Can we trust wellbeing surveys? A pilot study of comparability, linearity, and neutrality

Conrad Samuelsson
Summer Research Fellow, Happier Lives Institute

Samuel Dupret
Research Analyst, Happier Lives Institute

Michael Plant
Director, Happier Lives Institute

Caspar Kaiser
Assistant Professor, Tilburg University
Trustee, Happier Lives Institute

Happier Lives Institute

March 2023

Contents

Introduction

Subjective wellbeing (SWB) data, for example answers to life satisfaction questions, are important for decision-making by philanthropists and governments. Such data are currently used with two important assumptions: First, that reports are **comparable** between persons (e.g., that my 6/10 means the same thing as your 6/10) and second that reports are **linear** in the underlying feelings (e.g., that going from 4/10 to 5/10 represents the same size change as going from 8/10 to 9/10). Fortunately, these two assumptions are sufficient for analyses that only involve the *quality* of people's lives. However, if we want to perform analyses that involve trade-offs between improving *quality* and *quantity* of life, we would also need knowledge of the **'neutral point'**, the point on a wellbeing scale that is equivalent to non-existence.

Unfortunately, evidence on all three questions is critically scarce. We[1] propose to collect additional surveys to fill this gap. Our aim with this report is two-fold. First, we give an outline of the questions we plan to field and the underlying reasoning that led to them. Second, we present results from an initial pilot study (n = 128). Unfortunately, this small sample size does not allow us to provide clear estimates of the **comparability** of wellbeing reports. However, across several question modalities, we do find tentative evidence in favour of approximate **linearity**. With respect to **neutrality**, we assess at what point on a 0-10 scale respondents say that they are 'neither satisfied nor dissatisfied' (mean response is 5.3/10). We also probe at what point on a life satisfaction scale respondents report to be indifferent between being alive and being dead (mean response is 1.3/10). Implications and limitations of these findings concerning neutrality are discussed in Section 6.2. In general, the findings from our pilot study should only be seen as being indicative of the general feasibility of this project. They do not provide definitive answers.

In the hopes of fielding an improved version of our survey with a much larger sample and a pre-registered analysis plan, **we welcome feedback and suggestions on our current survey design**. Here are some key questions that we hope to receive feedback on:

1. Are there missing questions that could be included in this survey (or an additional survey) that would inform important topics in SWB research? Are there any questions or proposed analyses you find redundant?

2. Do you see any critical flaws in the analyses we propose? Are there additional analyses we should be considering?

[1] **Author note:** Conrad Samuelsson, Samuel Dupret, and Caspar Kaiser contributed to the conceptualization, methodology, investigation, analysis, data curation, and writing (original as well as review and editing) of the project. Michael Plant contributed to the conceptualization, supervision, and writing (review and editing) of the project.

3. Would these data and analyses actually reassure you about the comparability, linearity, and neutrality of subjective wellbeing data? If not, what sorts of data and analyses would reassure you?

4. What are some good places for us to look for funding for this research?

Of course, any other feedback that goes beyond these questions is welcome, too. Feedback can be sent to casparkaiser@gmail.com or to samuel@happierlivesinstitute.org.

The report proceeds as follows. In **Section 1**, we describe the challenges for the use of self-reported subjective wellbeing data, focusing on the issues of comparability, linearity, and neutrality. We highlight the implications of these three assumptions for decision-making about effective interventions. **Section 2** gives a description of the general methodology of the survey. In **Section 3**, we discuss responses to the core life satisfaction question. In **Sections 4, 5, and 6**, we describe how we will assess comparability (Section 4), linearity (Section 5), and neutrality (Section 6). We also discuss our substantive and statistical assumptions (stated informally). In **Section 7** we make some wider comments on the feasibility of fielding a scaled-up version of this survey. **Section 8** concludes.

We hope that this work will eventually help us and the wider research community to understand whether, and to what extent, we can reliably use survey data as a cardinal and comparable measure of people's wellbeing.

1. Linearity, comparability, and neutrality as challenges for wellbeing research

If we want to know how people's lives are going, an obvious, if not widely-used, method is to rely on individuals' own self-reported assessment of their own wellbeing. Philanthropists and policymakers are increasingly using such data in their decision making (Durand, 2018). Previous research showed that subjective wellbeing measures are reliable (i.e., giving similar results across multiple measurements, OECD, 2013; Tov et al., 2021) and valid (i.e., succeed in capturing the underlying phenomenon they seek to measure, Kahneman & Krueger, 2006). Still, there are doubts about the reasonableness of using subjective wellbeing reports to evaluate intervention effects on wellbeing and to make interpersonal welfare comparisons. These doubts are largely due to a lack of extensive research about how such reports are generated.

In this report, we engage with three important areas regarding self-reported subjective wellbeing:

1. The **comparability** of reports between persons (is my 6/10 the same as your 6/10?).

2. The **linearity** of subjective wellbeing reports (is going from 4/10 to 5/10 the same amount of increase as going from 8/10 to 9/10?).

3. The position of the '**neutral point**', which refers to the wellbeing level at which existence and non-existence for someone are equivalent in value.

Comparability and linearity are required to allow for the consistent estimation of relative wellbeing effects of interventions (see, for example, Kaiser, 2022). The neutral point is an additional requirement for estimating trade-offs between the quantity and quality of life (McGuire et al., 2022b). We therefore believe these three issues to be particularly important. For an overall theoretical analysis and a review of the existing (limited) empirical literature, see Plant (2020), who tentatively concludes we can assume SWB scales are comparable and linear.

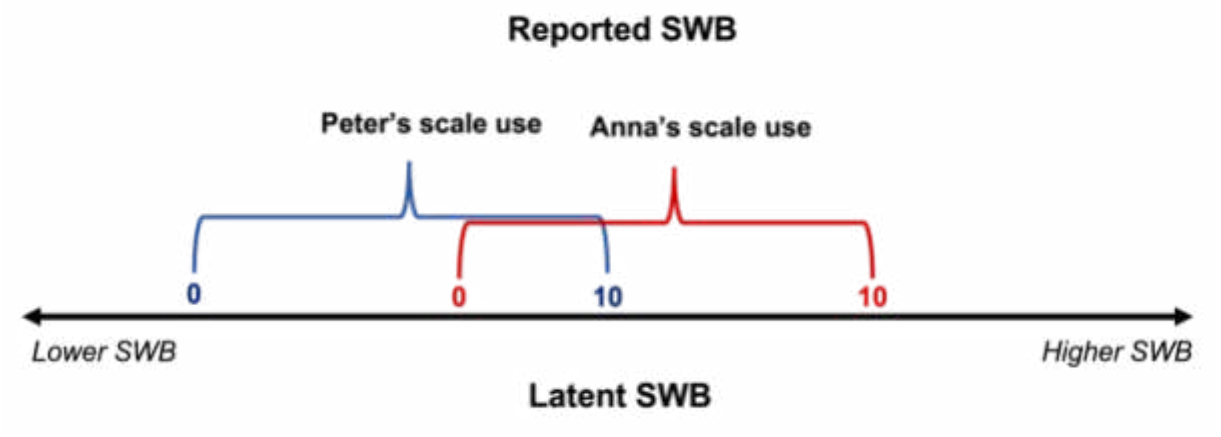
We will be fielding a large-scale survey to evaluate comparability, linearity, and neutrality from multiple perspectives. As a first step in this process, we have fielded a small-scale pilot, from which we will report some initial results.

Whether SWB scales are interpreted in a linear and comparable manner is an open question. Should it turn out that SWB scales are not (approximately) linear and comparable, knowledge about the ways in which these assumptions fail will allow us to correct for these deviations in subsequent analyses. Below, we discuss some of the previous work on these issues.

1.1 Comparability

With "*comparability*" we here mean that identical SWB reports refer to identical levels of SWB, regardless of the person and time. One way to think about comparability is in terms of *common scale-use*. Suppose we give you a scale for you to rate your life satisfaction between 0 and 10. In order for you to render a judgement about which digit best represents your life satisfaction, you need to decide where the threshold for each digit is. You then have to match your subjective feeling of life satisfaction to a number on the scale. Because this is a complicated cognitive process, we may expect that people will differ in their scale-use (see Figure 1). If differences in scale-use were random, they would 'wash-out' in sufficiently large samples and would not bias subsequent analyses.

Figure 1. An illustration of differences in scale-use for wellbeing reports



Note: *For a given level of underlying wellbeing, Peter will use larger numbers than Anna. Equivalently, a given response (e.g., a 6/10) corresponds to a lower level of wellbeing for Peter than for Anna. Peter's and Anna's scale-use, therefore, differs.*

However, if scale-use differed systematically, so that there were differences in scale-use between groups, then any reported differences in life satisfaction between groups would be confounded by differences in scale-use. In the literature, differences in scale-use are sometimes referred to as *scale shifts*, which may either occur between people or between points in time for a given person. Although most research assumes comparability, that assumption has been doubted. One reason for this are observed changes in response behaviour that are caused by factors that are unrelated to the content of the survey questions (Bertrand & Mullainathan, 2001). For example, question order, differences in the phrasing of questions, and ordering of answer options can yield substantial differences in responses between randomised groups of respondents.

Other researchers have argued that incomparability may explain counterintuitive differences between groups with similar objective circumstances. For example, French respondents tend to report surprisingly low life satisfaction (Angelini et al., 2014; Kahneman et al., 2004); women report higher life satisfaction than men despite having worse outcomes on many objective measures (Montgomery, 2022); and ageing populations report higher life satisfaction despite having poorer health and more loneliness.

Similarly, apparent inconsistencies in how respondents rate their current SWB, their improvement in SWB, and their memories of past SWB across time suggest that intrapersonal scale shifts occur (i.e., people change the scale they use over time; Fabian, 2022; Kaiser, 2022; Prati & Senik, 2020). Intrapersonal scale shifts may be an explanation of the well-known Easterlin paradox, which states that long-term country-wide increases in GDP do not improve wellbeing, despite the fact that

income is robustly correlated with wellbeing at an individual level (Kahneman & Deaton, 2010; Jebb et al., 2018; McGuire et al., 2022a).

However, even if SWB reports are incomparable, that need not be the end of subjective wellbeing data. For example, researchers have used *vignettes*, short descriptions of (fictional) persons to which people should give common answers, to estimate the differences in scale-use between groups, and to subsequently correct for these differences (e.g., Montgomery, 2022; Angelini et al., 2014; Wand, 2012; King et al., 2004). Of course, this test for comparability relies on the assumption that respondents have a common perception of the wellbeing of the persons described in the vignettes. This assumption may not always hold, and may depend on the design of vignettes. Similar methods have been suggested for intrapersonal scale shifts using data on memories (Kaiser, 2022). Generally, these studies find that survey responses are not perfectly comparable, but that biases arising from such non-comparability are typically too small to, for example, impact estimates of the *sign* of an effect. That said, the validity of current methods for assessing comparability has been questioned, and there is a lack of cross-method validation; namely, do the results from different methods, which have different underlying assumptions, converge?

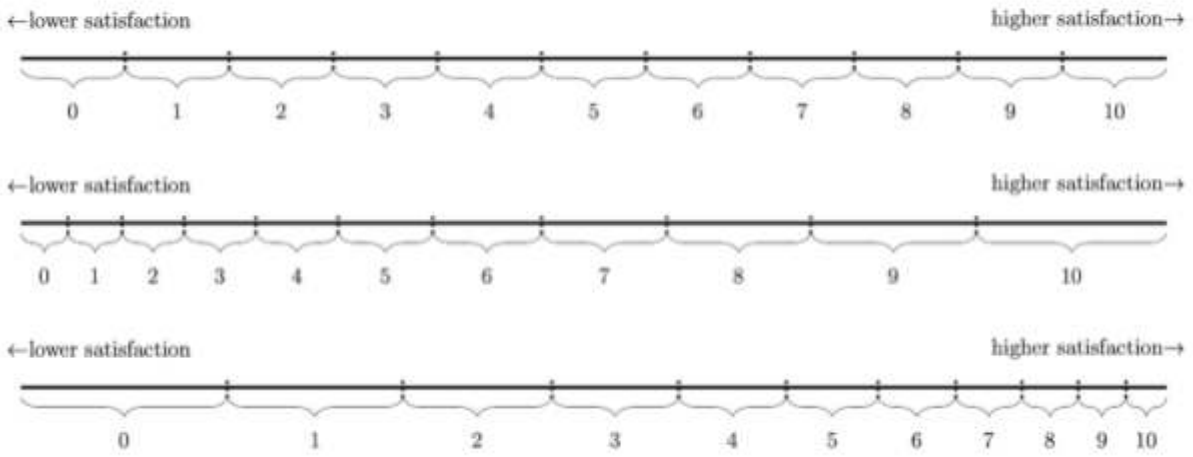
1.2 Linearity

As we use the term, "*linearity*" refers to the assumption that the relationship between latent and reported SWB is linear, so that any step on a SWB scale indicates the same increase or decrease of latent SWB, regardless of where on the scale the person is. To illustrate, consider Figure 2 which shows a linear, convex, and concave relationship between latent and reported SWB.

Currently, most researchers treat SWB data as though the linearity assumption was met. As in the case of comparability, violations of this assumption would bias analyses of SWB data, and could even reverse conclusions – as shown in several pieces of recent work (Bond & Lang, 2019; Schröder & Yitzhaki, 2017).

In response to this, Kaiser and Vendrik (2021) demonstrated that non-linear transformations of SWB scales would have to strongly deviate from linearity in order for such reversals to occur, and such strong deviations seem improbable. Few papers have sought to quantify how non-linear scale-use is. As one example, Oswald (2008) showed that the relationship between objectively measured and subjectively reported height is roughly linear. However, it remains unclear whether these initial results generalise sufficiently well to justify the linearity assumption for *wellbeing* data. Our survey design seeks to test this.

Figure 2. Linear, convex, and concave relationships between latent and reported life satisfaction (as one example of SWB)



Note: *The top of the figure shows a linear relationship between reported and latent wellbeing. In the middle, a convex relationship is shown. That is, the difference in latent wellbeing is larger between the higher response options than between the lower response options. The opposite concave pattern is shown in the bottom of the figure.*

1.3 Neutrality

The *neutral point* refers to the level on a SWB scale at which existence has a neutral value, compared to non-existence for that person (assuming this state is perpetual and considering only the effects on that person). Above this point life is 'better than death'; below it life is 'worse than death'. This is conceptually distinct, but possibly closely related, to what we call the *zero point*: the level on a SWB scale at which that type of SWB is overall neither positive nor negative (e.g., someone is neither overall satisfied or dissatisfied). A natural thought is that the *zero point* and the *neutral point* coincide: if life is good(/bad) for us when it has positive(/negative) wellbeing, so a life has neutral value if it has zero wellbeing.[2]

Locating the neutral point is essential in some decision-making contexts, such as prioritising healthcare, where we must compare the relative value of improving the quality of life with the value of increasing the quantity of life.[3] To date the location of the neutral point remains an open question. We think that potential answers can be informed by a combination of theoretical reasoning and empirical research.

[2] A standard philosophical explanation for what makes death bad for us (assuming it can be bad for us) is *deprivationism*, which says that death is bad because and to the extent it deprives us of the goods of life. Hence, death is bad for us if we would have had a good life and, conversely, death is good for us if we would have had a bad life. Here we take it that a good(/bad/neutral) life is one with overall positive(/negative/neutral) wellbeing. See, for example, Nagel (1970).

[3] The need to, and difficulty of, assigning values to both various states of *life* and to *death* is also a familiar challenge for measures of quality- and disability-adjusted life years (QALYs and DALYs). For discussion, see, for example, Sassi (2006).

To motivate the problem, consider two interventions **A** and **B**. For simplicity, assume that we simply seek to maximise total wellbeing. Under intervention **A**, 100 people's wellbeing is raised from a baseline score of 6/10 to 8/10 for a single year, but there is no change in each person's length of life. Under intervention **B**, there is no change in each person's baseline wellbeing level, but each person's length of life is increased by one year.

Define one 'Wellbeing Adjusted Life-Year' - a WELLBY - as a one point gain on the 0-10 wellbeing scale for one year. Intervention **A** yields 2*1*100=200 WELLBYs. We can do this without reference to the neutral point, because we know the counterfactual: without the intervention, they will live at 6/10.

For **B**, we need to assign a neutral point (i.e., a score equivalent to non-existence). Suppose we place the neutral point at 0/10. Under intervention **B**, we count a gain of 6*1*100=600 WELLBYs. Hence, intervention **B** seems more effective than intervention **A**. Now instead assume that the neutral point is located at 5/10. Under intervention **B**, we now merely gain (6-5)*100=100 WELLBYs. Under this alternative assumption, intervention **A** seems more effective. Hence, the location of the neutral point matters when attempting to decide between life-extending and life-improving interventions.

In previous works, the neutral point has been chosen in an *ad hoc* manner (also see Plant et al., 2022, for discussion). Layard et al. (2020), for example, set the neutral point at 0/10. However, in this case there would be no way of reporting a level of latent wellbeing below the neutral point. This violates the intuition that people can ever have overall bad lives (i.e., overall negative wellbeing), or lives where it would be rational for them to wish to die. One might instead think that the neutral point is at 5/10 (c.f. Diener et al. 2018). This, however, would counterintuitively imply that a vast proportion of the world's population lives below the neutral point (as many report life satisfaction levels below 5/10; Our World In Data, 2020).

Hence, there is no obvious and uncontroversial *a priori* choice here. The limited research done so far indicates that people place the neutral point somewhere between 0 and 5 on SWB scales. A small (n = 75) survey in the UK found that, on average, respondents would prefer non-existence over a life satisfaction level of about 2/10 (Peasgood et al., unpublished, as referenced in Krekel & Frijters, 2021). The IDinsight Beneficiary Preferences Survey (2019, p. 92; n = 70), estimated the neutral point to be 0.56. The think-tank Rethink Priorities (unpublished) ran pilot studies about the neutral point using a 0-100 scale (from the worst pain, suffering and unhappiness to the best pleasure, positive experience, and happiness[4]). When participants (n = 35) are asked at what level they prefer to be alive rather than dead, the mean answer is 24.9/100.

[4] The fact that the scale mixes three concepts into one seems problematic.

It is unclear what to make of these answers. An important first step is to get a better understanding of what respondents believe and why. To be clear, this is only a first step: decision-makers will not necessarily want to take such beliefs at face value if they seem mistaken or irrelevant. We consider limitations of our measures in Section 6.2.

As noted, a natural thought is that the *neutral point* and *zero point* will coincide.[5] We test whether respondents put the neutral point and zero point for life satisfaction in the same place, and whether respondents interpret the zero point for life satisfaction as somewhere between 0 and 5. If respondents do both, that would provide an explanation of previous works' findings. However, there are several reasons why respondents might not believe that the neutral point and the zero point coincide. Some of these issues are discussed in Section 6.2.

With these preliminaries in place, we next outline the general features of our pilot survey. Thereafter, we present the life satisfaction question we asked participants. Then we discuss, in that order, our questions on comparability, linearity, and neutrality. Throughout, we will report some tentative and preliminary results.

2. General outline of the survey

The survey contains 50 questions, which can be found in this external appendix with the question. We list the different types of questions used and the topics they address in Table 1 below.

We ran a pilot of the survey using Qualtrics for the implementation and Prolific Academic for the recruitment (chosen for its reasonably high data quality, see Peer et al., 2022). We used Prolific filters to recruit participants who lived in the UK, who spoke English as their first language, and had a Prolific Score between 95 and 100 (i.e., they were rarely rejected from studies). We used the balancing option from Prolific to recruit a similar number of men and women.

We recruited 128 participants. The median time to complete the survey was 9.82 minutes. In our sample, 64 participants report being females, 63 males, and 3 others. The mean age was 39.15 years old (median = 35.00, SD = 14.26). For additional summary statistics, see our external Appendix A.

[5] This is entailed by, for instance, a standard formulation of utilitarianism. In classical utilitarianism, the value of an outcome is the sum total of wellbeing in it, where wellbeing consists in happiness. On this view, *ceteris paribus*, extending an overall happy life is good, whereas extending an overall unhappy life is bad. 'Good' and 'bad' are understood either in terms of being good/bad for the person or good/bad 'for the world'. We are not endorsing classical utilitarianism here, but merely point out that aligning the neutral point with the zero point on the *appropriate* wellbeing scale (whatever that happens to be) would be a textbook view in ethics.

Table 1: Different question types and how we use them

Question type	Comparability	Linearity	Neutrality
Life satisfaction question	+	+	+
Vignettes (from Montgomery, 2022)	+		
Vignettes (ours)	+		
Psychophysical calibration (inspired by Benjamin et al., 2021)	+	+	
Objective-Subjective Method (inspired by Oswald, 2008)	+	+	
Neutral point questions	+	+	+
Endpoint questions	+		
Reporting linear scale use		+	
Continuous life satisfaction question		+	

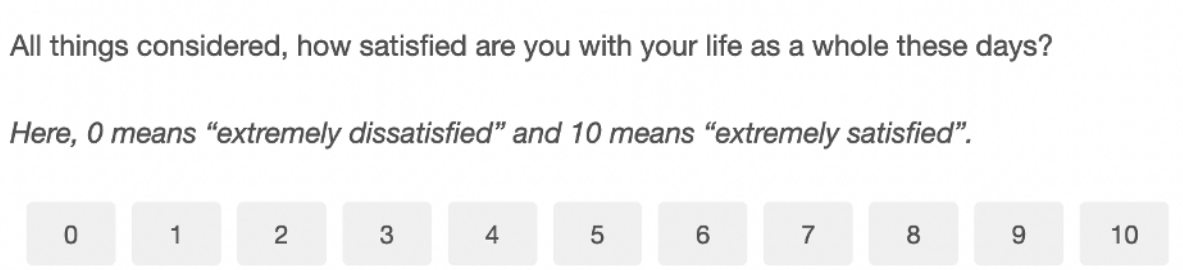
3. Life satisfaction question

At the start of the survey, we ask participants how satisfied they are with their life, on a scale from 0-10 (i.e., an 11-point scale). See Figure 3 for an illustration.

Life satisfaction is a typical measure of subjective wellbeing which best fits with a global desire theory of wellbeing (Plant 2020). It forms the basis for subsequent questions (e.g., the vignettes will be asking participants about the life satisfaction of the persons in the vignettes). In that sense, our survey is – for now – aimed at establishing neutrality, comparability, and linearity in life satisfaction reports in particular. In the future, we may extend our analyses to other kinds of subjective wellbeing data (e.g., affect or happiness).

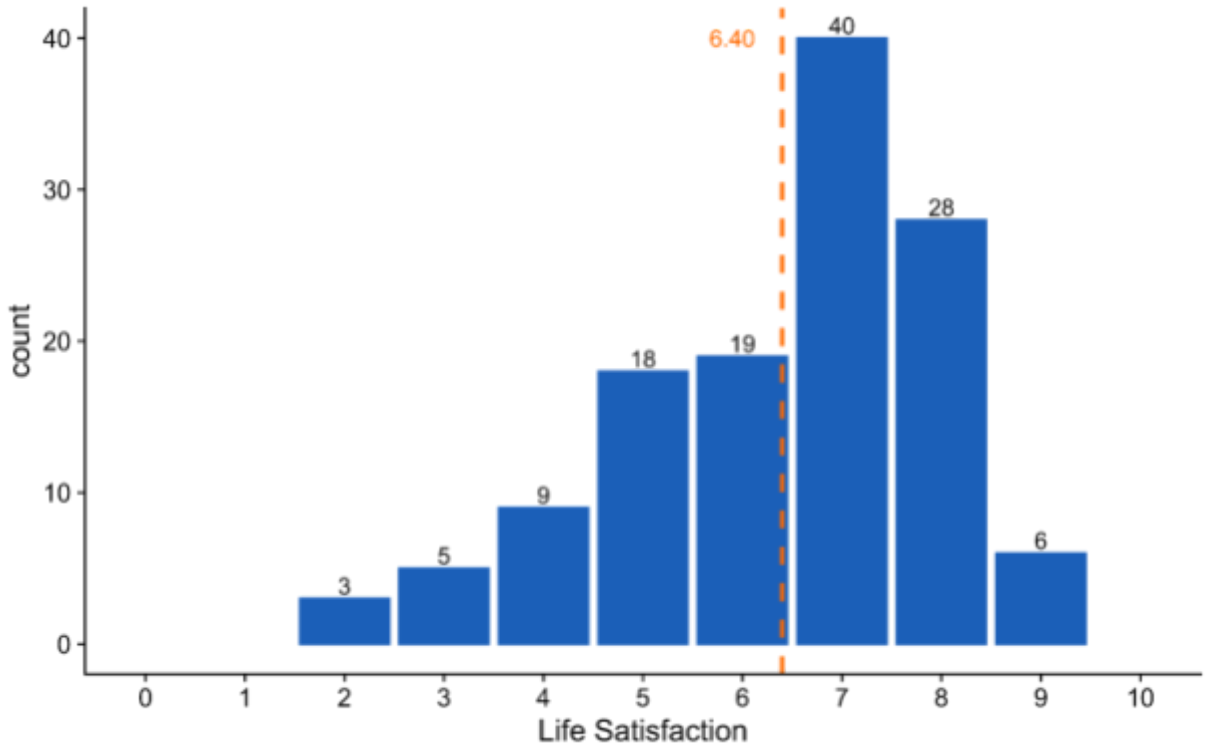
In our pilot, participants reported an average life satisfaction of 6.40 (SD = 1.62) points. This is slightly lower than the UK's 2020 average life satisfaction of 6.94, as reported by Our World in Data. See Figure 4 for the observed distribution of responses.

Figure 3. Screenshot of our life satisfaction question.



Note: *This is a screenshot of the question in the exact format presented to respondents.*

Figure 4. Distribution of life satisfaction responses



Note: *As is common in the field, responses to the life satisfaction question displays a left-skew. No respondents selected a 0, 1, or 10 as their response. Total N = 128.*

4. Comparability questions

By 'comparability', we mean that the same responses to wellbeing questions by different respondents refer to the same underlying level of wellbeing across respondents. To probe comparability, we are piloting five different kinds of methods. For each method, we give a description, an outline of the analyses to be conducted, and a brief discussion of the assumptions underlying each method.

Our methods are:

1. **Vignette method:** Adjust respondents' SWB-scores based on their ratings of short descriptions of people's lives, called vignettes. This is the standard method in the literature (King et al., 2004).

2. **Psychophysical calibration method:** A version of a method proposed by Benjamin et al. (2021), which consists of adjusting scores based on judgements about qualitative properties of presented stimuli.

3. **Objective-subjective method:** A generalisation and elaboration of an approach by Oswald (2008), which involves adjusting SWB-reports based on how respondents make subjective and objective judgements about their own characteristics that are observable on a cardinal and comparable scale.

4. **Neutral point method:** Adjust respondents' SWB-reports based on where they put the neutral point.

5. **Endpoint questions:** Simply ask respondents two questions about how they assign the endpoints on their life satisfaction scale.

Each of these methods requires different sorts of assumptions. We plan to assess the extent to which these different methods yield convergent results about the comparability of scale-use. However, in order to estimate the extent to which scale-use depends on respondents' socio-economic and other characteristics, we need a much larger sample than currently available. We therefore do not yet report on results from the kind of models sketched in the section below.

4.1 How comparability can be assessed

The first four of our methods all have a similar structure: We administer so-called 'anchoring' questions that are answered on the same kind of scale as the SWB questions. Using these answers, we will estimate a set of parameters to indicate whether or not there are differences in scale-use (i.e.,

whether there are issues with comparability), and what socio-demographic characteristics might determine these differences.

To do so, we rely on two general kinds of assumptions, though the specifics slightly differ across methods. The first may be called ***interpersonal equivalence***: respondents all have the same interpretation of the anchoring questions and all intend the same latent level on the construct when responding to the questions we pose. If this assumption is satisfied, any differences in responses to the anchoring questions are due to interpersonal differences in scale-use. Unfortunately, the equivalence assumption is untestable. Fortunately, the assumption can be relaxed to merely say that any differences in interpretation of the anchoring questions between respondents are not correlated with the characteristics (income, relationship status, health status, etc.) we use to predict wellbeing.

The second assumption required is ***intrapersonal scale consistency***. Loosely speaking, this assumption entails that an individual's patterns of scale-use are common across both anchoring and wellbeing questions. The next subsection will introduce some notation to make these two assumptions more explicit.

4.1.1 How comparability can be assessed (with notation)

If you are not interested in notation, feel free to skip to section 4.2.

Suppose that respondent i's underlying life satisfaction w_i is additively determined by characteristics X_i:

$$w_i = \beta X_i + \varepsilon_i$$

For simplicity, assume that the error term is distributed as $\varepsilon_i \sim N(0,\ 1)$.

Underlying (or latent) life satisfaction w_i is never observed. All we observe is participants' choice of response category r_i. Respondents choose from some number of K ordered response categories (e.g., K = 11 for a 0-10 scale). In turn, the response $r_i \in \{1, 2, \ldots k, \ldots, K\}$ that respondents choose is determined by their underlying life satisfaction w_i, as well as a set of thresholds τ_i:

$$r_i = k \; iff \; \tau_{i,k-1} < w_i \leq \tau_{i,k}$$

Comparability of satisfaction reports would entail that all thresholds $\tau_{i,k}$ are common across respondents, that is $\tau_{i,k} = \tau_k$ for all k response categories. If that were the case, models of underlying

wellbeing would be identified and could be estimated by maximum likelihood (conditional on the assumption of a normal error being correct).

However, suppose that the thresholds themselves are also determined by the characteristics X_i:

$$\tau_{i,k} = \gamma_k + \delta X_i$$

To gain some intuition, suppose that for some particular characteristics $X_{i,m}$, the associated coefficient δ_m were negative. This would mean that the variable $X_{i,m}$ is associated with less stringent scale-use. For example, imagine that $X_{i,m}$ is a dummy-coded variable indicating gender, with $X_{i,m} = 1$ denoting that the respondent identifies as a woman. In that case, $\delta_m < 0$ would imply that, for a given level of latent wellbeing, women tend to use a higher response category than men.

Unfortunately, without further data, the coefficients β and δ are not separately identified (Greene & Henser, 2010). When we observe that respondents' answers to wellbeing questions are associated with their characteristics (gender), we don't know whether that's because of a genuine difference in wellbeing, or a difference in scale-use, or a mixture of the two. 'Hierarchical ordered probit models' have been proposed as a potential remedy to this problem (see Wand, 2012; King et al., 2004). We plan to make use of (variants of) such models.

These kinds of models all depend on the availability of a particular kind of additional 'anchoring data', which we hope our survey questions will supply. In particular, we will field questions in which respondents face some stimuli s_i, about which respondents are then asked to provide a rating $r_i^{(s)}$. Respondents are asked to give these ratings on a scale with the same number of K response options as the original life satisfaction question.

Under the following two assumptions, such reports would then enable separating the determinants of scale-use (δ) from the determinants of underlying wellbeing (β):

(1) ***Interpersonal equivalence.*** Respondents' perceptions of the stimulus s_i are uncorrelated with the characteristics X_i. More formally, $\beta^{(s)} = 0$ in $s_i = \beta^{(s)} X_i + \varepsilon_i^{(s)}$.

(2) ***Intrapersonal scale consistency.*** Respondents use (up to a linear transformation) the same scale when answering questions about their latent life satisfaction w_i as about the stimulus s_i. More formally, when $r_i^{(s)} = k \; iff \; \tau_{i,k-1}^{(s)} < s_i \leq \tau_{i,k}^{(s)}$ and $r_i = k \; iff \; \tau_{i,k-1} < w_i \leq \tau_{i,k}$, we also have $\tau_{i,k} = a + b\tau_{i,k}^{(s)}$ for some a and some $b > 0$ and all $k \in 1, 2, ..., K$.

Vignettes of fictional persons (e.g., Montgomery, 2022) have previously been proposed as potentially meeting these assumptions. As we see it, more sources of anchoring data should be tried. It is unlikely that these assumptions strictly hold for any of the methods we consider below. However, by comparing results generated from multiple different kinds of stimuli, we hope to at least be able to *bound* the extent to which differences in scale-use are likely to occur. We are still in the process of working out the exact details of such a bounding method.

4.2 Vignette method

4.2.1 Methodology

As noted above, correcting subjective scores based on vignettes is a common method among researchers using subjective scales (e.g., Montgomery, 2022; Angelini et al., 2014; Wand, 2012; King et al., 2004). The idea is simple: We administer participants the same fictional account of a person (a vignette) and ask respondents about the expected life satisfaction of that fictional person. If people have similar interpretations of each vignette, any differences in reported assessments are likely to be driven by differences in scale-use.

In our survey, we administer half of the respondents three vignettes proposed by Montgomery (2022). Here is an example:

> *Think of a female who is 40 years old and happily married with a good family life. Her annual family disposable income is about £31,400. She has severe back pain, which keeps her awake at night.*
>
> *How satisfied with life do you think this person is?*

The other half of the participants receive a different set of three vignettes created by us. These newly designed vignettes focus more on subjective mental states than on objective variables. We hope that this will help us to meet the aforementioned *interpersonal equivalence assumption*. An example of this kind of vignette goes as follows:

> *Christina is 30 years old. Although generally happy, she is disappointed with some aspects of her life. She is well compensated but finds her work meaningless. She has several close friends, though she sees them seldom. Christina has been dating for a year, which she finds exciting. She has no major health concerns but is bothered by headaches twice a week.*
>
> *How satisfied with life do you think this person is?*

For a full list of the vignettes we plan to field, see this external appendix with the question.

Each set of vignettes is designed to be ordered. That is, in each set, one vignette represents a person we think of as having the lowest life satisfaction of the three, one with a medium level of life satisfaction, and one with the highest life satisfaction of the three.

4.2.2 Specific assumptions

As with all the comparability methods, the vignettes rely on the assumptions of *interpersonal equivalence* and *intrapersonal scale consistency*.

For the vignette method, the *interpersonal equivalence* assumption amounts to the claim that differences in interpretations of vignettes are not conditional on the characteristics of the participants that are used to predict their life satisfaction (income, relationship status, health status, etc.).

Intrapersonal scale consistency here means that participants are rating the vignettes using the same scale that they use to rate their own life satisfaction.

4.2.3 Analyses

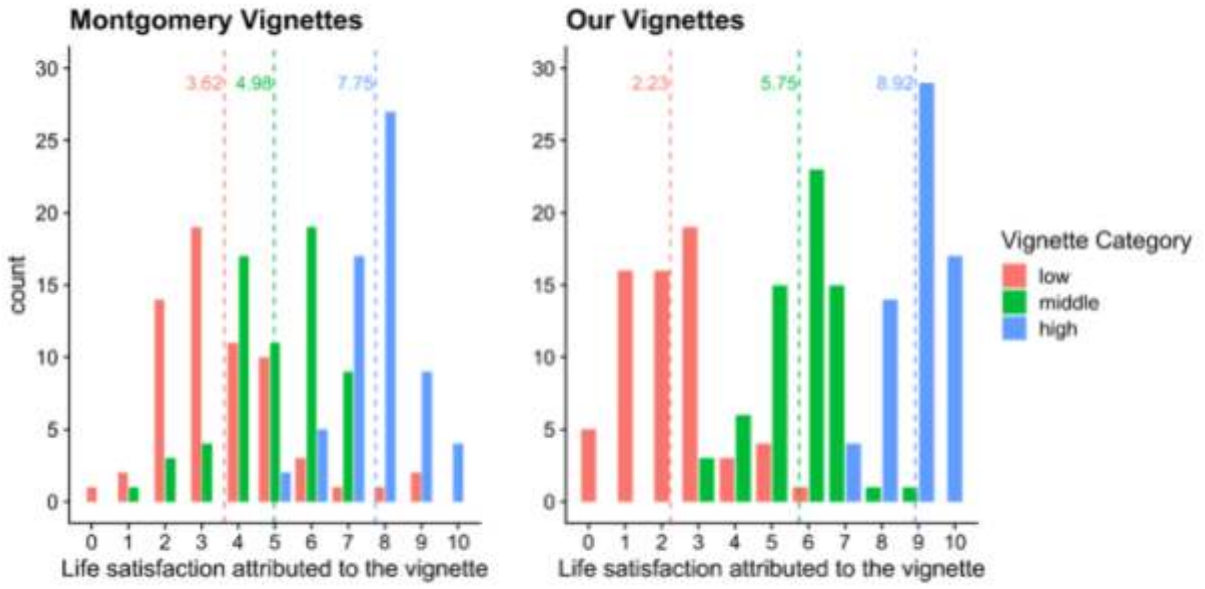
We intend to conduct three main analyses with the vignette data:

- Define a model along the lines detailed in Section 4.1. Such a model would jointly predict life satisfaction and variation in scale-use, following the approach of King et al. (2004). We will assess whether this model replicates previous findings such as Montgomery's (2022) findings about gender differences in scale-use (i.e., that women are more lenient in their scale-use). Also see Wand (2012).

- As a partial test of the vignette equivalence assumption, we will examine the orderings of the vignettes and the share of people who agree on a particular ordering.

- Estimating OLS regressions that predict the ratings of the vignettes using the same demographics variables that are used to predict life satisfaction, then comparing the models for each vignette and examining whether the demographic variables have different coefficients for the different vignettes. If so, this would indicate that vignette interpretation is not equivalent conditional on the demographic variables, thus yielding another partial test of vignette equivalence. See d'Uva et al. (2011) for an implementation of this idea in an ordered probit model.

Figure 5. Distribution of life satisfaction attributed to the vignettes



Note: *Colours distinguish between the vignettes. The red bars indicate the vignettes that were, within each set, intended as worst off, the green bars indicate vignettes intended to be in the middle, and the blue bars represent the vignettes that were described in the most positive terms. The numbers indicate mean assessments for each vignette across all respondents. These means follow the intended ordering for both sets of vignettes. However, means for our vignettes are more spaced-out than is the case for the vignettes used by Montgomery (2022). There is also less overlap in the vignette distributions. We take this as indicative of our vignettes performing better.*

Table 2. Frequencies of orderings of vignettes

Ranking	Montgomery vignettes		Our vignettes	
	n	%	n	%
Low/middle/high	50	78.13%	64	100.00%
Low/high/middle	1	1.56%	0	0.00%
Middle/high/low	1	1.56%	0	0.00%
Middle/low/high	12	18.75%	0	0.00%

Note: *The "Ranking" column on the far left shows the possible patterns of rankings that respondents selected. The words "low", "middle", "high" indicate the intended ranking of vignettes. As shown in the first row, not all respondents followed Montgomery's intended ranking of vignettes. In contrast, all respondents followed the ranking we intended for our own vignettes.*

4.2.4 Feasibility results

As noted above, we do not yet have sufficiently large samples to test the effects of particular socio-demographic characteristics on scale-use. In this section, we therefore limit ourselves to merely presenting some results that bear on the feasibility of fielding vignette questions.

Figure 5 shows the distributions of how participants ranked the vignettes for both the Montgomery vignettes and our own vignettes. In Table 2, we compare how often participants correctly rank the vignettes in the intended order (i.e., they give the lowest score to the lowest vignette, the middle score to the middle vignette, and the highest score to the highest vignette). For the Montgomery vignettes, participants do seem to rank them in the correct order, giving - on average - higher scores for the vignettes that are intended to have higher scores. However, ~20% of participants do not rank vignettes in the intended order.

For our own vignettes, which are more focused on the mental lives of the persons described, participants do seem to rank them in the correct order, giving - on average - higher scores for those vignettes that are also intended to have higher scores. Indeed, 100% of participants rank vignettes in the intended order. This tentatively suggests that participants interpret our vignettes in more similar ways than the vignettes used in Montgomery (2022). If so, the *interpersonal equivalence* assumption is more plausibly met in our vignettes.

4.3 Psychophysical calibration method

4.3.1 Methodology

Suppose that differences in vignette ratings are not due to differences in how participants understand questions about subjective wellbeing, or how they assign endpoints to life satisfaction scales, but due to more general differences in scale-use that occur across most or all subjective scales. How could we identify such non-specific differences in scale-use?

One approach, which has been suggested by Benjamin et al. (2021), is to show affectively neutral stimuli to respondents and ask them to rate certain qualities of the stimuli on a subjective scale (in our case, 0-10). The stimuli used take inspiration from the literature on psychophysics (e.g., Shepard, 1981). Because the objective value of the stimuli is cardinal, subjective scale shifts can be estimated.

Figure 6. Examples of calibration stimuli

Note: *Relating to the picture on the far left, respondents are asked to evaluate the acuity of the angle shown. For the blue dot shown in the middle, we vary the darkness of the colour shown to respondents. Concerning triangles on the right, we vary the number of triangles and ask respondents to rate how much of the square is filled by these.*

We intend to collect data on three kinds of stimuli: circle darkness, angle acuity, and square coverage (see Figure 6). Each stimulus modality (circle darkness, angle acuity, and square coverage) has three versions of varying levels. The distances between each version are equal (e.g., the three angles' obtuseness are 20, 80, 140 degrees, thereby, a 60 degrees difference between each version). Participants rated all three versions of all three kinds of stimuli on a 0 to 10 scale (e.g., "How dark is this circle? Here, 0 means "extremely light" and 10 means "extremely dark"").

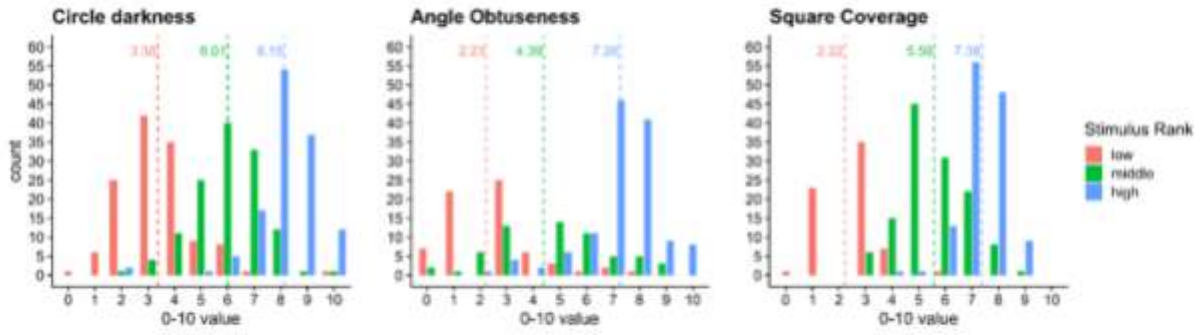
With this data, we intend to estimate similar models as described in the section on vignettes. One reason to estimate models on both kinds of data (vignettes or psychophysical stimuli) is to determine whether the calibration merely corrects for a smaller, but more general scale shift, or whether vignette-based corrections yield essentially equivalent results. The latter case would suggest that SWB judgements are just 'ordinary' subjective judgements, and that we should assign them the same level of credence that we assign to other subjective judgements.

4.3.2 Specific assumptions

Here, *interpersonal equivalence* means that participants have similar interpretations of the calibration stimuli. If the assumption is met, differences in subjective reports are attributable to differences in scale-use.

Intrapersonal scale consistency here implies that the spacing of thresholds of the response options are a linear transformation of those used for the life satisfaction question. In other words, scale-use patterns are assumed to be similar across the different subjective scales. *Prima facie*, we expect there to be similar cognitive processes involved in both kinds of scale-use, but the extent to which the processes are analogous is an open question.

Figure 7: Distribution of scores for each calibration stimuli



Note: *Colours distinguish between the rank of the stimulus within each type (e.g., low, medium, or high angle acuity). The dotted lines indicate mean subjective ratings for each type and rank combination. These means follow the intended orderings.*

4.3.3 Analyses

Primarily, we intend three analyses with these data:

- A joint model of life satisfaction and scale-use, exploiting individual differences in responses from each of the calibration questions. This model will be essentially analogous to the model used for vignette data.
- Replicating the analyses of Benjamin et al. (2021).
- Analysing the equivalence assumption and consistency assumptions using analogous statistical methods to the ones we described in the vignette section.

4.3.4 Feasibility results

To assess the feasibility of our calibration questions we present the distributions of how participants ranked the three versions of each three stimuli. For all three stimuli (circle darkness, angle obtuseness, and square coverage), participants seem to place them in the correct order, giving — on average — higher scores for versions that are higher (more dark, more obtuse, and more covered). See Figure 7 for the distributions.

However, the means for the different versions of each stimuli are not perfectly equidistant. For example, the distance between the means of the lower darkness circle and the middle darkness circle is 2.63 and the distance between the means of the middle darkness circle and high darkness circle is 2.14. In terms of individual rankings, 93% or more of participants rank the different versions of the stimuli in the intended order. See Table 3 for the frequencies.

Table 3: Frequency of different rankings of calibration stimuli

Ranking	Circle		Angle		Square	
	n	%	n	%	n	%
Low/middle/high	124	96.88%	120	93.75%	125	97.66%
High/low/middle	1	0.78%	2	1.56%	1	0.78%
Low/high/middle	1	0.78%	3	2.34%	2	1.56%
Middle/high/low	0	0.00%	2	1.56%	0	0.00%
Middle/low/high	2	1.56%	1	0.78%	0	0.00%

Note: *The "Ranking" column on the far left shows the possible patterns of rankings that respondents selected. The words "low", "middle", "high" indicate the objective rankings of the stimuli. As shown in the first row, across all kinds of stimuli most respondents followed the objective ranking.*

4.4 Objective-subjective method

4.4.1 Methodology

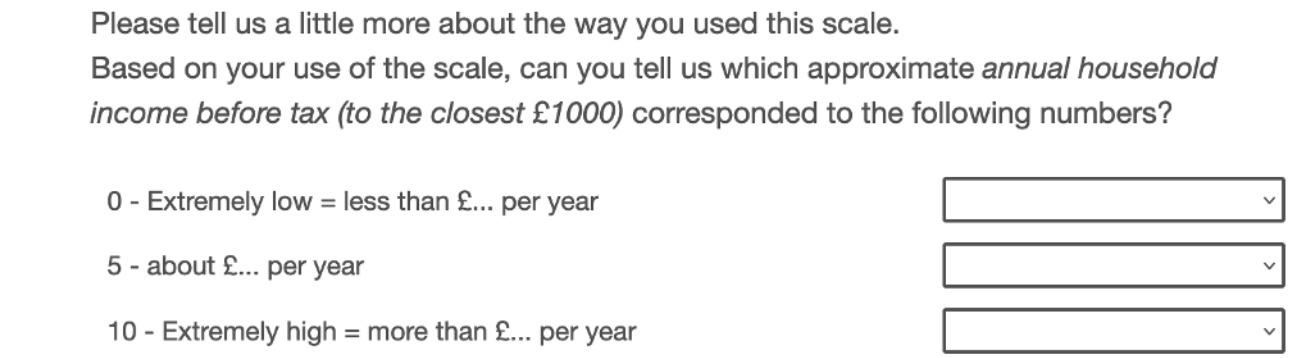
This approach is inspired by Oswald (2008), where participants were asked to subjectively report their height on a slider before measuring their objective height. Oswald then plotted the subjective height reports against the objective measurements and found a near-linear relationship. We use a variant of this method. We discuss our use of this method for linearity in Section 5. Here, we use it to assess comparability. We refer to these as *'objective-subjective' (OS) judgements*.

In the survey, we ask respondents to subjectively rate their level on three cardinal variables: height, income, and number of friends[6]. We then ask participants to report their cardinal value. We also ask participants to estimate population's averages of each of these variables. This is to examine whether scale-use differences are associated with individuals' knowledge about the population distribution of the underlying concept.

Objective-subjective judgements are analogous to SWB judgements in many ways: respondents need to adopt reasonable endpoints on the scale and 'translate' their objective value into a subjective judgement between zero and ten – precisely what respondents to SWB surveys do. However, unlike SWB judgements, Objective-subjective judgements are clearly about externally observable cardinal quantities (with clearly positioned and meaningful 0 points), allowing us to anchor the subjective rating against the objective cardinal quantities.

[6] Because SWB questions rarely include such qualifiers, questions are posed without any qualifiers i.e., we do not ask participants "what is your height *in comparison to your gender*?", or "what is your income *relative to other UK citizens*?".

Figure 8. The middle and endpoints question for the objective-subjective question on income



Note: *This is a screenshot of the question in the exact format presented to respondents.*

We also ask participants, for each *objective-subjective* question, to indicate the middle and endpoints of the subjective scale in terms of objective numbers. As sketched out further below, this will aid us in getting better individual-level estimates of both comparability and linearity. See Figure 8 for an example.

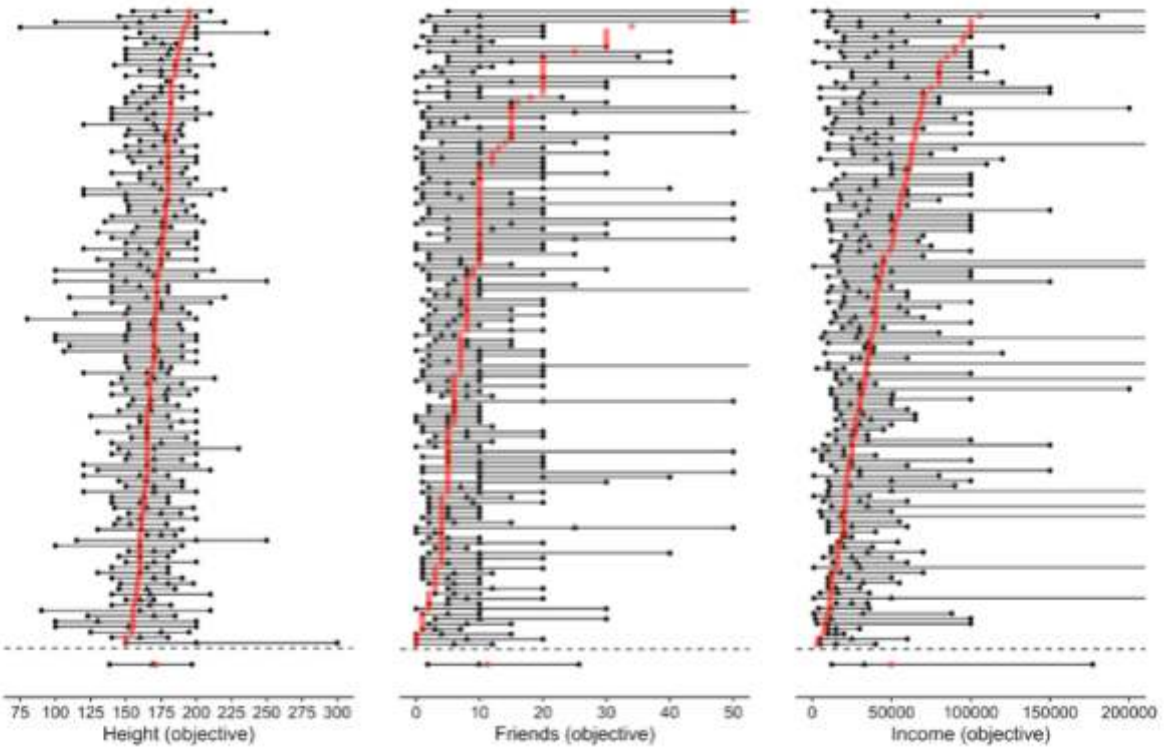
4.4.2 Specific assumptions

In the *objective-subjective* method, respondents report their cardinal quantity on a given variable as well as their rating of their objective quantity on a subjective scale. We thus derive, for each respondent, a "translation function" from the objective quantity (e.g., 170 cm) to the subjective rating (e.g., 6/10). The *interpersonal equivalency* assumption states that respondents make similar interpretations of the questions, so that differences in translation functions must be due to differences in scale-use.

The *intrapersonal scale consistency* assumption states that scale-use patterns are identical across scales, so that relative threshold locations calculated using objective-subjective questions are applicable to questions about subjective life satisfaction. Although neither assumption is directly testable, we can test whether estimated scale shifts are common between types of cardinal quantities (i.e., height, income, and numbers of friends). If they are, we should increase our credence in believing that similar shifts occur in the case of life satisfaction.[7]

[7] Of course, this depends on assuming that life satisfaction is ultimately a cardinal quantity. See Plant (2020) for some arguments in favour of this assumption.

Figure 9. Reported scale-use in objective-subjective questions.



Note: *Responses to each of the objective-subjective questions are shown (see Figure 8 for an example of the exact phrasing of the questions). Each row represents a respondent's answers. The red circles represent the cardinal quantity reported by each respondent (e.g., 180 cm). The black triangles represent the objective value that respondents assigned to a 5/10. The black circles represent the objective values that respondents assigned to a 0/10 and to a 10/10. In this figure, the respondents are ordered by the cardinal quantity they reported. The line at the very bottom shows sample averages of each value. Substantial variation in scale-use is visible. However, across each of the domains (height, friends, income), the cardinal quantity that respondents assign to the mid- and end-points looks to be uncorrelated with respondents' own levels of that quantity (e.g., respondents' own height appears uncorrelated with what respondents judge the 0/10, 5/10, and 10/10 points). Some of the top end-points on the income question are very large and therefore note shown.*

4.4.3 Analyses

We intend at least three main statistical analyses using these questions:

- A joint model of life satisfaction and scale-use, exploiting individual differences in responses from each of the calibration questions. This model will be essentially analogous to the model used for vignette data.
- A test of the equivalence assumption analogous to those in the vignette and calibration methods.

- A test of whether respondents' stated scale-use is correlated with socio-demographic characteristics and/or their own (cardinal) height, income, and number of friends.

4.4.4 Feasibility results

The graphs below summarise the responses to our objective-subjective questions (see Figure 9). The variables do not appear to have a systematic relationship between individuals' actual objective height/friends/income and the scale that they report using. In other words, as height/friends/income increases – as the red circle goes further right – the other elements of the scale (the 0, 5, and 10 points) do not systematically move in a certain direction. We see this as tentative evidence pointing towards comparability.

4.5 Neutral point method

4.5.1 Methodology

As discussed in Section 1.3, the neutral point designates a level of latent SWB that is comparable to non-existence. This may be related to the 'zero point', where SWB is neither positive nor negative. We ask respondents about both the neutral point and the zero point, and investigate whether they coincide. Specifically, we ask respondents:

1. Suppose you were just as satisfied as you were dissatisfied with your life overall. Where on the life satisfaction scale would you put yourself? Here, 0 means extremely dissatisfied and 10 means extremely satisfied.

2. At what level of life satisfaction would your life cease being worth living for you, disregarding any effects your life might have on others? Here, 0 means extremely dissatisfied and 10 means extremely satisfied.

Further details on these questions are provided in Section 6. We are especially interested in whether this novel neutral point method for comparability yields similar results to the older and more established vignette method.

4.5.2 Assumptions

With the neutral-point method, assumptions are analogous to those in the vignette method. The *interpersonal equivalence assumption* states that participants have equivalent interpretations of the neutral point, so that neutral-point reports all refer to the same latent level of life satisfaction. Any differences in these reports would then be due to differences in scale-use.

The *intrapersonal scale consistency* assumption states that respondents adopt the same scale-use patterns when answering the neutral point question as when rating their own life satisfaction.

4.5.3 Analysis

The neutral point method is analysed in exactly the same manner as the vignette method, replacing the vignette assessments with responses to either neutrality question.

4.6 Endpoint questions

4.6.1 Methodology

One potential source of interpersonal incomparability that could be especially important is the possibility of participants assigning different meanings to the endpoints on their individual life satisfaction scales. In order to evaluate this possibility, we simply ask participants how they would define both endpoints. We offer respondents some pre-made options. We also allow for an open text field. The choices are ordered by how encompassing they are. In other words, the first option we offer is the most conceptually encompassing (defining the endpoints in terms of counterfactual possibility) while the last option we offer is least encompassing (defining the endpoints in terms of the respondent's own past).[8]

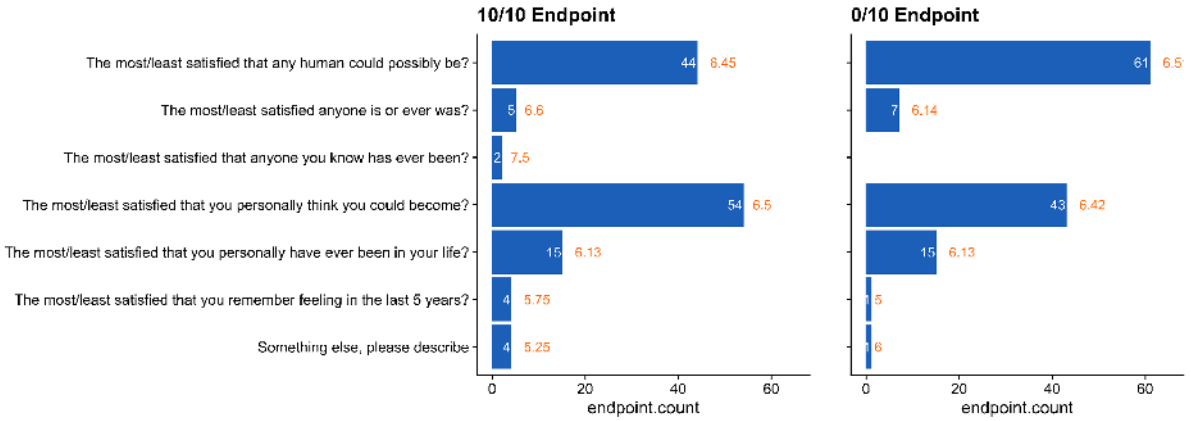
The question for the upper endpoint (i.e., 10/10) is shown below. The question for the lower endpoint is structured analogously.

Earlier, you indicated your level of life satisfaction on a 0 to 10 scale. Do you think of a score of ***10 out of 10*** *as:*

- *The most satisfied that any human could possibly be?*
- *The most satisfied anyone is or ever was?*
- *The most satisfied that anyone you know has ever been?*
- *The most satisfied that you personally think you could become?*
- *The most satisfied that you personally have ever been in your life?*
- *The most satisfied that you remember feeling in the last 5 years?*
- *Something else, please describe:* _____

[8] We are aware that our first option ("most/least satisfied any human could *possibly* be") leaves it open what exact sense of 'possibly' we mean. Some specifications: actually possible (possible within the real world), nomologically possible (possible within the laws on nature; for example, including development of future technology), logically possible (possible in principle; there is no reason to think there is any logical limit to satisfaction, just as there is not logically largest number). We do not provide a more precisely specified response option as we feared that this would overburden participants. We might distinguish between these senses of 'possible' in future work.

Figure 10. Interpretation selected for the endpoints of the life satisfaction scale



Note: *For both the bottom and top endpoint, responses cluster around two options: (1) the most/least satisfied any human could possibly be (34.4% for 10/10 and 47.7% for 0/10) and (2) the most/least satisfied the respondent could personally become (42.2% for 10/10 and 33.6% for 0/10). The numbers in orange represent the mean life satisfaction score of participants selecting the related endpoint option.*

4.6.2 Analysis

Those who adopt the most encompassing endpoints should rate their life satisfaction closer to the middle of the scale than those who report less encompassing endpoints. Similarly, respondents who report a more encompassing 10 and a less encompassing 0 - for example by reporting that 10 refers to the most satisfied any person could possibly be, and the 0 the least satisfied they have ever felt - should have lower average life satisfaction, and vice versa. We will therefore assess whether response behaviour is in line with these intuitions.

Building on that, we will regress endpoint answers on demographics, in order to see if variables typically associated with higher life satisfaction, such as income or employment status, are associated with using more or less encompassing endpoints, which may in turn be indicative of scale shifts.

We may also assess the extent to which endpoint answers correlate with answers on the vignette questions of Section 4.2.

4.6.3 Feasibility results

In Figure 10, we present the endpoint responses that participants selected. In general, participants either select the end point relating to how "*any human could possibly be*" (34.4% for 10/10 and 47.7% for 0/10) or relating to how satisfied "*they themselves could possibly become*" (42.2% for 10/10 and 33.6% for 0/10). The numbers in orange represent the mean life satisfaction score of participants selecting the related endpoint option.

The mean life satisfaction for the participants selecting the most frequent endpoint options do not differ greatly. In this pilot, we do not have sufficient power to test our hypothesis that more encompassing endpoints will lead to lower average levels of life satisfaction.

Only 70% (n = 89) of participants select the same scope of endpoint as their answer for 0/10 and 10/10. It is unclear if this is because participants truly have different scopes for both endpoints or answer poorly. Inconsistent answers may come from low motivation and/or attention. However, another possible problem is that participants do not carefully think about endpoints when asked a life satisfaction question and that this question pushes them to provide an answer to a concept they had not first thought about.

5. Linearity questions

To test whether the relationship between latent and reported SWB is **linear**, we use four approaches.

(1) We present a number of possible relationships between response options and underlying wellbeing – including a *linear* relationship – and ask participants which one they think is most representative of their own scale-use (see Figure 11). In future versions we may also ask respondents whether they use a mixture of the options presented to them.

(2) We use the psychophysical calibration questions described in Section 4.3 to estimate linearity. Each stimuli has three different versions with differing intensity on the variable, and the cardinal values of the stimuli are equidistant. Therefore, we can directly test the assumption that scale-use is linear at the individual level.

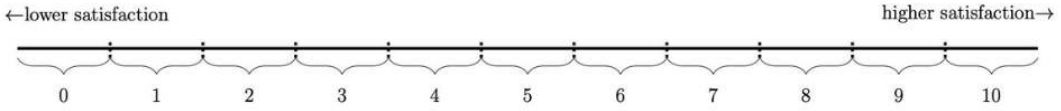
(3) Similarly, we use the objective-subjective questions described in Section 4.4 to assess linearity of translations from objective to subjective reports.

(4) Finally, we ask participants to indicate their life satisfaction again at the end of the survey, this time with a continuous slider. We test whether the relationship between discrete and continuous scales is linear. If so, it would indicate that the mere discreteness of the 10-point scale does not induce respondents to adopt non-linear scale-use.

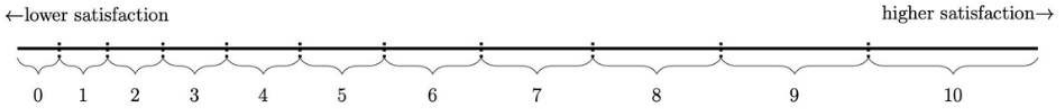
Figure 11. Question asking participants to indicate how they use the life satisfaction scale

Earlier, we asked you about your life satisfaction. Different people might answer this question in different ways.

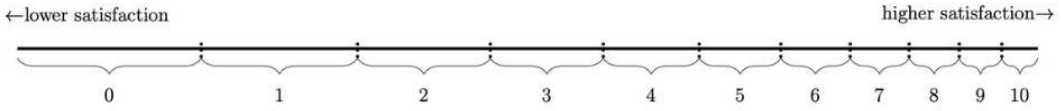
(Option 1) Some people might answer questions about life satisfaction **"linearly"**. These people assign the same "distances" between response categories: For these people, the difference in life satisfaction between for example the 3rd and the 4th response category, is the same as the difference between the 5th and the 6th response category. This sort of response behaviour is pictured below:



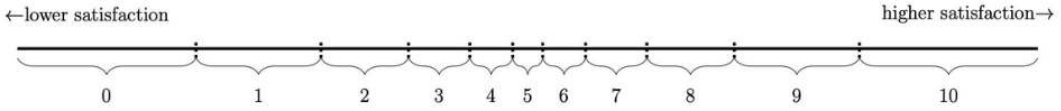
(Option 2) Some people might answer in a way where differences between response categories **increase**:



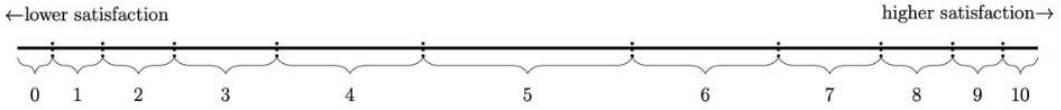
(Option 3) Some people might answer in a way where differences between responses categories **decrease**:



(Option 4) Some people might give **large differences to the outer ends** of the response scale, and small differences to options in the middle:



(Option 5) Some people might assign small differences to the outer ends of the scale, and **large differences to the middle of the scale**:



Which of these options comes closest to your way of answering the question?

Note: *This is a screenshot of the question in the exact format presented to respondents.*

5.1 Pilot results

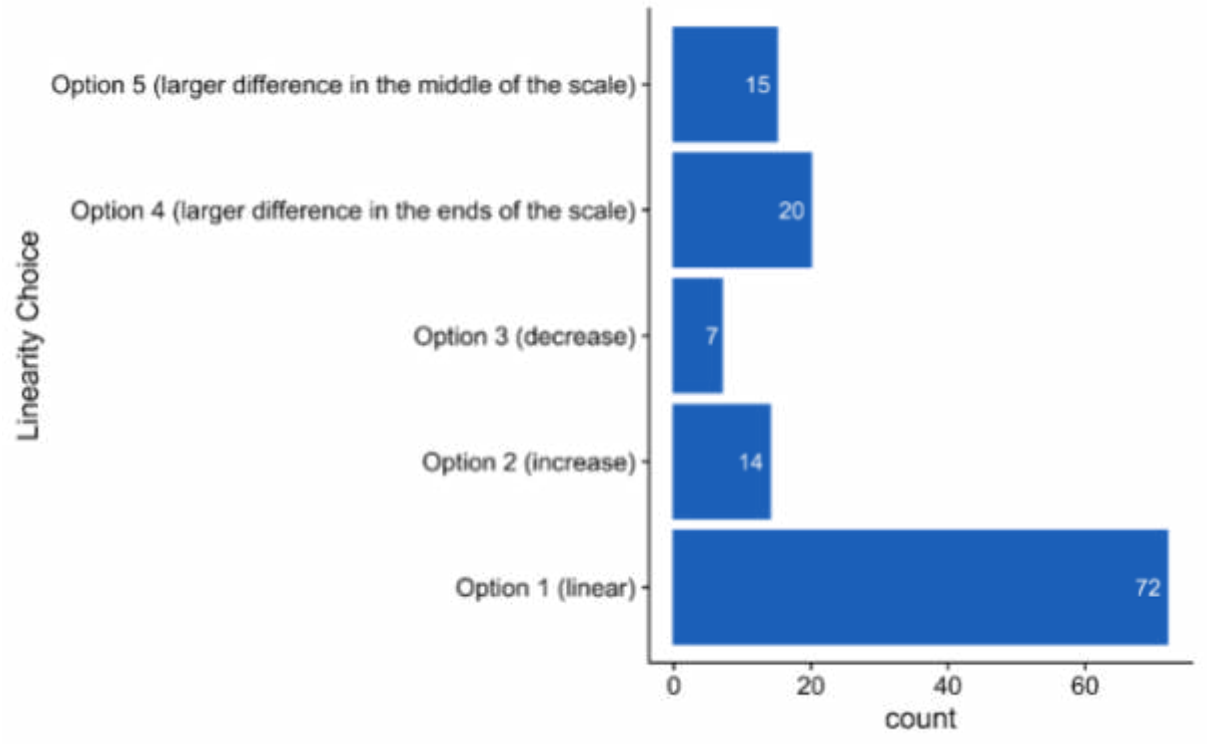
(1) When asked how they use the life satisfaction scale, 56.2% of participants reported that they used the scale linearly (see Figure 12). This is evidence in favour of linearity for at least a subset of respondents. That said, just because participants *report* that they do something does not mean they necessarily do so.

(2) For the calibration questions we plot the cardinal (objective) value of each kind of stimuli on the y-axis (e.g., how obtuse the angle is in degrees) and the subjective response (0-10) on the x-axis (see Figure 13). The size of the dots represents the number of participants giving this subjective answer for that cardinal value. For the largest part of the scale, we do not see strong departures from linearity.

(3) For the objective-subjective questions, Figure 14 shows the cardinal (objective) value of the height, number of friends, and income on the y-axis and the subjective response (0-10) on the x-axis. Each point represents a different participant response. Similarly to Figure 12, we do not see strong departures from linearity. Table 4 shows what cardinal values respondents assigned to a "0", a "5", and a "10" points on the objective-subjective questions. Using this information, we can observe how linear these interpretations of the scale are. For height, it seems that scale-use is linear: the average middle point (5/10) is approximately at the same distance from both average endpoints. For number of friends and income these averages are somewhat further away from linearity.

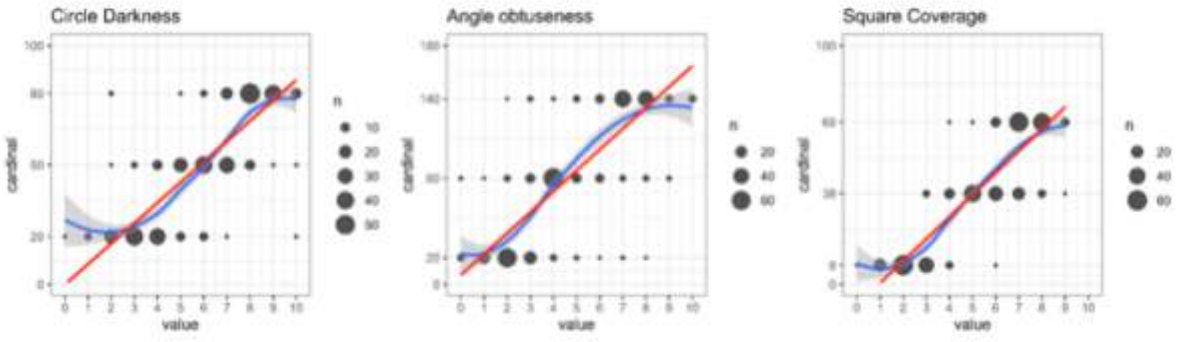
(4) Finally, we plot each participant's response on a discrete 11-point life satisfaction scale against their response on a continuous scale (which really has a resolution of 10,000 discrete values). Based on Figure 15 below, we see that participants give similar life satisfaction answers when they are given a continuous slider as when they are given discrete integers to choose from. For example, respondents who gave a 5 on the 11-point scale, tended to give - on average - a 5072 on the 10,000-point scale. If participants gave exactly the same answers, their answers would align on the black diagonal line. These pilot results suggest linearity in the use of subjective scales. However, it might be that respondents interpret the continuous scale non-linearly in the same fashion as they might be using the discrete scale.

Figure 12. Distribution of responses on how participants use the life satisfaction scale



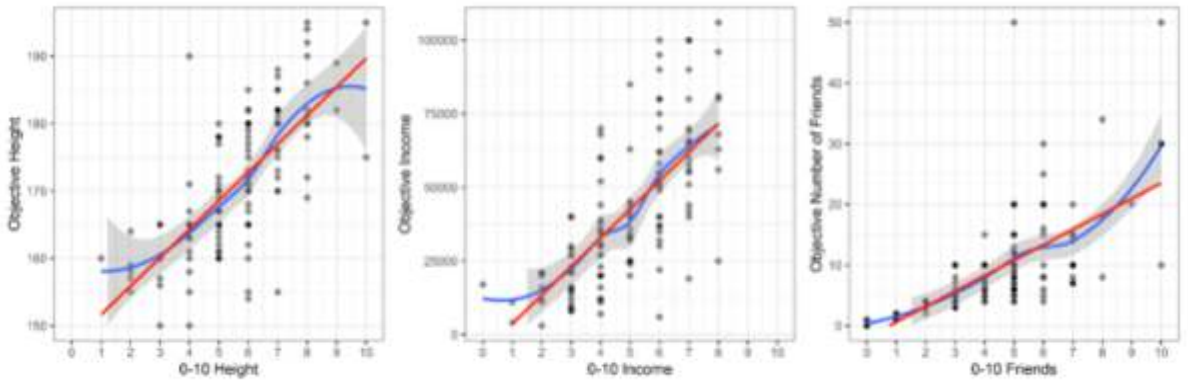
Note: *Most respondents (56.2%) report that they use the scale in a linear manner.*

Figure 13. Linearity in responses to calibration stimuli



Note: *Responses to each of our calibration questions are shown. The cardinal value of each kind of stimuli is shown on the y-axis. The subjective responses, measured on a 0-10 scale, are shown on the x-axis. The blue line and grey shading represent a fitted lowess function and the red line represents a linear fit. For most response options, the lowess function generally follows the linear fit, indicating linearity of scale. However, for the top and bottom response categories, the lowess fit flattens, resembling a sigmoid function. This may be due to a genuine non-linearity in scale-use or due to the limited variation in the cardinal values of our calibration question. The size of the dots represents the number of participants giving this subjective answer for each cardinal value.*

Figure 14. Linearity in responses to the objective-subjective questions



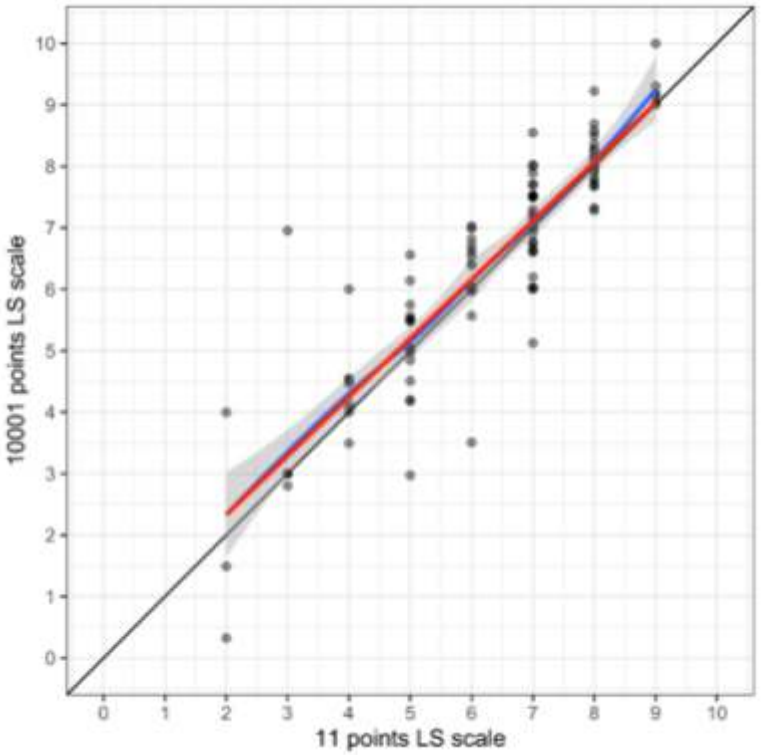
Note: *Responses to each of our objective-subjective questions are shown. The cardinal values for height, income, and number of friends are shown on the respective y-axes. The subjective responses, measured on a 0-10 scale, are shown on the x-axes. The blue line and grey shading represent a fitted lowess function and the red line represents a linear fit. For most response options, the lowess function follows the linear fit, indicating linearity of scale-use. The dots represent raw responses. For clarity of presentation, one participant who reported a height below 130cm, one participant who reported to have more than 55 friends, and one respondent who reported an income above 250,000 GBP were removed.*

Table 4. Summary averages of participants 0, 5, and 10 points to the objective-subjective questions

variable	0 point	5 point	10 point
Height	138.46 (19.08)	169.03 (10.37)	196.92 (16.72)
Friends	1.82 (1.63)	9.89 (6.90)	25.68 (27.16)
Income	12352.00 (7431.02)	32888.00 (13204.37)	176960.10 (278397.48)

Note: *Average responses for the endpoints and the middle point of the scale are shown. Standard deviations are shown in parentheses.*

Figure 15. Comparing life satisfaction on an 11-point scale and a 10,001-point scale



Note: *The dots show life-satisfaction responses of participants to both the continuous and the discrete response scale. The blue line and grey shading represent a fitted lowess function and the red line represents a linear fit. The discrete responses are clearly linear in the continuous responses.*

6. Neutrality questions

To better understand the neutral point, that is the place on a SWB scale at which continued existence is neutral in value for the respondent, we ask two questions.

First, we ask where respondents would place themselves if they were neither satisfied nor dissatisfied with their life on the life satisfaction scale. Our hope is that this question will give valid estimates of the zero point (i.e., the point at which a person feels zero satisfaction).

Second, we ask respondents to indicate the point on a life satisfaction scale where life would no longer be worth living for them (considering just themselves and ignoring the effects on others).

As noted, one theoretically coherent position is that it is good for you to live if your wellbeing is positive, and bad for you if your wellbeing is negative. Hence, if life satisfaction is taken to be the correct measure of wellbeing, we might expect the *zero point* for life satisfaction (i.e., being neither satisfied nor dissatisfied) to coincide with the *neutral point* on the life satisfaction scale (where your existence is comparable in value to non-existence).

The first question (**N1**) reads:

> *Suppose you were just as satisfied as you were dissatisfied with your life overall. Where on the life satisfaction scale would you put yourself?*
>
> *Here, 0 means "extremely dissatisfied" and 10 means "extremely satisfied".*

The second question (**N2**) is asked in two different ways[9], randomised between respondents. Half receive the following prompt (N2free):

> *At what level of life satisfaction would your life cease being worth living for you, disregarding any effects your life might have on others?*
>
> *Here, 0 means "extremely dissatisfied" and 10 means "extremely satisfied".*

The other half receive a binary counterpart (**N2binary**) where they are provided with a value between 0 and 5 out of 10 and asked if a life at that level would be worth living:

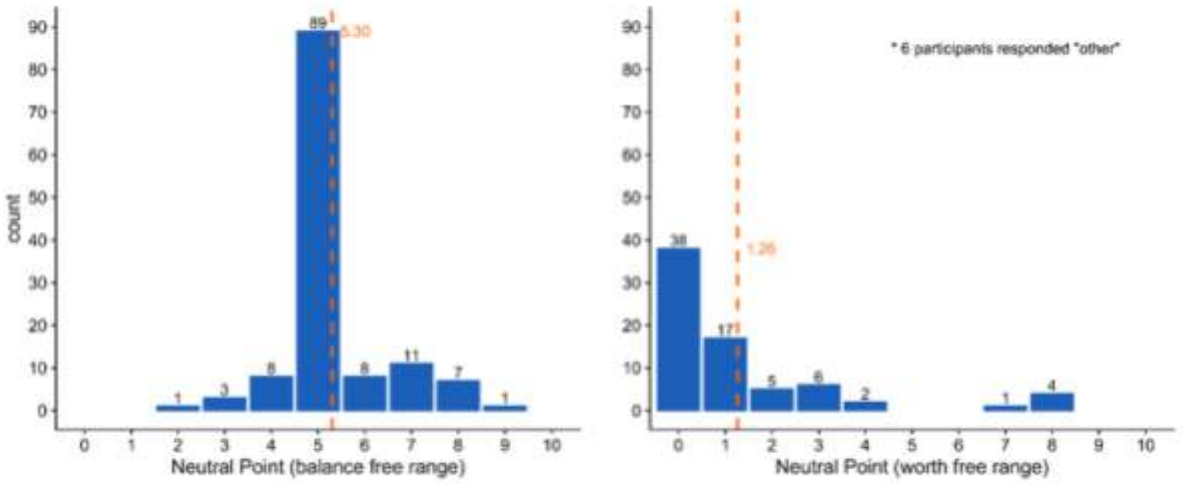
> *Suppose your life satisfaction would be [0/1/2/3/4/5] out of 10 for the rest of your life.*
>
> *Disregarding any effects your life might have on others, would this life be worth living for you?*
>
> *Yes/No/Other:___*

Both questions have an "Other" option, so that participants are not forced to indicate a neutral point.

In analysing the neutral point, we intend, apart from basic descriptive statistics, to model the predictors of the neutral point using the same sorts of demographics that are typically used to study SWB. We will also investigate the relationship between reported life satisfaction and the neutral point.

[9] We split the question because a series of pilot iterations of the survey suggested that neutrality questions are very cognitively taxing and need to be as simple as possible in the survey format. The binary question is a further simplification.

Figure 16. Distribution of responses to neutral point questions



Note: *Distributions of responses to our two main neutrality questions are shown. The dashed lines and numbers show sample means. The left panel shows responses to the question "Suppose you were just as satisfied as you were dissatisfied with your life overall. Where on the life satisfaction scale would you put yourself?". The vast majority (69.5%) select 5/10 as the point of neutrality in this question. The right panel shows responses to the question "At what level of life satisfaction would your life cease being worth living for you, disregarding any effects your life might have on others?". Although most respondents (48.1%) select 0/10, the distribution of responses is clearly left-skewed, thereby pushing up the mean.*

6.1 Pilot results

All participants answered **N1**. Via random allocation[10], 79 participants answered **N2free** and 49 participants responded to **N2binary**. Regarding the latter question, there are unfortunately too few participants to draw out detailed findings. We no longer intend to use this question in future iterations of the survey[11]. In Figure 16 we present the distribution of responses to **N1** and **N2free**.

For **N1**, participants set the balance of satisfied and dissatisfied - on average - at 5.30 (SD = 1.08). They answered this question in very similar ways, with 69.5% answering that the point of neutrality was 5/10. This pattern suggests that the zero point, being neither satisfied nor dissatisfied, is approximately located at 5/10. However, we cannot assess the extent to which respondents chose this option merely due to demand effects (i.e., due to believing that this is the 'correct' answer or the answer they think we expected of them). If we regress the neutral point on people's life satisfaction,

[10] 51 participants were allocated via random allocation. We then recruited the next 28 participants only for the free ranging question (and not the binary question) because the binary question would demand too many more participants to give initial insights and we preferred to focus on the free ranging question.

[11] A survey focusing solely on uncovering the neutral point might explore all sorts of binary choice questions (like this one) as well as trade off questions.

we find that higher levels of life satisfaction are significantly related to higher neutral points. On average, an increase in life satisfaction of 1 point is associated with a 0.19 increase in the neutral point (p = .001). Related to the arguments of Section 4.5, this may be indicative of variation in respondents' scale-use. There was no significant effect of the randomisation (whether N1 or N2 came first; p = .818). There was no significant effect of whether participants answered our maths proficiency question correctly or not (p = .197).[12]

For **N2free**, most participants selected a score of point of 0 (48.1%) or 1 (21.5%). If we disregard the 'other' responses and treat the responses as a continuous variable, then the average score reported is 1.26 (SD = 2.08). When we regress this point on people's life satisfaction, the coefficient is not statistically significant (coefficient = -0.16, p = .290). There was no significant effect of the randomisation (whether N1 or N2 came first; p = .421) and no significant effect of whether participants correctly answered our maths proficiency question correctly (p = .457).

What we tentatively conclude from this is that people think their own life would only stop being 'worth living' if they were very dissatisfied with it overall and that people do not align the neutral point with zero point of life satisfaction. The next section sets out some possible interpretations of this initial finding.

6.2 Limitations

Survey questions about the neutral point cannot settle the debate about when a life is not worth living on their own. At most, we can elicit respondents' beliefs and preferences about their own lives, which, at least on some moral theories, will be an important determinant of the neutral point.

As noted in Section 1.3, an intuitive thought about the value of extending lives is that living longer is good for you if you would have positive wellbeing, and bad for you if you would have negative wellbeing. Yet, the above result seems in tension with this. We might expect someone who is overall dissatisfied with their life would say that they have negative wellbeing. However, if you have negative wellbeing, shouldn't you also believe that living longer would be bad for you? Note that we have specifically asked respondents to disregard the effects of their existence on others. This should rule out participants thinking their life is "worth living" merely because their existence is good for other people (e.g., their family). In light of these arguments, how might we explain the divergences between the reported neutral point and the reported zero point?

One reason participants might not align the neutral point with the zero point on a life satisfaction scale is that participants are not endorsing a life satisfaction theory of wellbeing. That is, respondents

[12] We asked participants "Which risk is larger: '1 in 100' or '2 in 1,000'?", adapted from GiveWell (2019).

do not believe that their wellbeing is wholly represented by their overall satisfaction with life. If, for instance, participants ultimately valued their happiness, and they expect that they would be happy even if they were very dissatisfied, then they would justifiably conclude that even if they were dissatisfied with their life, it would be worth it, for them, to keep living.

A distinct possibility is that there are special theoretical considerations connected to the life satisfaction theory of wellbeing. Part of the motivation for such a theory is that individuals get to choose what makes their lives go well (Sumner 1999, Plant 2020). Hence, perhaps individuals could coherently distinguish between the level of satisfaction at which they would rather stop existing (the neutral point) and the level at which they are neither overall satisfied or dissatisfied (the zero point).[13]

What the previous two comments reveal is that investigations into the neutral point may substantially turn on philosophical assumptions. We may need a 'theory-led' approach, where we decide what theory of wellbeing we think is correct, and then consider how, given a particular theory of wellbeing, the location of the neutral point would be determined. This would contrast with a 'data-led' approach where we strive to be theory agnostic.

Another concern is that our survey questions are too cognitively demanding for participants. Perhaps respondents do not, in fact, account for the stipulation that they should discount the effects on others. Alternatively, respondents might answer our questions on the assumption that their life would only temporarily, rather than permanently, be at that level of life satisfaction. In future iterations of our survey, we may try to get a better understanding of this possibility.

Finally, there may be experimenter demand effects. Respondents may think it is important to express the value of human life, and therefore it is wrong to say a life is 'not worth living' unless a life is truly miserable.

For these reasons, we remain unsure whether a level of 1.3/10, which is the sample mean on our neutrality question, indeed gives a valid estimate of the 'true' neutral point.

7. Discussion of feasibility

Before this pilot, we recruited a smaller sample of 24 participants where we discovered that participants were giving nonsensical responses to some of the more cognitively demanding questions (e.g., large proportions of very high answers to the neutrality questions). This led us to reduce the size and cognitive demandingness of our survey – by reducing the number of questions and simplifying

[13] Note that on a hedonist theory of wellbeing, where wellbeing consists in happiness, it does not seem possible to make this move: your (continued) existence is good if you are happy, bad if you are unhappy, whatever your opinions on the matter.

the text of some questions – before starting again with the current version of the survey. We find that participants were giving much more reasonable answers in this version.

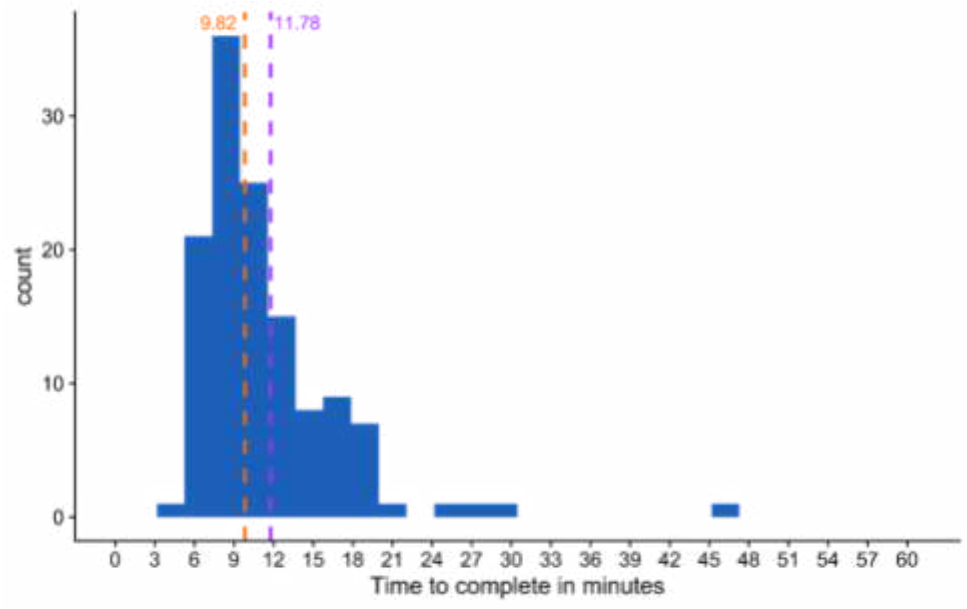
One factor that reassures us about the feasibility of this survey is participants' response times (see Figure 17). Whilst 29 participants completed the survey in 8 minutes or less - which surprised us - there is no significant relationship between completion time and answers on essential questions like life satisfaction or answers on cognitively demanding questions like the neutral point. The timing for individual questions follows expected patterns. Answering the life satisfaction question (median 6.51 seconds) is faster than answering arguably more cognitively demanding questions like the neutral free ranging question (median 12.5 seconds). Participants are slower on the first endpoint question (how they represent 10/10 life satisfaction; median 24.87 seconds) than on the second endpoint question (how they represent 0/10 life satisfaction; median 11.35 seconds), which suggests that they took the time to understand the first question before transferring that understanding to the next question.

Moreover, the distributions of answers to core questions are all reasonable. The distribution of life satisfaction responses is close to other UK samples and exhibits the well-known and commonly observed left-skew. Answers to the neutral point questions seem sensible; very few participants gave very high neutral points. For the free-ranging neutral point question (N2free), 82.19% of participants gave a neutral point equal to or below 2/10. Responses to vignettes (especially those of our own creation) and calibration stimuli seem to be distributed as expected and correctly ordered. We encountered some other issues concerning the Objective-Subjective questions, but not that would change our mind about the feasibility of the survey (see external Appendix B).

We have collected a range of demographic variables. Their summary statistics look reasonably representative (see our external Appendix A). This suggests that we can collect good enough samples from Prolific to run regressions over standard demographics. We do not do so in this pilot because of the small size of our sample, but we will be able to do so with a larger sample size.

Finally, in this survey we use a life satisfaction question. Versions of this survey could be made using other wellbeing questions to test whether the findings hold for other measures (and theories) of wellbeing, such as happiness questions for example. Using more than one type of wellbeing question in the same survey would overwhelm respondents.

Figure 17. Distribution of times to complete the whole survey in minutes.



Note: *The distribution of time respondents to complete the survey is shown. The orange line presents the median, the purple line presents the mean.*

8. Conclusion

Assumptions of linearity and comparability are key for research and decision-making based on subjective wellbeing data. It seems that disbelief in these assumptions is currently hindering the wider adoption of wellbeing data for cost-effectiveness analyses. The neutral point is crucial when comparing life-extending and life-improving decisions.

We proposed a survey design with a range of methods and questions, and we collected pilot data to assess the feasibility of this survey. Although we cannot yet provide detailed results on comparability, our initial pilot findings suggest that respondents appear to interpret response scales in a close-to-linear manner. Concerning neutrality, our results suggest a neutral point around 1 on a scale from 0 to 10.

Our next steps will be to collect a much larger sample in order to provide more definitive results. Before then, we are interested in feedback about our survey design (see the start of this report for some questions we are interested in).