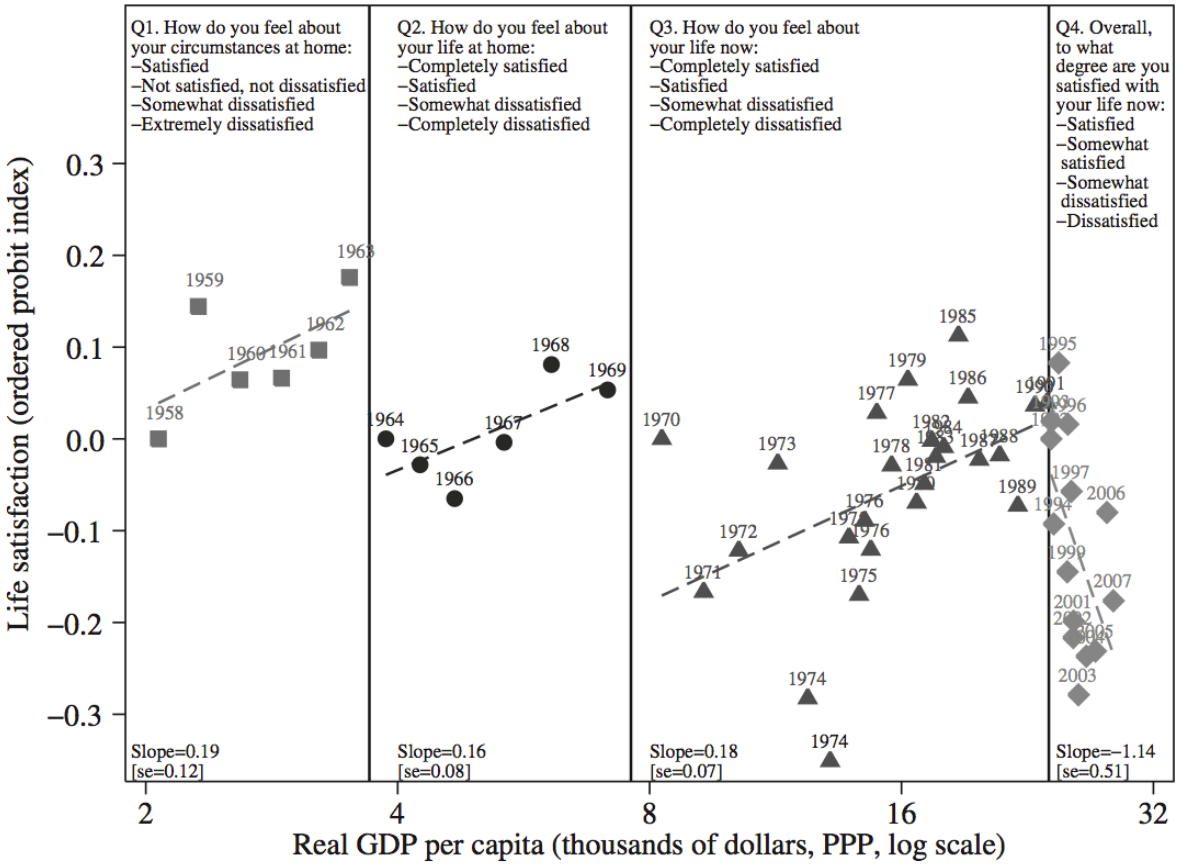

Figure 1. GDP per capita vs Life satisfaction across survey questions, Japan, 1958-2007 — Stevenson and Wolfers (2008)

This is a familiar feature of other measures in the social sciences. For instance, the 'Flynn effect'[16] denotes the observation that continual scale renormalisation conceals

[16] Trahan et al. (2014)

sustained and significant improvements in intelligence quotient (IQ) test aptitude. Since the generated scores are always standardized around an average score of 100 by convention, comparing unadjusted results at different times obscures these changes, just as in the figure above. Instead, the Flynn effect is observed when older tests are taken in *conjunction* with newer ones by the same people.

From these putative causes, it is possible to derive characteristic and contrasting features of adaptation and scale-norming. Firstly, adaptation typically takes place over the course of months or years: *instantaneous* adaptation would not confer an evolutionary advantage, presenting no incentive at all. However, scale-norming might occur within an indefinitely short time since observer effects, question wording, and reference classes can all change overnight. Secondly, we should expect adaptation to closely integrate with preferences—genuine adaptation might make a person willing to make larger sacrifices or take greater risks to further improve their circumstances. By contrast, merely updating the scale I use would be unlikely to generate new preferences. Thirdly, adaptation typically ought to be reflected in *recalled* SWB. Ignoring recall biases for simplicity of illustration, I have adapted to new circumstances if I recall being no happier before my improvement in circumstances than I am now (after a short-lived bump). If I do become happier but correspondingly update my scale upwards, then I will still recall being *less* happy before my change of circumstances. Further, scale-norming ought to affect my projected ratings of *hypothetical* circumstances; preserving their relative ranking with my actual life. On the other hand, adaptation is more likely to hold fixed my raw scores for imagined alternatives, while *varying* the relative ranking of my actual life. Lastly, adaptation might be expected to reflect changes in day-by-day affective experience, while scale-norming might occur absent any actual changes in affect.

Three Solutions

In the previous section, I considered plausible causal mechanisms for scale-norming, which suggested contrastive indicators of scale-norming and adaptation. In this section, I will make use of the latter three of these distinguishing features to describe three tentative kinds of procedure to identify and estimate the extent of scale-norming. First, a ‘recall’ approach for comparing scales over time. Second, a ‘reference class’ approach for comparing scales across populations and cohorts. Thirdly, and most tendentiously, a ‘frequential’ approach for relating global estimates of affect to more immediate self-reports. In turn, these procedures suggest how existing measures might be renormalised to better reflect the underlying SWB constructs.

Recall

The reason the results from the longitudinal study of Tongan migrants so strongly suggest scale-norming is because the migrants reported large improvements in life-satisfaction not fully reflected in their actual life-satisfaction scores across times. Since

there is no easily measurable and agreed-upon conception of subjective well-being which can mediate disputes about the extent and severity of scale-norming, it is at least coherent and consistent to attribute these discrepancies as artefacts of warped memory. The result is two competing explanations of discrepancies in recall: that these migrants are prone to systematically misremembering their life-satisfaction in even their recent past, or that they have adopted new criteria for reporting life-satisfaction. I suggest the second (scale-norming) explanation is far more plausible. Firstly, it has the virtue of taking seriously respondents' reports about their changes in life-satisfaction, and it avoids prescribing significant biases in recall which might be criticised as patronising or paternalistic. More pointedly, a bias for reporting *improvements* in life-satisfaction run counter to the typical direction of biases in recall of subjective well-being, which tend to *idealise* the past[17]. Suppose an individual P records numerically equal self-reports of SWB between successive survey rounds at t_1 and t_2, but recalls a large increase in SWB between t_1 and t_2:

1. Either P did not use the same scale at t_1 and t_2, or P's recall at t_2 of its subjective SWB at t_1 is radically biased.
2. P's recall at t_2 of its subjective SWB at t_1 is not radically biased.

C. Therefore, P did not use the same scale at t_1 and t_2.

If the argument is sound, it can be generalised from individuals to samples and populations—suggesting a procedure for normalising successive self-reports of the same individual over time. Consider two successive measurements x and y from the same individual at times t_1 and t_2 respectively, where the numerical measure ranges from 0 to 10. Ask the individual at t_2 to rate both their present SWB (specifically, life-satisfaction) *and* their recalled SWB at t_1, using the **same** (present) scale. This should be made explicit: e.g. "how would you have scored your life-satisfaction at t_1 if you were to use the scale / criteria you used to score your present life-satisfaction?". Call this second rating $y*$. If $y*$ is significantly different from y, infer scale-norming. On a new scale, take the initial rating y at t_1 as a fixed point, and adjust the new measure for t_2 to $x = x + (y - y*)$. This can be applied iteratively, over multiple successive survey rounds. Interpreted graphically, the result is a single line within a kind of 'worm' which tracks changes in SWB on a single scale, and displays the scales used at successive times in terms of a single scale; enabling comparisons across time. The resulting scale might no longer range from 0-10 on y's original scale:

[17] See Mitchell et al. (1997) for evidence of the so-called 'rosy view', and Prati and Senik (2020) for an examination of recall bias in affective self-reports.

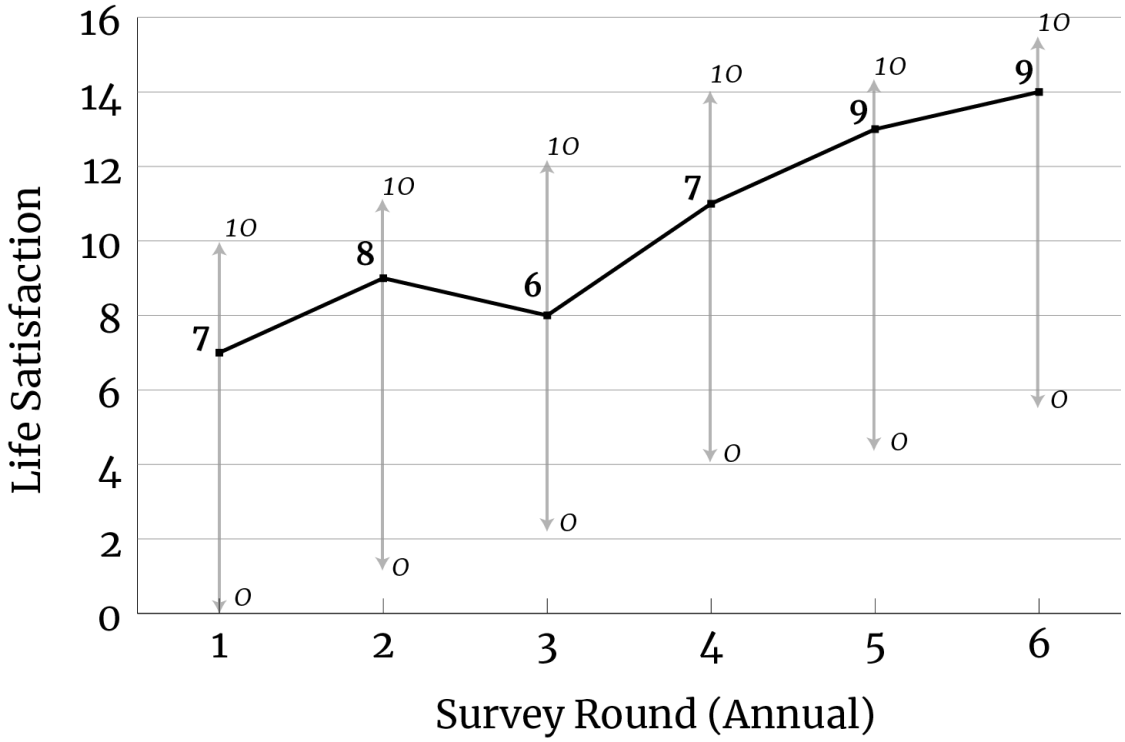

Figure 2. Visualising scale-norming over successive survey rounds. Adapted from Fabian (2019).

As in the study of Tongan migrants, the recall approach gives a way of identifying scale-norming in the accident victims surveyed in the Brickman et al. study. This is because while the paraplegic group's self-reported 'general happiness' was only marginally lower than the control group, they also reported being significantly happier with life in the past compared to the controls (with net changes of -1.45 and +0.5 out of 5 respectively). The authors conclude that the accident victims inaccurately idealised their past, since their actual past happiness should not in fact differ from the controls'; both being drawn from the same cohort[18]. By contrast, the recall approach uncovers the more realistic explanation of scale-norming.

Where there are uncontroversial, 'objective' measures, it is easy to identify and even correct for recall bias. For instance, self-reports of behaviours which are judged to be socially undesirable like smoking and drug-taking are often underestimates which can be adjusted towards their real values by comparisons with e.g. cigarette excise taxes[19]. It is possible to ascertain the accuracy of such reports because there exist noncontroversial and objective ways to measure their real values—unlike for SWB. Indeed, there is enough room for maneuver within the intuitive conception of SWB to

[18] See Drews and Greeland (199) for a discussion of the impact of differential recall on the results of case-control studies in an epidemiological context.
[19] Hatziandreu et al. (1989)

accommodate for different attitudes to recall bias, and in turn to the relative prevalence of scale-norming and adaptation. Differences between measured and recalled SWB can always be attributed entirely to recall bias (rather than scale-norming) without outright contradiction, even if this is implausible. Therefore, the recall approach should not be expected to provide a decisive and objective means to identify and quantify scale-norming.

While the recall approach aims to identify and correct for scale-norming across time (longitudinally), it may be possible draw inferences about *cross-sectional* comparisons across persons and populations at any one time. For instance, the example of the Tongan migrants might suggest that Tongans employ lower criteria for the same numerical rating of life-satisfaction than New Zealanders. By contrast, a second method for scale renormalisation prioritises cross-sectional comparisons, from which longitudinal inferences might be drawn. I turn to this next.

Reference Class

Suppose a researcher wanted to identify SWB scale-norming for people with some congenital disease or disability. In this case, she would have no use for the recall approach: you cannot ask a congenitally blind person how much happier they were when they could see. Instead, the researcher can ask the participant to forecast their well-being under different circumstances—answering a battery of *counterfactual* life-satisfaction questions: how happy *would* you expect to be under circumstances x, y, and z?

The researchers might then pose the same set of questions to a comparison group, and compare the two groups' results. If the average counterfactual SWB scores are lower for one group, then that group plausibly has higher criteria for the same numerical ratings of SWB: their scales have raised. Next, participants are asked to rank both their own SWB and that of the other group among the counterfactual circumstances (using the same scale they use for self-reports). On this approach, statistically significant differences in scores for the counterfactual reference class count as evidence for scale-norming. After correcting for such differences if they are present, the difference between one group's estimate of the other group's actual self-reported SWB counts as an instance of surprising or unexpected *adaptation*. Just as in the recall approach, the measures derived from the two groups might be rendered comparable by adjusting the SWB scores for one or both groups so that their scores for the reference class match.

To give a real example in the medical context, Lacey et al. (2008) asked patients with chronic illness (either lung disease or diabetes) and a control group of non-patients to answer two kinds of question. Firstly, they asked both groups to rank and numerically score their projected SWB for a varied set of adverse circumstances. Secondly, they asked the control group to estimate the rank and numerical SWB score for the patients' condition. A discrepancy in ratings was observed: patients rated their own SWB higher than the estimates of the non-patient control group. This difference between forecast

and actual SWB ratings is, of course, a prima facie indicator of adaptation. If instead it is entirely an artefact of *scale-norming*, then the patient's scales have been calibrated downwards: the same underlying SWB now yields higher absolute numerical scores. In this instance, we should expect the rank-order of the patients' reported SWB and the non-patients' corresponding estimates of SWB to remain fixed relative to one another and to the reference class, while the ratings for *all* the situations (actual and counterfactual) should have moved up for the patients. In the limiting case of full scale-norming and zero adaptation, patients' higher than estimated SWB self-reports ought to correspond to equivalently higher projected ratings for the counterfactual situations. On the other hand, if the discrepancy between actual and expected SWB can be fully accounted for by *adaptation*, then the ratings for the counterfactual reference class might be expected to stay fixed, while the relative ranking of the patients' own condition will be higher than the non-patients expectations. Where differences between reported and expected SWB are caused by some combination of scale-norming and adaptation, the extent of scale-norming can then be gauged by the extent to which the reference class is ranked relatively higher or lower, and the extent to which the relative rankings remain the same. While Lacey et al. present their method absent explicit theoretical motivation, interpreting it as an instance of this more general 'reference class' approach explains how it can be extended to other domains.

This approach is vulnerable to similar indeterminacies as the recall approach. Changes in a person's actual SWB might influence their *projected* SWB for other hypothetical situations—as in 'projection' or 'forecast' biases. For instance, the patients with chronic diseases might indeed scale-norm downwards while *also* underestimating how much *better* life is without that disease. These countervailing effects might thereby approximately cancel out, preserving numerical ratings and relative rankings consistent with the adaptation explanation while concealing the true scale-norming explanation. Again, there is no uncontroversial and objective way to adjudicate between these two explanations. Nonetheless, *agreement* between these two approaches may build a case that scale-norming is occurring. Where recall bias is a possibility in a psychological measure, it is often corrected for by case-control studies just like the reference class approach. Similarly, the possibility of projection bias might be corrected by relating to known instances of biases in affective forecasting[20] . The two approaches are thus mutually supportive: where both indicate scale-norming, it would be sensible to conclude that it is occurring. In searching for a robust measure of SWB that can accommodate for scale-norming, there are no obvious fixed points to build from—better instead to find the most coherent view that survives a range of tests.

[20] See, for instance, Wilson and Wheatley (2000)

To recapitulate, consider how these two approaches might be applied to the islands example. Recall that the successive self-reports of SWB before and after migrating from X to Y yielded no absolute change. Suppose that an additional question about recall revealed that the migrants to Y also report experiencing an improvement in SWB following migration; and that questions about projected SWB revealed that both migrants and controls ranked the SWB of living in Y over living in X. Here is the clearest possible indication of scale-norming: a convergence of the two approaches. If scale-norming is not occurring, the migrants must be making systematic errors in recalling their past SWB; and the controls must be inaccurately predicting the SWB associated with life in Y. Thus, it is possible to deny the reality of scale-norming in cases like these *only* at the significant cost of postulating radical and systematic errors in both recall and forecasting.

Frequential

So far, I have considered two kinds of procedure for identifying scale-norming: the 'recall' and 'reference class' approaches. These methods were best suited to measures of life-satisfaction over measures of affect. Both make use of claims about how ratings of SWB relate to past or counterfactual ratings given changes or differences in the 'actual' or 'underlying' SWB constructs. I suggested that large discrepancies between recalled life-satisfaction and past measures of life-satisfaction are better explained by scale-norming than purely as instances of biased recall; and that changes in scales ought to alter raw life-satisfaction scores while *preserving* relative rankings for counterfactuals. In this third approach, I will consider how scale-norming can be specifically identified in measures of affect. This is the most contentious approach, because it explicitly relies on normative claims about the relation between instant and recalled measures of affect.

Affect can be measured both retrospectively and in the present moment. An example of a retrospective measure is the 'Positive and Negative Affect Schedule' (PANAS), which instructs participants to indicate the extent to which they have felt each of a list of both positive and negative emotions over the past week, such as 'active', 'irritable', and 'interested'. A score is generated by taking a weighted average of the separate positive and negative scales; the resulting numerical value ranging from 0 to 10. An example of a measure of momentary or 'instant' affect is Daniel Kahneman's 'experience-sampling method' (ESM)[21]; where an electronic device prompts the participant at random intervals to report on their immediate experience. Between the PANAS and the ESM in immediacy lies the 'day reconstruction method' (DRM) asks participants to describe the previous day's activities and their affective character.

[21] Kahneman et al. (2004)

Corresponding to these methods is a construct that Kahneman has called 'objective happiness': the temporal integral of these momentary reports. Thus, the *accuracy conditions* for recalled affect are given by the balance of momentary (hypothetical or actual) reports of affect. The picture is one of a continuity between immediate and increasingly long-term recall of affective experience.

The 'objective happiness' view of the relation between instant and retrospective reports of affect therefore assumes that "instant ratings must contain all the relevant information required."[22] While I will assume this is the case for clarity of presentation, note that Alexandrova (2005) criticises such a claim on the grounds that "assessing happiness moment by moment leaves no place for both cognitive and moral ex post evaluation of our own inner states."

Suppose Annie spends most of her days in boredom or frustration but tends to systematically over-estimate her average happiness when prompted; while Bob lives an overall more contented life but tends to give more stubbornly realistic estimates of his average happiness. When both are asked in a survey to recall the balance of their affective experience over the past week, Annie and Bob give identical answers. The 'objective happiness' construct is able to make sense of the intuition that Annie's response is less accurate than Bob's, because Annie's numerical score for life-satisfaction less accurately reflects the average score she would have given in more frequent questions asking about her current happiness.

In order to get a procedure for identifying scale-norming from a claim about the accuracy conditions for measures of recalled affect, I will first assume that measures of recalled affect are vulnerable to scale-norming, as in the foregoing example. Secondly, I assume that momentary reports of affect are likely to be more vulnerable than longer-term reports to a familiar battery of recall biases and distortions of memory, and therefore far *less* vulnerable to scale-norming[23]. This just reflects Kahneman's claim that more frequent measures of affect over shorter time scales are (all other things being equal) more reliable measures than longer-term recall. In turn, a method linking momentary and recalled measures of affect may constitute a legitimate procedure for identifying affective scale-norming.

For illustration, suppose a cohort's recalled affect over the past month increases between successive survey rounds by 10%, while aggregated momentary affect improved by 20%. This might (tentatively) suggest that the cohort have renormalised their scales downwards by 10% relative to their 'objective' happiness. In this way,

---

[22] Alexandrova (2005)

[23] This is also highly plausible for exogenous reasons: Perez-Truglia (2012) makes a convincing case that immediate and acute affective sensations are differentially less adaptive than global assessments of affect.

identifying scale-norming in recalled affect is made more tractable than for life-satisfaction.

Naturally, shorter-term techniques are also more expensive and less convenient. It is not feasible to distribute experience sampling or DRM methods to anywhere near the size and diversity of samples which simple survey questions can. More people can be reached at less expense by a few items on a yearly survey, such as the World Happiness Report. Therefore, the appropriate role of more frequent measures of affect is not to replace longer-term and less frequent measures, but for small-scale controlled experiments using those measures to inform and adjust those larger-scale measures in light of what they tell us about the potential prevalence and severity of scale-norming in different contexts.

This 'frequential' approach suggests that the lottery winners surveyed by Brickman et al., once touted as paradigms of pure adaptation, may have adjusted their scales upwards. This is because they reported taking less pleasure in 'mundane' events than the control group (3.33 vs. 3.82)—where mundane pleasure ratings might be construed as a rough proxy for average momentary affect.

Although the 'frequential' approach most naturally applies to measures of affect, it might also be extended to measures of life-satisfaction. For instance, Newman et al. (2020) found that global reports of well-being overestimate aggregated daily states of well-being, across a range of measures including life-satisfaction. The authors suggest that global evaluations of life-satisfaction provoke respondents to privilege significant, recent, or memorable experiences; while omitting the "quotidian ebbs and flows" of ordinary experience. The *extent* to which global reports of well-being differ from aggregated daily states could then be compared between times and cohorts, which in turn may indicate differences in life-satisfaction scales. However, this begs the question against life-satisfaction being legitimately independent of aggregated affect. Yet, life-satisfaction reports are attractive precisely because the respondent is free to decide what counts as relevant or important considerations. Thus, straightforwardly inferring scale-norming in measures of life-satisfaction from 'objective happiness' is not a credible option. To treat life-satisfaction as a proxy for aggregated momentary affect is like readjusting a two-star restaurant review because the food was actually very good, and the reviewer just failed to appreciate it.

However, it does not follow that measures of life-satisfaction cannot be *partially* constrained by similar accuracy conditions. This is because it is equally possible to make mistakes in aggregating non-psychological conceptions of the good such as the number and strength of close relationships, social status, attainment of skills or 'flourishing', and altruistic impact. In this way, life-satisfaction reports can accommodate weak accuracy conditions.

In this way, the challenge is to carve out a conceptual space which accommodates both the democratic virtues of life-satisfaction measures, and the practical usefulness of the frequential approach in identifying scale-norming. I suggest that much confusion in this

regard may derive from equivocating two definitions of "measure of subjective well-being":

> **SWB*** — A measure of subjective well-being is an individual's subjective assessment of their own well-being.
>
> **SWB**** — A measure of subjective well-being is an assessment of an individual's subjective well-being.

Life-satisfaction reports are best described by SWB*, since they needn't be informed solely by aggregated affective states: neither descriptively, nor as a matter of accuracy conditions. By contrast, measures of affect are best described by SWB**, on which biases in recall about aggregated affect might legitimately be corrected. When life-satisfaction questions are instead interpreted on SWB**, they are mostly useful as indirect measures of affect[24] (the only non-circular psychological candidate). This importantly does not rule out that life-satisfaction judgements (accurate or inaccurate) might themselves provide grounds for affective states: reflecting on one's life can itself be an occasion for positive or negative emotional states.

The 'frequential' approach, where scale-norming in long-term measures of SWB is identified by their relation to shorter-term measures which are less vulnerable to scale-norming, is applicable in the context of SWB**; while I suggest the recall and reference-class approaches are agnostic between SWB* and SWB**.

The distinction arises from an unfortunate pun on 'subjective' that is rarely made explicit: the difference between its *psychological* use as in 'subjective experience', and its *evaluative* use as contrasted with 'objective'. It is possible that a great deal of unnecessary disagreement has been generated in the SWB literature by this confusion between an understanding of measures of subjective well-being as a person's assessment of their own life using their own conception of the 'good life', and an assessment of the aggregate or average subjective (qua psychological) quality of their life.

Insofar as a poor community with expanded capabilities and economic freedoms does in fact become happier, but the change registers as a decrease in raw measures of SWB, SWB** can describe that change as an instance of scale-norming while SWB* cannot[25]. Therefore, if measures of SWB should track a conception of SWB that matters for

[24] Note that the existence of valid SWB constructs does not imply that self-reports are *about* such constructs.
[25] See Sen (1970) and (1999) and Haybron (2007).

policy, then SWB** may be a better interpretation than SWB* of what counts as a 'measure of subjective well-being'.[26]

In this section, I considered three procedures for identifying scale-norming. The emphasis has been to link scale-norming and adaptation to other kinds of judgement about what they entail. In turn, this allows attitudes about those other judgements to inform and update best guesses about the extent of scale-norming. Since there is no a priori recourse to an objective standard of SWB by which to evaluate claims about the extent of scale-norming[27], I intend these three approaches to contribute to a kind of mutually supportive, coherent whole. Finally, I considered a distinction in the meaning of "measure or SWB" which may help clarify where these approaches are applicable.

Conclusion

Even assuming some established understanding of its actual or underlying constructs, the literature almost entirely neglects the possibility of scale-norming. Yet, if it is real, then intertemporal and interpersonal comparisons of unadjusted measures of SWB are potentially invalid, and such measures cannot obviously underpin all the ambitious policy proposals they have inspired. To this end, I described three potential ways scale-norming can be identified and adjusted for. The 'recall' and 'reference class' approaches suggest additional survey questions to distinguish hedonic adaptation from genuine scale-norming in self-reports of both life-satisfaction and affect. The 'frequential' approach specifically targets measures of positive and negative affect, appealing to Kahneman's notion of 'objective happiness' to link scale-norming to discrepancies in recalled and reported affect. Given the typical direction of scale-norming, the effect sizes of interventions on subjective wellbeing are likely to increase in light of these adjustments.

This leaves at least three broad directions for further inquiry. Firstly, although I have repeatedly eluded to (and implicitly assumed the existence of) the 'real' or 'underlying' constructs of subjective well-being, I have not described them in any detail. Indeed, the difficulty in identifying scale-norming owes in part to the opaque, normatively infused, and controversial status of these SWB constructs. Therefore, work remains to be done to reveal the processes and mechanisms underlying both life-satisfaction reports and positive or negative affect[28]. Secondly, it remains to determine the actual, empirical extent and severity of scale-norming in measures of subjective well-being. I hope the

[26] Note also that SWB** does not imply that the balance of affective states is all that *matters.* This is compatible with a range of pluralistic conceptions of well-being.

[27] For an extended defence, see Alexandrova (2017).

[28] See Reed and Csikszentmihalyi (2014), Kahneman, Diener, and Schwartz (1999), and de Lazari-Radek (forthcoming).

approaches I have described go some way to describing how such a project might be undertaken. Finally, the normative significance of SWB measures ought to be reassessed in light of scale-norming: an evaluative controversy cannot be answered without evaluative conclusions. I therefore suggest that scale-norming forces us to seriously reconsider the relevance of unadjusted life-satisfaction reports for policy.

Bibliography

Alexandrova, Anna. *A Philosophy for the Science of Well-Being*. Oxford: Oxford University Press, 2017.

—. "Subjective Well-Being and Kahneman's 'Objective Happiness'." *Journal of Happiness Studies* (2005): 301-324.

Boyce, Christopher J. and Alex M. Wood. "Personality Prior to Disability Determines Adaptation: Agreeable Individuals Recover Lost Life Satisfaction Faster and More Completely." *Psychological Science* (2011): 1397–1402.

Brickman, Philip, Dan Coates and Ronnie Janoff-Bulman. "Lottery Winners and Accident Victims: Is Happiness Relative?" *Journal or personality and social psychology* (1978).

Busseri, Mechael, Sadava, Stan. "A Review of the Tripartite Structure of Subjective Well-Being: Implications for Conceptualization, Operationalization, Analysis, and Synthesis." *Personality and Social Psychology Review* (2010).

Cowen, Tyler. Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals. Stripe Press, 2018.

Diener, Ed, et al. *Well-Being for Public Policy*. Oxford University Press, 2009.

Diener, Ed, Richard E. Lucas and Christie N. Scollon. "Beyond the Hedonic Treadmill." *American Psychologist* (2006): 305-314.

Dolan, Paul and Robert Metcalfe. "Measuring subjective wellbeing: recommendations on measures for use by national governments." *Journal of social policy* (2012): 409-427.

Drews, Carolyn D and Sander Greeland. "The Impact of Differential Recall on the Results of Case-Control Studies ." *International Journal of Epidemiology* (1990): 1107-1112.

Easterlin, Richard A. "Does Economic Growth Improve the Human Lot? Some Empirical Evidence." *Nations and Households in Economic Growth*. 1974. 89-125.

Easterlin, Richard A., et al. "China's life satisfaction, 1990–2010." *Proceedings of the National Academy of Sciences* (2012).

Fabian, Mark. Adaptation or Scale Norming? 2019.

Gallup. *The World Gallup Poll*. n.d. <https://www.gallup.com/analytics/232838/world-poll.aspx>.

Hatziandreu, E.J., et al. "The reliability of self-reported cigarette consumption in the United States." *American Journal of Public Health* (1989).

Haybron, Daniel. "Life satisfaction, ethical reflection, and the science of happiness." *Journal of Happiness Studies* (2007): 99-138.

Haybron, Daniel. "The Nature and Significance of Happiness." Boniwell, Ilona, Susan A. David and Amanda Conley Ayers. *Oxford Handbook of Happiness*. Oxford: Oxford University Press, 2013.

Helliwell, John, et al. "World Happiness Report." 2020. <https://worldhappiness.report/ed/2020>.

Kahneman, D, et al. "A survey method for characterizing daily life experience: the day reconstruction method." *Science* (2004): 1776-80.

Kahneman, Daniel and Angus Deaton. "High income improves evaluation of life but not emotional well-being." *Center for Health and Well-being* (2010).

Kahneman, Daniel. "Objective Happiness." Kahneman, Daniel, Ed Diener and Norbert Schwartz. *Well-Being: The Foundations of Hedonic Psychology*. 1999.

Lacey, Heather, et al. "Are they really that happy? Exploring scale recalibration inestimates of well-being." *Health Psychology* (2008): 669-675.

Larson, Reed and Mihaly Csikszentmihalyi. "The Experience Sampling Method." *Flow and the Foundations of Positive Psychology*. 2014. 21-34.

Layard, Richard. *Can We Be Happier?* 2020.

McClimans, Leah, et al. "Philosophical perspectives on response shift." *Quality of Life Research* (2012): 1871-1878.

Mitchell, Terence R., et al. "Temporal Adjustments in the Evaluation of Events: The "Rosy View"." *Journal of Experimental Social Psychology* (1997): 421-448.

Newman, David, Norbert Schwarz and Arthur Stone. "Global reports of well-being overestimate aggregated daily states of well-being." *The Journal of Positive Psychology* (2020).

Ortiz-Ospina, Estaban and Max Roser. "Happiness and Life Satisfaction." May 2017. *Our World in Data.*

Oswald, Andrew J. and Rainer Winkelmann. "Lottery Wins and Satisfaction: Overturning Brickman in Modern Longitudinal Data on Germany." Rojas, Mariano. *The Economics of Happiness*. Springer, 2019. 57-84.

Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Paul, Laurie A. *Trannsformative Experience*. Oxford University Press, 2014.

Perez-Truglia, Ricardo. "On the Causes and Consequences of Hedonic Adaptation." *Journal of Economic Psychology* (2012).

Plant, Michael. Doing Good Badly? Philosophical Issues Related to Effective Altruism. Oxford: PhD Thesis, 2019.

Prati, Alberto and Claudia Senik. "Can people remember how happy they were?" (2020).

Rayo, Luis and Gary S. Becker. "Evolutionary Efficiency and Happiness." *Journal of Political Economy* (2007).

Sen, Amartya. Collective Choice and Social Welfare. 1970.

—. Development as Freedom. 1999.

Stevenson, Betsey and Justin Wolfers. "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox." *Brookings Papers on Economic Activity* (2008).

Stillman, Steven, et al. "Miserable Migrants? Natural Experiment Evidenceon International Migration and Objective and Subjective Well-Being." *World Development* (2015).

Trahan, Lisa, et al. "The Flynn Effect: A Meta-analysis." *Psychological Bulletin* (2014).

van Praag, Bernard. "The Relativity of the Welfare Concept." Nussbaum, Martha and Amartya Sen. *The Quality of Life*. Oxford: Oxford University Press, 1993.

Veenhoven, Ruut. "Is Happiness Relative?" *Social Indicators Research* (1991).

Wilson, Timothy D. and Thalia P. Wheatley. "Focalism: A Source of Durability Bias in Affective Forecasting." *Journal of Personality and Social Psychology* (2000).