



IARIW 2024

IARIW 2024

Thursday, August 22 – Friday, August 30

Evidence Against the Simple Validity of Life Satisfaction Scales from Long Cognitive Interviews

Mark Otto Fabian

(University of Warwick)

Caspar Kaiser

(University of Warwick)

Paper prepared for the 38th IARIW General Conference
August 26-30, 2024

Session 6B-2, SWB Measurement Challenges

Time: Friday, August 30, 2024 [11:00-12:30 GMT]

Evidence against the simple validity of life satisfaction scales from long cognitive interviews

Mark Fabian, University of Warwick (corresponding author: mark.fabian@warwick.ac.uk)

Caspar Kaiser, University of Warwick

Sofia Panasiuk, University of Toronto

Sabina Funk, University of Warwick

Liana Pountney, University of Warwick

Caroline Brett, University of Edinburgh

ABSTRACT

Statistical analysis of life satisfaction data standardly assumes 1) linear scale use 2) interpersonal comparability 3) intertemporal comparability. To interrogate the credibility of these assumptions, we need to understand the reporting function: the process by which individuals make life evaluations and then map them to a category on the response scale. We develop a formal model of the reporting function that informs questions in long cognitive interviews of a diverse sample of 100 residents of the United Kingdom. Our results bear out previous efforts to validate life satisfaction scales. However, we also find widespread, severe, systematic, and non-random violations of all three standard assumptions. In particular, many respondents do not use the top of the scale, major shocks appear to alter the meaning of the points on the scale, older people are more likely to interpret life satisfaction questions as referring to how life *went* rather than how it is going, and respondents rarely describe their own scale use as linear.

Keywords: life satisfaction, measurement, subjective wellbeing, psychometric validity, cognitive interviewing

JEL Codes: I31, C18

THIS PAPER IS A FIRST DRAFT AND ABSOLUTELY NOT FOR CITATION OR DISSEMINATION. PLEASE CONTACT THE LEAD AUTHOR FOR ANY QUERIES AND AN UP-TO-DATE VERSION.

1. Introduction

Statistical analysis of life satisfaction scale data standardly makes three assumptions:

1. That respondents use the response scale in a linear way (linearity)
2. That different respondents use the scale the same way (interpersonal comparability)
3. That respondents use the scale the same way over time (intertemporal comparability)

The first of these assumptions was recently challenged forcefully in the economics literature by Schröder & Yitzhaki (2017) and Bond and Lang (2019).¹ These authors assert that although responses from life satisfaction questions may contain ordinal information, such data does not carry cardinal information. Concretely, they claim that although a reported satisfaction level of e.g. 7/10 indicates a greater latent level of life satisfaction than e.g. a 6/10, there is no information on whether the difference in latent satisfaction between e.g. a 3/10 and a 4/10 is just as large as the difference between e.g. a 5/10 and a 6/10. They then show that in many instances – though not all (Kaiser & Vendrik 2023) – making different assumptions about the relative sizes of differences in response categories can alter the sign of estimated coefficients in regression models.² This is crucial, as it implies that most basic results from the literature on life satisfaction cannot be relied upon. More generally, their arguments imply that results relying on *any kind of* subjective survey data may be sensitive to nonlinear scale use.

These arguments echo an old and much litigated debate in psychometrics concerning whether Likert scale data can be treated as interval (i.e. linear/cardinal) or only ordinal (Larroulet Philippi 2022). That is, whether the response categories on such scales are equidistant from each other in a quantitative sense or merely constitute an ordered ranking. One of the most recent broadsides in this debate is Liddell & Kruschke (2018). They review papers across three top psychology journals and find that all of the papers therein utilising ordinal Likert scale data used methods only appropriate for interval data. They demonstrate that this can lead to biased statistical estimates and recommend ordered probit methods be used instead. In the context of life satisfaction scale research, Frijters & Ferrer-i-Carbonell (2004) conducted an influential study that found little if any noteworthy effects on life satisfaction analysis from assuming cardinality or ordinality of the underlying data. Happiness economists have nonetheless tended to use ordered probit techniques in their work, but concerns about cardinality were long ago thought put to bed.

However, it can be shown that ordinary least squares and ordered probit methods rely on assumptions that, in most empirical applications, are closely related. In particular, as we show in Appendix A0, the ordered probit assumption of a normal distribution for latent satisfaction implies *linear scale-use when reported satisfaction is normally distributed*. However, as is well-

¹ Strictly speaking, Bond & Lang (2019) contains two arguments. One argument relates to the fact that responses should only be interpreted ordinally. The other argument focuses on the possibility that there may be variation in latent satisfaction within a given response category. We here focus on the former of the two arguments.

² To put the claim slightly more formally: Suppose one obtains responses to a life satisfaction question with K ordered response options $\{a_1, a_2, \dots, a_K\}$. In many parts of the literature, we now label each response option in their rank-order, i.e. we set $a_k = k$ for all $k = 1, 2, \dots, K$. We then run a regression of these labelled responses, call them r_i , on a number of covariates $x_i, \dots, x_{mi}, \dots, x_{Mi}$ to obtain coefficient estimates $\hat{\beta}_1, \dots, \hat{\beta}_m, \dots, \hat{\beta}_M$. The claim is that it is often possible to obtain coefficients $\hat{\beta}_m$ with $\text{sign}(\hat{\beta}_m) \neq \text{sign}(\hat{\beta}_m)$ when applying some non-linear positive monotonic transformation $f(r_i)$ to obtain a new set of labelled responses $\tilde{r}_i \in \{f(1), \dots, f(k), \dots, f(K)\}$.

known, life satisfaction data is rarely, if ever, normally distributed. Instead, reported satisfaction tends to be left-skewed in most samples from rich Western countries. As also shown in appendix A0, an ordered probit model will then implicitly assume that differences in latent satisfaction across response categories tend to decrease (i.e. the difference between, say, a 1/10 and a 2/10 will be assumed to be larger than the difference between, say, an 8/10 and a 9/10). Indeed, in general, it holds that the ordered probit assumptions (or similar assumptions in a structural equation modelling framework) always imply some particular assumption about respondents' scale use.

What needs to be tested then is not just whether estimates from analysing life satisfaction data are stable across methods, but the data generating process itself, namely the interpretation life satisfaction scale questions by respondents. This paper seeks to fill this gap. It does so not only to test the linearity assumption (1), but also to investigate assumptions (2) and (3) regarding interpersonal and intertemporal differences in scale use. Violations of these assumptions are potentially much more problematic for statistical analysis than violations of non-linearity. We begin by reviewing existing work in psychology on the psychometric properties of life satisfaction scales. While encouraging, these tests are general whereas a thorough investigation of assumptions 1–3 requires modelling the life satisfaction reporting function. This is the process by which individuals assess their life satisfaction and then report it on a scale with a limited number of response categories. To this end, we develop a formal model and conceptual framework for understanding life satisfaction reporting. We use it to inform long (30–60 minutes) cognitive interviews that we conduct with a diverse sample of 93 residents of the United Kingdom. The results of these interviews can be fed back into our formal model and thereby used to check the credibility of assumptions 1–3. We find substantial, systematic, non-random violations of all three assumptions

2. Conceptual Framework

2.1. Differences in scale use

Psychometric testing of psychological measures and instruments are standard operating procedure before their widespread acceptance and adoption. The life satisfaction scale is no exception. The single-item life satisfaction measure and its multi-item predecessor, the Satisfaction with Life Scale (SWLS; Diener et al. 1985), have been subjected to a battery of tests assessing their psychometric properties since their inauguration as the flagship measures of subjective well-being. At first these tests evaluated validity, i.e. whether the life satisfaction scale indeed measured what it alleged to. Results were encouraging. Differences in life satisfaction are correlated with objective circumstances (Oswald & Wu 2009), life satisfaction is associated with other measures of subjective well-being (Diener et al. 1985), and evaluations of life satisfaction changed in a predictable direction following significant life events (Clark et al. 2008). Establishing measurement invariance for life satisfaction proved to be more difficult. Measurement invariance allows for meaningful comparison of scores between individuals and groups because differences in scores reflect differences in the latent construct as opposed to differences in question interpretation or reporting. In a review of measurement invariance evidence, Emerson et al. (2017) found equivalence of the SWLS for mean comparisons across gender, but not across age or cultural groups. For a more general review of the first round of life satisfaction validation tests, see Diener et al. (2013).

The second wave of tests followed the growing popularity of alternative theories of psychological measurement. Whereas previous tests involving correlations and descriptive statistics were underpinned by the dominant framework of Classical Test Theory (CTT), the new tests emerged from the rival frameworks of Item Response Theory (IRT) and Rasch Measurement Theory (RMT). Most importantly, these frameworks did not presume that psychological scales were cardinal. Instead, both frameworks consider variability in item difficulty and individual ability which can lead to misleading interpretations if left unaccounted for. For these more stringent tests, the life satisfaction scale could not clear the threshold for acceptability. Notably, the instrument was bad at detecting life satisfaction differences towards the top of the scale (Schutte et al., 2019). In other words, the top of the scale was less sensitive to life satisfaction differences than the bottom of the scale. This does not mean that we cannot assess whether someone is happier than they were last year. But as people get happier, estimates of life satisfaction become less precise. If indeed ‘most people are happy’ (Diener & Diener, 1996), this should be simultaneously a cause for celebration and a cause for concern regarding measurement.

2.2. The reporting function

Why might life satisfaction scales be less precise for happier people? To begin to answer this question, we need to investigate the life satisfaction reporting function. The reporting function is the affective, cognitive, social, and linguistic process that maps subjective judgements to numerical response categories on a scale question. For the case of life satisfaction scales specifically, we can specify it as follows, building on Oswald (2008), and Fleurbaey and Blanchet (2013). Let $\mathbf{l}_{it}$ represent a vector of possible lives l for individual i at time t to be satisfied about, running from relatively less to more satisfying. Then let l_{it}^* be a respondent's actual life. Finally, let $r_{it} = \{1, 2, \dots, k, \dots, K\}$ denote the response category chosen to a single life satisfaction question with K response options. The reporting function, ξ_{it} , is then a function of the individual's actual life, l_{it}^* , relative to the vector of possible lives, $\mathbf{l}_{it}$ and maps that life into r_{it} . Thus:

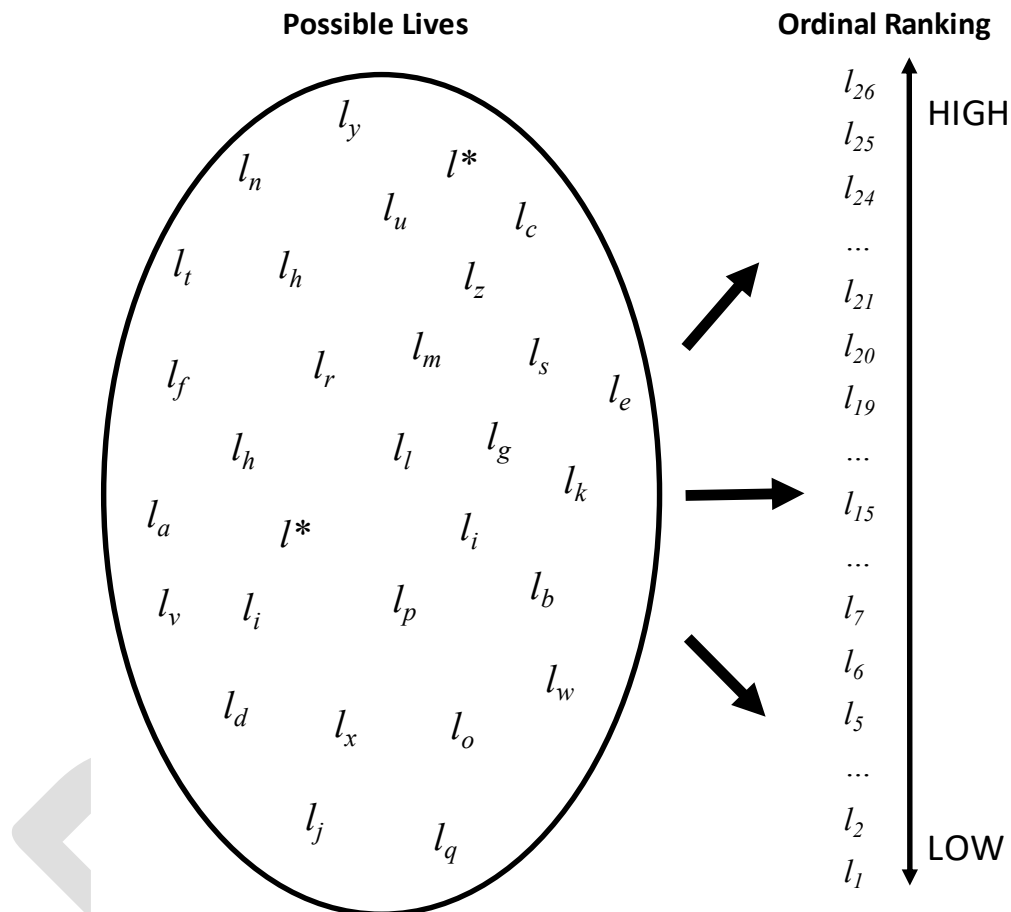
$$r_{it} = \xi_{it}(l_{it}^*; \mathbf{l}_{it}) \quad (1)$$

As researchers, we only observe life satisfaction judgements that have passed through the reporting function. It is important to conceptually distinguish the process of forming a life satisfaction judgement from the process of mapping that judgement to a response category i.e. from the reporting function. This is illustrated by Fleurbaey & Blanchet (2013), who identify three “problems” that respondents need to solve when answering life satisfaction questions.

First, the “scope problem” requires that they decide on what aspects of life to consider in their judgements (Benjamin et al. 2023). There is abundant evidence that different issues can be salient across time and respondents. For example, when life satisfaction questions are preceded by questions about politics, respondents give on average lower life satisfaction scores (Deaton & Stone 2016). Even when such explicit priming is absent, respondents will likely anchor on salient issues in their life. For some this might be family, for others career, for others religion, and so on. This might vary widely and frequently across time – a tough week at work might bring job satisfaction to the fore, for example. An additional aspect of the scope problem is whether respondents consider how the lives of others, such as family or peers, are going when forming their judgement, or whether they focus narrowly on their own life.

Second, the “ranking” problem requires respondents to weight these different aspects of life that are in scope to form the vector l_{it} i.e. an ordered ranking of possible lives. This is illustrated in figure 1. We would expect weights to differ across respondents and over time. This is often touted as one of the appeals of life satisfaction scales in the context of welfare analysis: individuals are free to decide what matters to them, rather than analysts choosing weights to associate with typically objective variables like income or life expectancy.

Figure 1: The ranking problem

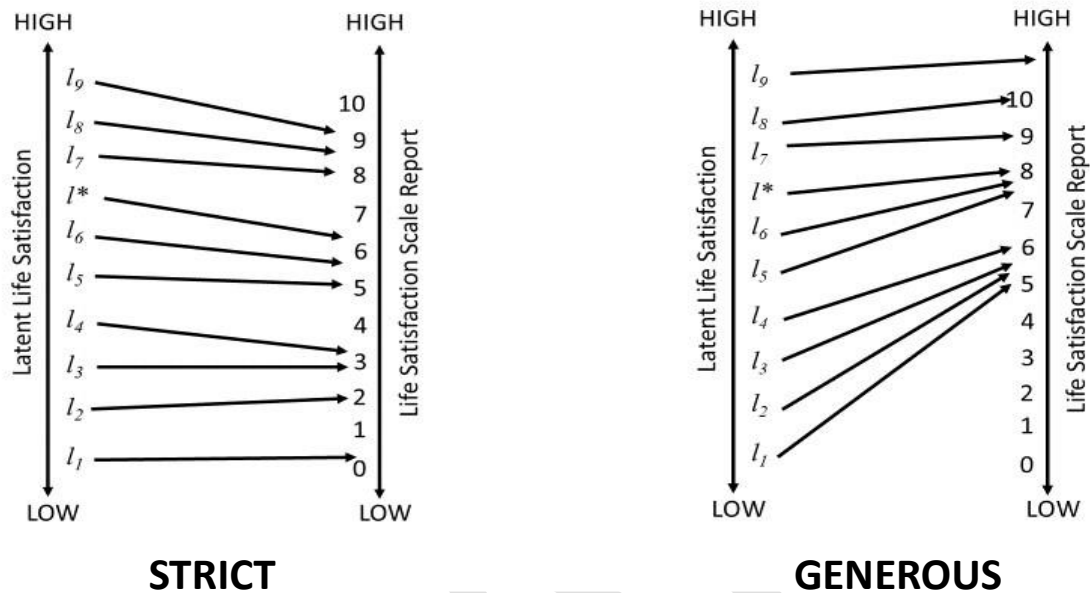


Notes: Respondents must arrange possible lives they could live, of which there are theoretically an infinite number, into an ordered ranking from least to most satisfying.

Third, the respondent must “calibrate” their ranking to the response scale – this is where the reporting function comes in. Calibration is the mapping of l_i^* relative to l_i to the scale, as illustrated in figure 2. Respondents may differ in how they do this, notably in terms of how generous or strict they are. This can be illustrated through the analogous case of generous and strict graders (Kaiser 2022). Most teaching academics will be familiar with a situation in which two graders assess a student paper in broadly the same way. They will concur on its strengths and weaknesses and its adherence or not to the marking criteria. They may even agree on its position in the class – identifying that it is in the top 10% for example (an A, or 1st). Nonetheless, they could each map this assessment into a different numerical grade. One might have a policy of giving first-class marks only near the cutoff (70 in most UK universities), while the other might reduce marks from 100 based on errors and, there being few, give the paper 88. As observers, we would be mistaken in thinking that the first marker didn’t rate the piece as highly as the

second marker. We could imagine a similar situation in life satisfaction. For example, while British respondents think of 10/10 as ideal, Clark et al. (2015) find that the “desired condition” for Japanese respondents is 8.3 on average. There are several papers using vignette methods that suggest similar differences in scale use between genders and nations (Montgomery 2022, Angelini 2014, Kapteyn 2010).

Figure 2: The calibration problem



Notes: respondents must map their ranking (figure 1) into the limited number of response categories on the scale. Some may err towards a more positive mapping than others.

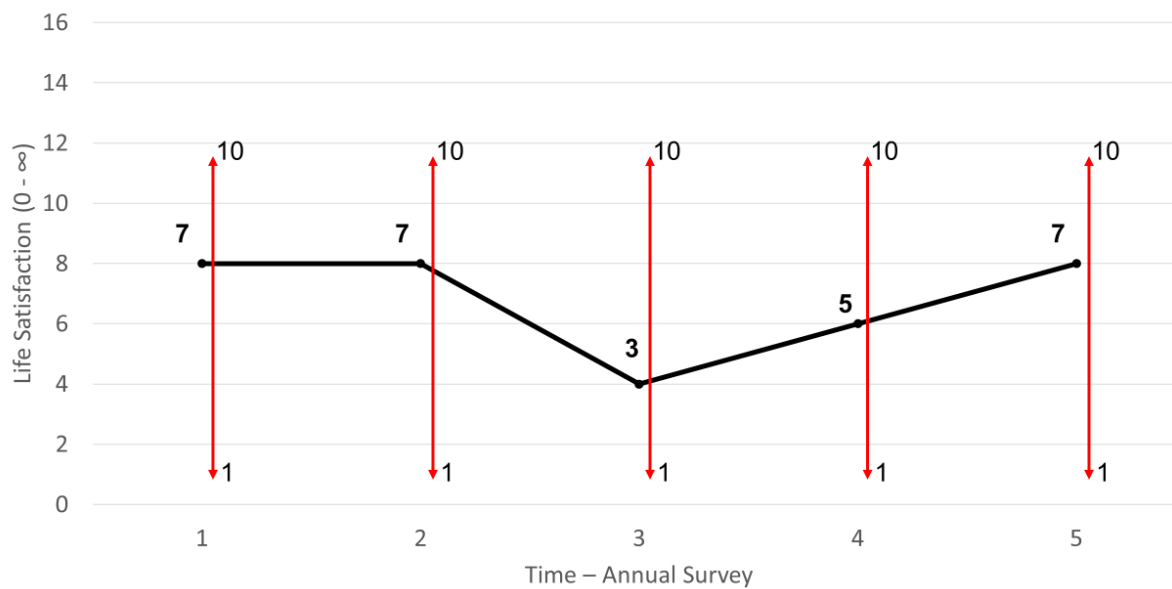
2.3. Response shift: adaptation, scale norming, and reference point effects

Many stylised facts of the life satisfaction literature are questionable in light of our underappreciation of the reporting function. In particular, much of what we now regard as adaptation might be better explained by scale norming (Fabian 2021). Adaptation involves a real change in judgements about life resulting in a return of life satisfaction to a long-run average (Lucas 2007). Adaptation could be a matter of acclimatisation, such as getting used to your appearance when your hair falls out, or a matter of working through something, as people often do after divorce. Both processes are likely involved in the conditions that originated the adaptation hypothesis, namely Brickmann and Coates’ (1978) study of lottery winners and people with spinal cord injuries. Both were seen to recover from these positive and negative shocks, respectively, within 5 years, returning close to their earlier levels of life satisfaction. Adaptation is depicted in figure 3. Note that the individual’s scale is unchanged over successive waves, only their life satisfaction judgement changes.

Brickman and Coates’ result is striking given that depressive symptoms, completed suicides, and willingness to pay for treatment are all markedly elevated among people with spinal cord injuries (Craig et al. 2009, Chan et al. 2018, Harris & Barraclough 1994). This would suggest lower subjective wellbeing. Concurring with this hunch, Lucas (2007) finds that disabilities, broadly defined, reduce life satisfaction by 0.78 points on a 0–10 scale on average. This is a large

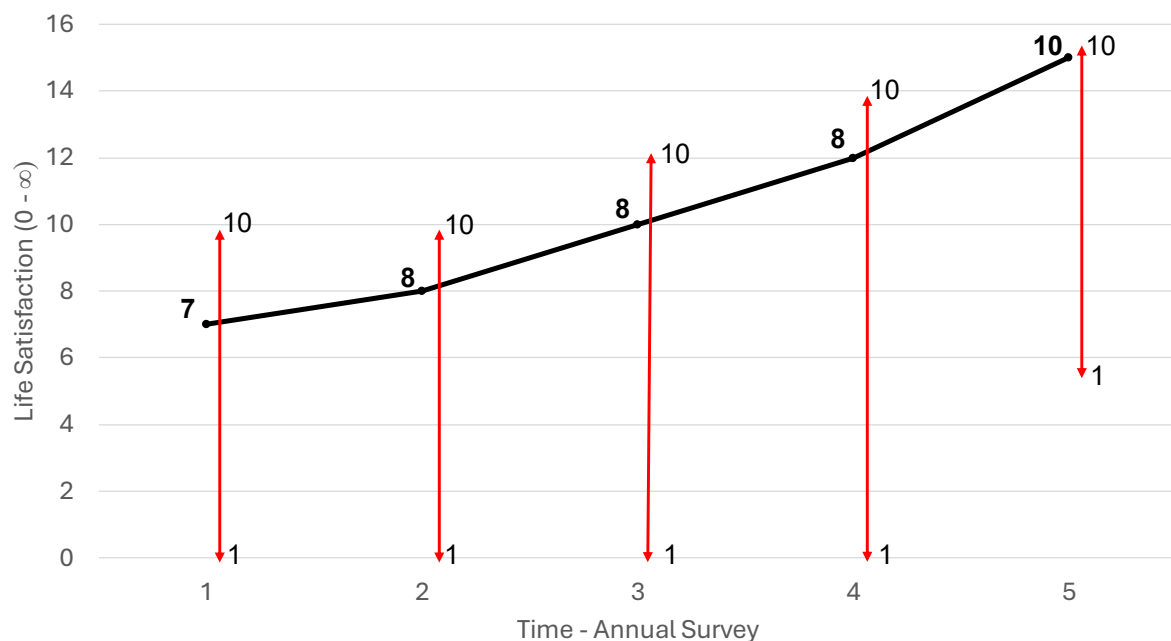
effect in the context of life satisfaction, which is notoriously insensitive to change, but seems small at face value. Might there be some measurement problem?

Figure 3: Adaptation



Notes: the respondent's scales are fixed over time, whereas their life satisfaction judgements first decline following a shock between wave 2 and 3, then adapt back to baseline by wave 5.

Figure 4: Ceiling effects



Notes: while life evaluations improve over time, the response categories do not communicate this owing to scale norming from wave 3 onward. In wave 3 and 4, the individual 'stretches' their scale, while in wave 5 they 'shift' it upward.

Scale norming is an alternate explanation for the apparently high life satisfaction of people with spinal cord injuries. It refers to a change in reporting style as distinct from a change in life satisfaction judgements. One example observed in Fabian (2021) is ceiling effects, depicted in figure 4. This is where an individual with high reported life satisfaction continues to experience improvements. They do not want to imply that they are ‘completely satisfied’ because they can foresee still further improvements, so they do not give a 10/10 response. Instead, they change the meaning of the points on their scale. Their scale either ‘stretches’ (see figure 4 between time 2 and 3) or ‘shifts’ (between time 4 and 5) such that 8/10 in time 3 and onwards now reflects a greater degree of life satisfaction than it did in time 1 and 2. The response categories chosen are not straightforwardly comparable, and doing so would bias statistical estimation. This issue is voluminously documented in the response shift literature in medical quality of life studies (see Sawatzky et al. 2024 for a review). Unfortunately, we are unable to distinguish between adaptation and scale norming in raw life satisfaction data without auxiliary questions.

A third phenomenon complicates the story further, namely reference point shifts. This is where an individual changes their reporting style (perhaps more so in the scope and ranking problem than calibration) and this changes their life satisfaction judgement. This is how gratitude and many other positive psychology interventions are supposed to work. Focusing on what is going well in your life is analogous to putting more weight in judgement on those things relative to what is going badly. You essentially introduce a cognitive bias into the judgement process, when a more ‘considered’, ‘dispassionate’, (or ‘rational’ in economics) approach might yield a different answer. This then is supposed to make you feel better. The challenge for positive psychology research is that in raw life satisfaction data we cannot tell whether that improvement in judgements has happened, or whether the individual has changed their reporting style – they might simply have shrunk or shifted down their scale. In this context, trying to prove the efficacy of interventions using statistical significance is difficult because the variable being measured hides multiple confounding effects.

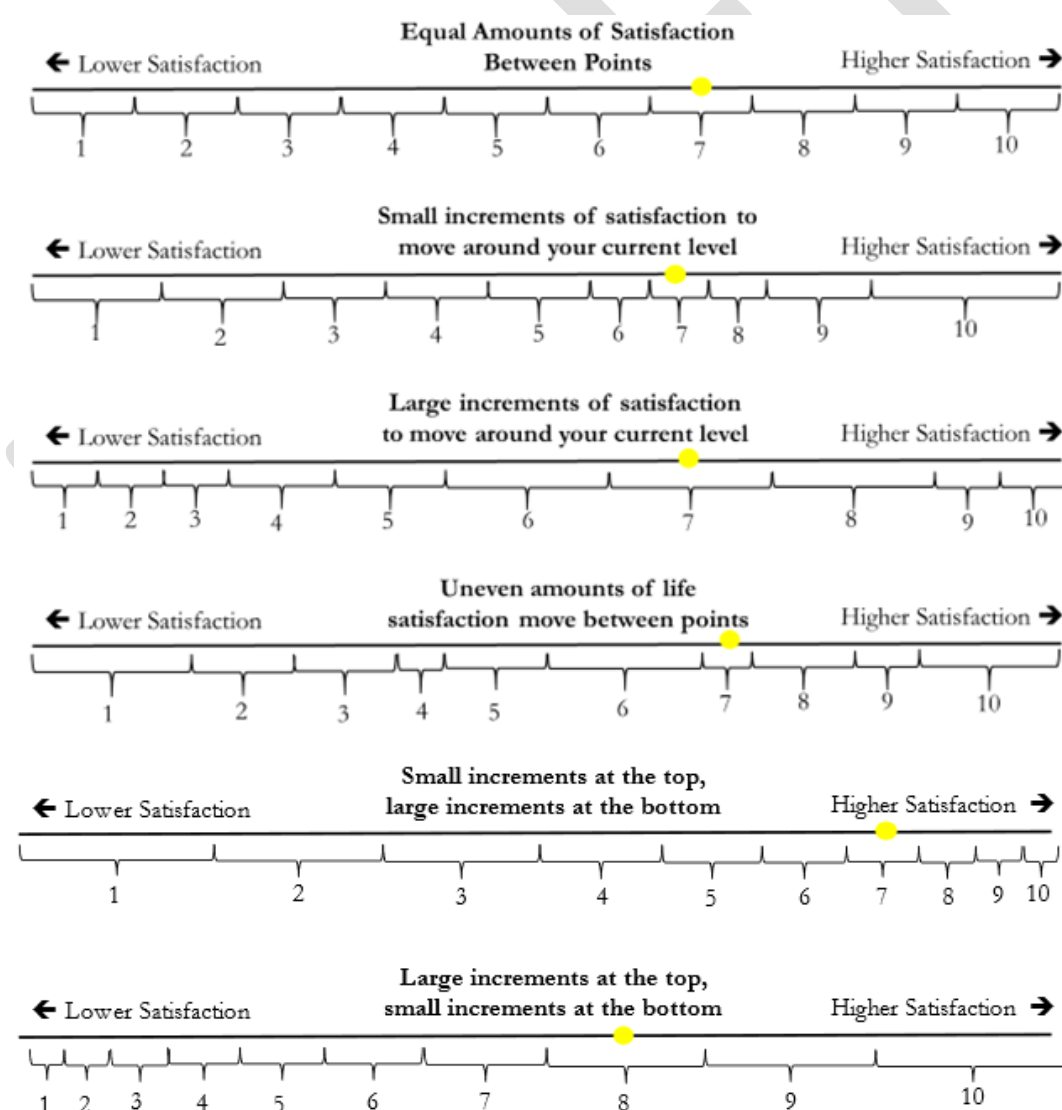
3. Methodology

Owing to our general ignorance of the life satisfaction reporting process, a range of methodologies will be needed to understand it holistically. We commence this research project here (though acknowledging that the response shift literature has already made substantial parallel progress in multidimensional quality of life studies) with cognitive interviewing (Beatty & Willis 2007). This is a qualitative method that involves having respondents “think out loud” while answering survey questions. Our approach goes beyond traditional cognitive interviewing in that we interview respondents for between 30 and 60 minutes and interrogate their answers and thought processes, rather than simply checking whether their cognitions at the time of responding align with expectations. Our methodology thus emphasises *interviewing*, and we consequently employ theory-led thematic analysis to our data (Braun & Clark 2006). The theory in question is the framework outlined in section 2, and a range of theories from philosophy and psychology about what constitutes and/or determines wellbeing, including self-determination theory, hedonism, and capabilities (see codebook in appendix A2). To develop our codebook, we first had authors 1, 3, 4, and 5 – who have different disciplinary backgrounds and views of wellbeing – all code a sub-sample of 10 transcripts and develop a harmonised codebook. We hoped that this would ensure few themes were missed. These authors then split the remaining transcripts between them for coding using this harmonised codebook. Our raw data and codebook are available open access for cross-checking.

Cognitive interviewing is a good place to start developing an understanding of life satisfaction reporting because it provides rich, contextual data that will aid in the interpretation of numbers emerging out of more quantitative approaches. Having said that, a strength of our approach is that our interview questions are designed to inform the formal model specified in section 2 and to allow direct testing of the statistical assumptions typically made when analysing life satisfaction data (see sections 5.2. and 5.3.).

The full interview schedule is provided in appendix A1. We begin with a life satisfaction question, randomly alternating between the GLS wording and Cantril's ladder of life. We then ask respondents to describe what determined their answer to that question, before investigating how they resolved the scope problem. We ask them to elaborate on the determinants of their life satisfaction and how they interpret words like 'satisfied' and 'possible'. This is to discern whether they believe life satisfaction has prudential value, which we ask them about directly. In the second half of the interview, we explore the ranking and calibration problems. We ask respondents what sort of life they would need to be living to choose different response categories, what changes would make them go up or down a point, and what 0 and 10 mean for them. We then ask them about their reporting function using figure 5.

Figure 5: Reporting functions



In the final stages of the interview, we follow up on their reporting function by asking whether there were other times in their life when they would have given the same number as today, and whether they are indifferent between these lives or prefer or other. We ask them how frequently the qualitative meaning of the points on their scale changes, and whether they nonetheless think the numbers mean the same level of satisfaction over time even if the associated lives change. Finally, we ask them whether there is anything they would like to tell us about life satisfaction, and whether they would like to change their original answer after the deliberations of the interview.

4. Data

Our data come from a purposive sample of 100 adult residents of the United Kingdom. As is typical with qualitative data, our objective was to achieve conceptual saturation and diversity rather than representativeness, as a representative sample would include a great deal of redundancy. Our aim is to glimpse all that is out there in terms of reporting style, not to establish what is typical. The sample is consequently highly diverse. With incomes ranging from <£15000 p.a. to > £110 000, age ranging from 19–75, residences in metropolitan areas, rural homes and everything in between, an even split on gender, and all major sexualities and racial groups represented. 70% of our sample is white compared to 83% of the UK, indicating our level of diversity. Our sample is especially diverse on occupation, including the unemployed and on benefits, doctors, software engineers, office managers, delivery drivers, teachers, accountants, celebrants, nurses, bartenders, gardeners, and a retired probation officer, among others. Full descriptive statistics are available in appendix A2. While 100 interviews is not a large sample for quantitative work, it is substantial for a qualitative project. Recruitment was piecemeal and opportunistic. We used personal networks, social media, the networks of our students, and word of mouth to recruit participants, often focusing on demographics we felt were missing from the sample, like tradespeople. We then asked participants to pass on the details of the study to anybody they knew who might be interested, and thereby achieved further recruitment through snowballing.

5. Results

5.1. Determinants of life satisfaction

We begin by briefly noting that the determinants of life satisfaction that were reported by our participants correspond to determinants covered by existing theories of wellbeing. Many of these theories are present, and our data thus support a hybrid of integrative account of what subjective wellbeing is (see Fabian 2022 or DeYoung & Tiberius 2022 for examples of such theories). For the most part, we found people's responses and explanations to be rational, coherent, and consistent. People with relatively better (worse) lives gave relatively higher (lower) numbers. This is in line with the extensive psychometric validation of life satisfaction scales. People's ranking of possible lives also seemed broadly stable over time. A longer discussion of the determinants of life satisfaction we identified is the subject of a separate paper. Here we focus on the reporting function.

5.2. Properties of the reporting function

Confabulation

There was some indication of respondents giving post-hoc rationalisations for their scale responses, or even confabulating. For example, N7 said:

Health, I suppose, is another one. Oh, yeah. You didn't say that initially, but it is. But now I'm sort of trying to see what's what went quickly through my mind, because I don't got health, then the other factors would have probably been influenced. So it's something that I probably subconsciously took for granted. But had it been different, it probably would have been up in the foremost factor.

This quote shows the ranking problem in action – health is consistently important, but whether it is given as a reason for a life satisfaction judgement depends on its salience at the time of responding. This is reminiscent of focussing illusions (Kahneman et al. 2006), but that strikes us as too pejorative a lens. Human attention and rationality are bounded. Respondents can only anchor on a few items when considering what matters in their life and how much. On the one hand, this raises some concerns about the comparability of responses over time if different items are salient over even short intervals. On the other hand, the quote from N7 suggests that health has a consistent place in their ranking at all times, and we generally find that respondents have time invariant solutions to the ranking problem.

Scope

Most respondents focused only on their own circumstances when making their life satisfaction judgements (note that this is a UK sample). Some respondents thought about their romantic partners and immediate family, but the circle of care never extended beyond that. For example, M21 said:

So I think I was the first few things that immediately came to mind, so physical health and the kind of, the immediate kind of good and bad of that. It's helpful to explain each one as I sort of go through or. Yeah. So I mean generally kind of I sort of thought about sort of feeling generally fit and active generally well, albeit with just one or two sort of individual things kind of, umm, kind of going on at the minute. And so those kind of generally good news there for a little bit. And then I kind of thought about, just things around me, so kind of home life, my partner and they are kind of generally kind of very good, albeit kind of just the challenges of life are going to come with that. But yeah, so I think immediately physical health and then emotional health and the immediate people around me, I think.

In contrast to this uniformity, respondents used a wide variety of time frames as shown in table 1, anywhere from “right now” to “how my whole life has gone”. The later was especially common among the elderly, whereas younger respondents tended to consider how life is going. It is questionable whether this variety of scopes affects the comparability of responses, but one result here is concerning: when asked directly: “would you change your answer if you considered just life right now?” some elderly respondents said that they would give lower numbers (M8, for example), typically because their health was worsening, or their friends were dying. This complicates the story of the u-shaped relationship between age and life satisfaction, one of the most prominent results in life satisfaction research (Blanchflower & Graham 2021). If

respondents are more likely to interpret scale questions as referring to ‘all of life’ as they age, then the u shape might reflect a changing reporting style more than changing evaluations of a consistent construct.

Table 1: Time frames used in determining scope of life evaluation

TIME FRAME	PERCENTAGE OF SAMPLE
Life right now/currently	38
The past few days	3
The past few weeks	6
The past few months	10
The past year	9
The past few years	17
The past few decades	3
Whole of life	12

Emotions and affective responding

We were somewhat surprised to find that respondent’s life satisfaction judgements were not much affected by emotion or recent events. This was both explicit in their answers to the question “are you feeling any strong emotions right now or has anything happened in recent days that might have affected your answer”, and implicit in the sorts of reasons they gave for their life satisfaction. Two respondents did delay their interviews and one pulled out of the study entirely due to sudden, emotionally destabilising shocks: in all cases a close family member being hospitalised. Perhaps if we had interviewed them at exactly that moment emotions would have been more salient. However, when interviewed less than a week after these incidents these respondents said that they weren’t influencing their life satisfaction, as that judgement was based on a longer-term view.

However, while emotions did not seem to exert a strong effect, several respondents engaged in ‘affective responding’, by which we mean that their scale response was determined by checking in with their affect. When asked what they thought about when coming up with their number, such respondents would say things like (N36, 9/10) “yeah things are good; financially I’m struggling a bit like everyone but I enjoy what I do, I’m in a happy relationship, things are good”. M14 (8/10) is another example:

I measured, I guess, instantaneously, what feels good, what feels left to be desired. But then kind of measured over everything? What comes in within my sense of agency, what's you know, what's within my grasp? And I guess just my life experience or wisdom of I guess you know, internal locus of control, accepting what you what you can and cannot change and having a deep fundamental faith in myself that anything that any gaps in satisfaction can be, you know, I have all the skills I have all the experience or whatever, I have all the wisdom to fill that. So leaving an element of two points, say of these still things that I would want to aim or achieve, but all things considered, yeah, actually, you know, my life's pretty damn good.

Note that both N36 and M14 can express their affective assessment in reasons and language, but the primacy of affect over cognition and thinking is clear in words like “feels good”, “things are good”, “happy relationship”, “sense of agency”. Compare this to more *cognitive* responses like M19 (5/10 – but perhaps surprisingly low given objective indicators):

So, I was thinking about the most common perhaps denominators, so my job role and particularly pay so I, you know, I know that I get paid higher than the average salary for example and, part of that makes me happy, but I think part of that also makes me feel that, I should be happy. And so I was thinking about pay in particular. I was thinking about largely physical circumstances. So the reason it wasn't higher is because you know, I currently don't live in my own property with my partner, I was renting previously with my partner or temporarily moved back in with my parents. Don't particularly like being the idea of a 30 year old man living at home. And so that that knocked a couple of runs off it, and then I suppose the other outside of the physical is the, you know, the social and emotional side, so. Lots of things that we all want to achieve in those domains, and perhaps they take longer than anticipated. So one example is that I often work lots more at the expense of some of the things that I find socially interested in, such as, you know, my hobbies, like hiking. So not having enough time to pursue my hobbies.

The answer here is much longer, more particular, reasoned, and relativising. It displays a greater degree of thinking rather than feeling. Now given the speed with which M19 and other cognitive respondents gave a numerical answer to the initial scale question, affect may be involved in their answer, but so may other heuristics. Affect itself may be based on cognitions done at one time or another that are operating in the background and inform affective life satisfaction.

Ceiling effects: 8/10 is the top of the scale for many people

Alongside being a miserable millionaire, N21 was unusual in another way: he was one of the few respondents to have a fantastical notion of what 10/10 means, specifically the phrase “best possible life”. He said that: “Up in my little study there are 1800 paperback science fiction books on the shelves...So possible for me is all the worlds that every science fiction author has ever invented.”. His comments about “best possible” lives articulated the difficulties of the scope and ranking problems:

So well what I will do is, I go into a little fantasy, have a look at it because it's quite an interesting exercise, and pick apart the stuff that would be good or bad. So for example the Formula One racing driver fantasy. Not sure I would like that fame. That is where I am now applied to a set of circumstances that are frankly inappropriate.

The vast majority of respondents had more realistic notions of ‘possible’, but they shared N21’s reluctance to say they were 10/10. A common explanation given for this was that life could always be better in some way, or as M9 put it: “Yeah, I don't know. I wouldn't go the whole to perfect. I think I would leave just a little gap in between.”. N30 was blunter: “I honestly wouldn't wanna have a 10 out of 10 satisfied life then I think there would be no fun”. Even some respondents who could not point to any specific factors nonetheless refused to give 10/10: “The only thing that my 8 would be better in would probably be the fact that hmmm, I don't actually think I... if I'm really honest, I think what could make my life better? I don't actually know. I really have no clue.” This seems a clear demonstration of ceiling effects in life satisfaction reporting.

Relatedly, the people who said that they were 10/10 or could at least see themselves being 10/10 had modest (though typically at least middle class) lives in mind. For example, M40, a 60 year old woman living in a pleasant part of the countryside, said:

I think I've got to the point... where I don't actually want anymore. This is fine. This is lucky. This is so good and some good people around me. Some nice places to be, a good job. The, you know, the right amount of money. I'm not rich, but a nice, you know, comfortably off. And. And I suppose that my age I would be thinking of retirement but retirement. Feels like a negative thing to me, rather than a positive. Yes, it will happen one day and I may work fewer hours, but the social life and the interactions that I have around me are quite interesting and sufficient, and I don't particularly want to change anything.

The themes here – security, comfort, nourishing relationships, and enjoyable activities – are the same ones we see in respondents who say that they are 8/10. Indeed, based on our affective interactions with respondents in the interviews, we believe that these people judge their lives the same as the people who give 10/10, they are simply stricter graders. The 10s are not more satisfied with their lives, they have a different reporting style.

The scientific and policy implications of this are substantial, and so we urge further study at scale using a range of methods. According to the Office for National Statistics (ONS 2019), in the year ending March 2019 (immediately prior to the COVID-19 pandemic), average life satisfaction in the UK was 7.5/10. If 8/10 is often the top of the scale, the UK was already close to maxed out. Using life satisfaction as a measure of ‘social progress’ may be setting ourselves up for failure. More perniciously, our results imply that the estimated treatment effects for interventions targeted at samples with high baseline levels of life satisfaction are likely to be biased downward. This is because respondents have no space in their scales to communicate improvements, and such changes are likely to result in scale norming rather than higher response categories being chosen. The corollary is that interventions targeted at low baseline life satisfaction groups may appear relatively effective simply because it is easier for respondents to communicate an improvement. This should be factored in when constructing treatment and control groups for life satisfaction experiments, or when using life satisfaction data for cost-benefit analysis.

Our results have mixed implications for the Easterlin paradox. That people do not like to give $r > 8/10$ suggests that income growth may increase life satisfaction among high earners but also induces ceiling effects and scale norming. However, we caution against this interpretation because another consistent theme of our data is that income is a constraint on rather than cause of life satisfaction. Financial means only acted strongly on wellbeing when money was so tight that respondents couldn't pursue their modest interests; as M16 put it: “I don't need to be a millionaire living in a mansion to be at 10. You know, jetting off to San Trope or Monaco, or wherever rich people go. That's not on my kind of list of desires.” What she did want to do was attend occasional workshops on mental health and exercise practices, and she felt she couldn't afford this as a single mother on a lower-middle class salary. Income *growth* may well have little effect once a life well-suited to an individual is financially available. This would be consistent with the result that life satisfaction and experienced wellbeing have a linear relationship with the *logarithm* of income (Stevenson & Wolfers 2013, Killingsworth 2021). Doubling income above middle class levels is very hard and produces only a marginal gain in subjective wellbeing.

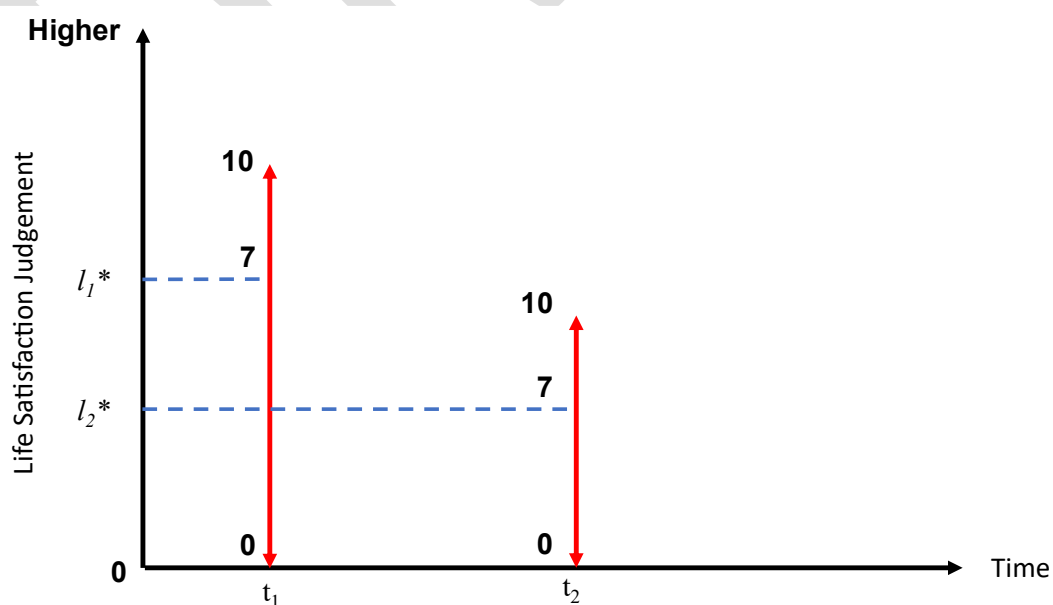
Scale norming

Along with ceiling effects, we find evidence for meaningful scale norming after major shocks. However, this evidence is limited because major shocks are rare. While a modest number of our respondents had such a shock in their life, very few had it recently. Nonetheless, the case that we do have very clearly involves scale norming of a type that should be concerning for science. At interview, M23 was suffering from severe long COVID. She said that she was 7/10 because she had been able to return to work part time and walk her dog for 20 minutes. She thought of 8/10 as working full time and walking her dog up a hill, 9/10 as being able to run again, and 10/10 as being able to run a local half marathon. We asked her what the points on her scale would have meant in 2019 before the COVID pandemic began. She said that she was also a “7 or 8” then, but she was regularly running half marathons and her 9 and 10 out of 10 related to finishing her PhD, which had to be put on hold owing to long COVID. This could be interpreted as adaptation, but this is not how the respondent saw it. When asked directly: “Is your scale from 2019 in any way comparable to your scale today?” her response was “totally different, because I would not have started from a point of extreme illness...I think I was a different person, actually”. M23 was clear that her life today was much less satisfying than pre-COVID, stating that:

My perspective has massively changed with having to go through such bad health and such gruelling recovery and affecting so many different areas of my life, not physically but mentally also academically, socially, in terms of work, so many different aspects. So I think that my perspective has changed completely and I think before 2019 then I think I was much more aspirational.

This cannot be interpreted as adaptation unless you believe the numbers more than the subject. M23’s 7 today and her 7 in 2019 do not correspond to commensurate magnitudes of satisfaction. Her scale has shrunk or shifted, as depicted in figure 6.

Figure 6: scale norming after a major negative shock



Notes: The respondent reports 7/10 satisfaction at both t_1 and t_2 , but the shock between these periods that caused her life evaluation to decline has also caused her to change the meaning of the points on her scales. As such, these 7/10s correspond to different levels of satisfaction – l_1^ and l_2^* .*

This sort of scale norming obviously introduces bias into longitudinal statistical analysis of life satisfaction change that assumes fixed scales over time. Far from resolving this issue, panel data makes it more likely to occur the longer the panel. These are not fixed effects associated with differing reporting styles across respondents, but rather an example of *changing reporting style over time* within a respondent. We also have cases where the meaning of the points on somebody's scale changes over time but the magnitude of satisfaction that each number corresponds to remains constant. Indeed, this is quite common as people age and their priorities change. For example, M25 said:

I think the answer for me is probably about an 8 or a 9 and to really stream of consciousness, one of the things that drew me to this research... was that I recently been thinking that I have quite a high level of life satisfaction, but I don't in any way have the life I thought I would have wanted when I was younger.

Unfortunately, it is impossible to tell these phenomena apart in raw data, which undermines the usefulness of life satisfaction scales for studying precisely the sorts of wellbeing interventions we are most interested in.

Adaptation

We also find evidence of adaptation. It tends to take the form of 'doing the work' – discussed in 5.1 – or 'making peace with it' rather than sensory acclimatisation. An example comes from M15:

"Or, I've got past the bulk of my grief relating to not being able to have a child which I'm going to know I'm going to be fine with I am this is going to give me either a miracle or closure... And if motherhood looks to me like being an emergency foster mum, instead of a biological mum then I haven't got a lot of choice but to accept it. It doesn't mean that grief will leave me entirely it just means that life's gonna just look a bit different and it means that I will take a few more risks with my career."

There were instances of things that people did not seem to adapt to, notably disability (see N15):

Well, it's interesting that because I was thinking about where my place is in society, as a person who's 46. And where I thought I might have been at this age, but due to different health conditions, disability, I feel I've not been able to get to maybe, where I probably think of as a more fulfilling life, I suppose. And, yeah, so when you when you asked me the question, I automatically thought "how do I fit into society now?". And I think you fall into that comparison trap a little bit sometimes as well. You think about yourself and where you might have been compared to someone else with a similar age. And I think it's probably more that when you have health conditions, you probably are not where you want to be due to some of the barriers that you actually face if your health condition or that, you know, the, the you start looking at the media and other things and you sort of place yourself in a bit of a situation where you think,

well, actually, the people of my age may have been, I don't know, living in a different part of the world, have more of an income and maybe have had a bigger family.

Combined with the scale norming evidence above, we think our data supports the view that the negative effects of disability on life satisfaction, especially major disabilities like spinal cord injury, are under-estimated, even in Lucas' (2007) study.

Reference point shifts

Finally, we also found evidence for reference point shifts. People describe cognitively reframing their life and how they assess it such that they then feel better about their lives. For example, M15: "Because historically all my self-esteem came from my career; and it definitely doesn't anymore because I realized that was a very unhealthy way to look at myself and my life." The most common such reframing is some form of gratitude, which is always associated with improvements in life satisfaction judgements. Take M30 for example:

I have got a chronic illness and I'd rather not have it, but that doesn't mean to say that that's pulled me right down. Do you know what I mean? Because I think whatever you're dealt with, there's positives on it, isn't there? And my sort of mantra as I've got, as I've learned as I've gone through life, is that there's always people worse off.

We did not come across anyone who had attempted positive psychology interventions of this sort, whether under instruction or out of their own intuition, who did not see improvements.

Neutral points and the bottom of the scale

Respondents rarely had a clear description of 0/10, though a couple of respondents said they had been at that level at times in their life. Some themes that emerged were suicidality, chronic and painful health conditions, extreme loneliness and poverty, and hopelessness. We did not find any data to support a clear neutral point level where respondents were ambivalent between living or not.

Is the reporting function real?

We found evidence to support a small subset of respondents anchoring on the bottom, middle, and top of the scale, as suggested by Barrington-Leigh (2024). N5, for example: "I know what you're saying. I didn't think about it in terms of that at all. I just thought about it as a top, middle and bottom.", or M14:

Five, that might be a little bit easier. Because you know, then there's a middle point, I'm not sure. Then there's that, you know, absolute worst or not quite. And then it's because it's subjective. It just so many divisions. It's like, Whoa, is it a 5, is it a 6, 7?

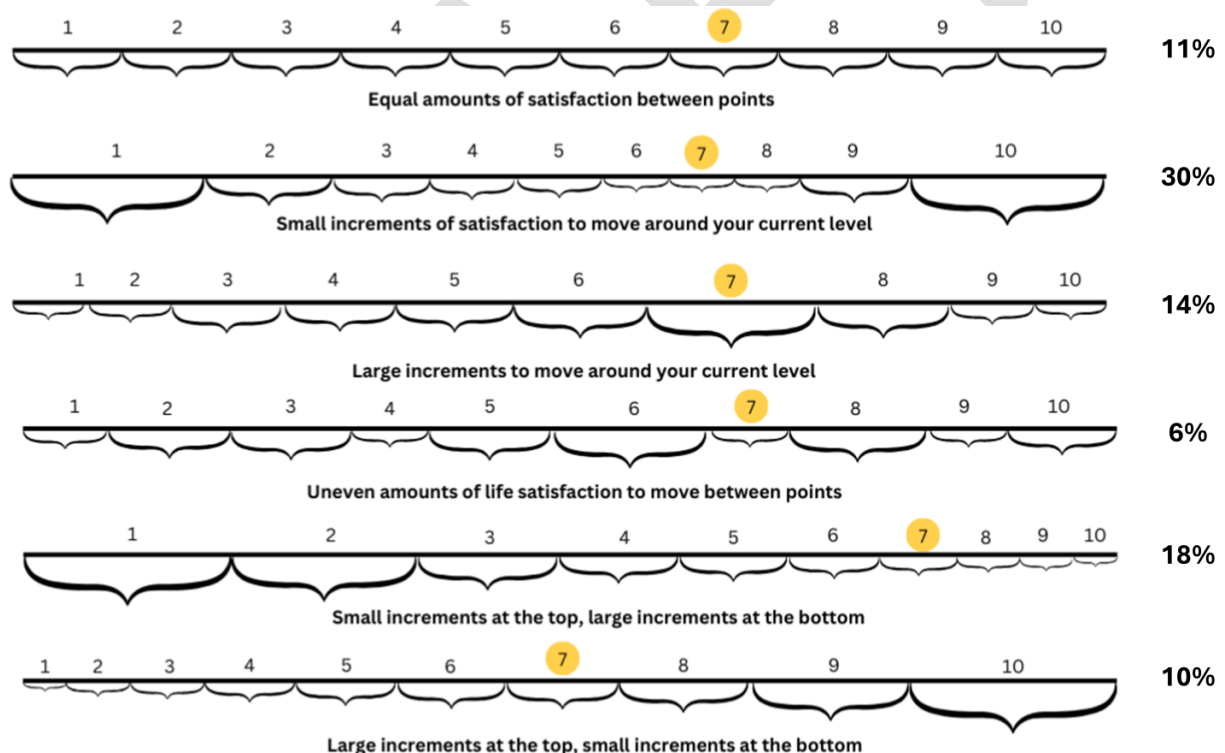
These sorts of responses raise the prospect that our model of life satisfaction judgement and reporting in section 2 is over-structuring the phenomenon. On the one hand, there is certainly some evidence for this, such as the following from M61: "I just think that 7, maybe it's a sort of nice round number". We certainly come across very few respondents who have distinct lives in

mind for *all* response categories – they tend to anchor on a few issues in scope and map these to only a few response categories. However, we do see evidence across our data, some of which is reported above and below, that people do engage in solving the scope, ranking, and calibration problems; that ceiling effects, scale norming, adaptation, and reference point shift exist; and that life satisfaction judgement is often a separate process from mapping that judgement to a response category.

5.3. The shape of the reporting function

By the time we show respondents figure 5 and explain its meaning, they have discussed what various numbers on their scale mean, what it might take to move them up or down a point, and what sort of life they would need to be living to give 0 or 10. Perhaps as a result, none of them had difficulty understanding the figure. However, 11% said that it did not reflect how they used their scales (or they were not asked because we judged this to be the case). For most of these individuals, the response process was relatively affective, and they did not have distinct lives in mind for most of the numbers on their scale. Across the other 89% of respondents, all varieties of reporting function were present, as shown in figure 7.

Figure 7: Self-reported reporting functions



Notes: Self-reported reporting functions. Note that 11% of respondents were either not asked the question or said that these graphics did not reflect how they answered the scale question.

Respondents often found this question difficult, or at least found justifying their answers difficult. Sometimes their answers did not quite make sense. Those that struggled often expressed frustration with the idea of putting their whole life into a single number. However, in most cases, respondents answered quickly and vehemently and provided sound reasons for

their answers. For example, starting from the bottom, an example of a concave reporting style is respondent M44:

So it's a sort of asymptotic and of so you get closer and closer and closer. And so I think the last steps are the hardest. You know ... I don't really see why there's a need to move on up anyway, you know, I mean obviously 8 is pretty good. So I think you know, so not unhappy with that. So if some of the things that I'll do lead to me being a nine fine, but if they're not, I wouldn't be devastated.

An example of a convex reporting function is respondent M39:

No, I think to go up, it wouldn't have to be a huge amount of change, really. Because the more we're talking, the more I'm thinking you know I maybe I am an 8, actually, you know, so even just even just openly discussing with you, and thinking about the past and thinking about how I would grade it, because no one's ever asked me before in my life, how happy I am on a scale of one to 10 in that way, so I've never really thought about it. So I don't think a huge amount would need to change to go up to an eight or nine or a 10...but for a permanent change to a four, I think a fair amount of things would have to, like I said, sort of in conjunction, like if I lost my job, and I lost my house, and I lost my husband, you know, and all my friends...I think one thing after another would just get too much. So I think, you know, to go up, wouldn't be very much, but to go quite drastically down, I think would be of a fair amount of things would change would need to change

Respondents with uneven amounts of life satisfaction to move between points typically had quite chaotic reasoning that makes sense in the context of our conversations, but in ways that are hard to reproduce in short quotes. Consider N30:

It would be a combination of different scales to be the ideal for me to scale my satisfaction...Probably it's, when we talk about satisfaction, it's a whole thing. It's a major part of our life. Like it includes like big factors of a person, how that person is living and who that person is living around. So it can't be, I'm pretty sure it can't be scaled on like an equal amount. It is always gonna be different; there are gonna be different increments. And that's why I think I would say, ohh, large increments at the top, and small increments at the bottom like that, a combination the last three would be ideal for me.

For wide central response categories, consider N23:

I guess, looking at these, what feels like my scale is a bunching up at 1, 2 and 3 and 8, 9 and 10. And then more spread out in the middle.

For “small increments of satisfaction to move around your current level” – which we call the logistic reporting function, consider M10:

I would say the second one with the very small differences between like 6,7,8 and nine, but a very large difference between one and two. I think that's probably the most apt.

And finally, for linear reporting function, we have M3:

I'd say I'd say I would use the top one of those, because it seems to me more, more sensible, more, more logical...I suppose I suppose I just think of a linear scale, I

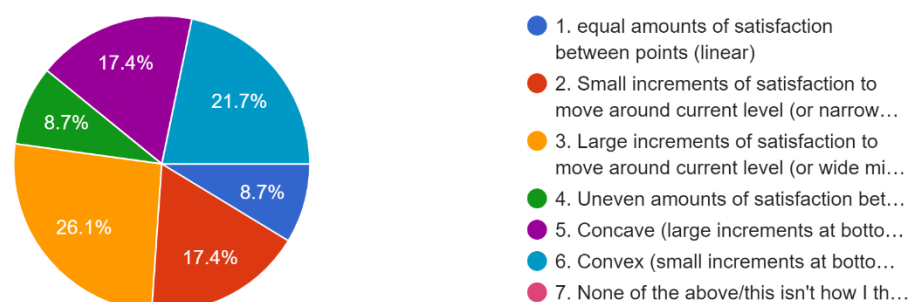
think a linear scale has to be, you know, the points have to be equally spaced. Otherwise, it's otherwise it becomes a bit of a cheat, really. I say, having said that, I rarely use the extremes of top and bottom of any scale, rightly or wrongly. But, you know, to, to sort of, to change the scale as you go along. I'm struggling a little bit to put a numerical value on a very complex thing anyway. And then to say that if you bend the scale that you can, you can make it better or worse. I'm not quite sure I go with that. I think I'd rather just stick to that linear, linear point thing.

In addition to our UK sample, we have conducted this exercise with two small classes of masters-level psychology students at a top German research university within a methods course and obtained similar results. At the start of the class, students are asked their life satisfaction and to describe, in a short paragraph, what sort of life they would need to be living in order to answer each response category. They are then given a half lecture on the conceptual framework outlined in section 1 and 2 of this paper, before being asked the shape of their reporting function question shown in figure 5. Their answers (N=23) are shown in figure 8 and mirror the variety of our UK sample. It seems that even among people adept at psychological science and knowledgeable about life satisfaction reporting, linearity is rare.

Figure 8: *Reporting functions of MSc in Psychology students*

Which of the following scales best describes the way you use your scale?

23 responses



Given our results, if we think we should take people's subjective reports seriously, then we should acknowledge that most people do not have a linear reporting function.

6. Discussion

Our results contain something to support the view of every scholar with a stake in life satisfaction research. Life satisfaction can be broadly explained by existing theories of wellbeing, though it requires pieces from many of them. Responses to life satisfaction questions combine cognitive and affective elements, seem broadly rational, and solutions to the ranking problem seem mostly stable over time. People doing relatively well tend to give higher numbers and vice versa. All this supports the existing psychometric validation of life satisfaction scales as a way of measuring subjective wellbeing, defined as people's own evaluations of how their lives are going.

However, we also found clear evidence of nonlinearity and heterogeneity in reporting functions, ceiling effects, scale norming, and reference point shifts, all of which suggest that status quo methods for estimating models with life satisfaction data are untenably biased. Naïve longitudinal interpretations of such data, where researchers assume that say, 8/10, means the same thing over time, also seem questionable. The implications of this for the adaptation hypothesis and longitudinal phenomena like the Easterlin paradox are substantial. Our results also suggest that using life satisfaction as the sole means of evaluating treatment effects on subjective wellbeing problematic. While it would be ideal to have a single-item indicator of subjective wellbeing for some applications like cost-benefit analysis, our results suggest that multiple items and qualitative process tracing methods are necessary to cross-check life satisfaction results and triangulate the effect of an intervention or shock on life satisfaction. Such mixed approaches, exemplified by the work of Graham (2017), Huppert and So (2013), and Martela et al. (2022), seem somewhat robust to the challenges our results present, though they do not provide precise effect sizes or magnitudes for straightforward cost-benefit analysis.

Something like the reporting function as a process of developing life evaluations and mapping them to scale response categories appears to exist and, while rough, the conceptual framework presented in section 2 seems a feasible foundation for building an understanding of it. Our data show clear cases of respondents solving the scope, ranking, and calibration problems, and doing so heterogeneously. We think replications of this project across cultures would be a valuable next step, as would more quantitative and longitudinal methods for unpacking the reporting function.

A final comment we want to make is that many respondents did not like life satisfaction scales as a way of understanding their own wellbeing, nor as a way for others to do so:

“I don’t go about life thinking of it in terms of a scale and where I am on that scale and I don’t think I’d want to.”

“It’s tricky, because having to put a number to it, to me it’s a sort of an alien thing. We can’t measure satisfaction by numbers.”

“Am I making any sense because it’s very difficult getting tangled up in these numbers. I find it very frustrating.”

“It’s not so much confusing but maybe limiting to the sort of boxes you’re in in a way that doesn’t really get to the bottom of it I think.”

“I don’t mean, what’s the point of living? But what’s the point of answering that question? It’s sort of meaningless.”

People don’t mind answering the scale question at the start of the interview, nor do they struggle to do so, but they resent attempts to rationalise their answer or to analyse their life on the basis of that number. They would certainly rather tell you about their life qualitatively. Many describe being frustrated by how difficult it is to communicate the complexity of their life through a single number.

We think this speaks against claims that life satisfaction scales are ‘democratic’ in the context of policymaking (e.g. Clark et al. 2018). Democracy implies voting, or at least voice. If respondents were empowered to communicate their wellbeing directly to policymakers, they would not do so in a single number. Scepticism of respondents’ verbal explanations compared to numerical responses on the grounds of cognitive biases (O’Donnell et al. 2014) do not seem

warranted to us given the rationality of the majority of answers in our data. Nor does such scepticism seem congruent with taking subjective judgements seriously. We suspect that much of the supposed evidence for cognitive biases in life satisfaction is evidence of bias *in the scale instrument* rather than the respondent, as many of our results suggest. Thus to make progress in understanding *subjective* wellbeing, and to make it an objective of policy, we will need to enable individuals to express their wellbeing in natural language and to take such expressions as the fundamental type of data to be analysed and responded to.

An obvious next step in this research is to work with a much larger, representative sample of the UK population and determine how widespread are the phenomena we have identified here. We did not discern anything peculiar about the elderly we interviewed, the people who refused to use the top of their scale, or those who engaged in scale norming, but that does not mean that these concerns will show up at scale. Developing survey questions that can tap these issues quickly and without an interviewer on hand to explain will be challenging but surmountable. It would be ideal to study the reporting function longitudinally, and to work with cross-cultural samples.

7. Conclusion

Our results suggest the need for caution at the frontiers of subjective wellbeing research, especially efforts to integrate life satisfaction data into policy analysis like WELLBYs (Frijters & Krekel 2021). Given the measurement error and biases that result from scale norming and reference point shifts, stylised facts about subjective wellbeing derived from statistical analysis of life satisfaction data, especially longitudinal analysis, need to be considered unsettled. Understanding the reporting function and its effects on scientific analysis will be very challenging and require mixed methods and interdisciplinary perspectives. An attitude of circling the wagons against critiques will be unhelpful in this context. As a field, we need to develop a richer philosophy of science in our approach to understanding subjective wellbeing. The present dominance of regression analysis and experiments needs to be combined with qualitative work, grounded theory, theory-led quantitative analysis, and a variety of measurement strategies.

We encourage all scholars interested in subjective wellbeing, especially those who focus strictly on quantitative methods, to read our data. It is publicly available along with our codebooks, which makes searching it easier. To understand what we are observing in numerical analysis of life satisfaction, it is very helpful to understand the rich qualitative texture of the phenomenon that goes into those numbers. There is a tremendous amount of complexity that people try to cram into a single number.

Works Cited:

- Angelini, V., Cavapozzi, D., Corazzini, L., and Paccagnella, O. (2014). Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases. *Oxford Bulletin of Economics and Statistics*, 76(5), 643–666.
- Barrington-Leigh, C. (2024). The econometrics of happiness: Are we underestimating the returns to education and income? *Journal of Public Economics*, 230(1): e105052.
- Beatty, P. C. & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2): 287–311.
- Benjamin, D., Guzman, J. D., Fleurbaey, M., Heffetz, O., & Kimball, M. (2023). What do happiness data mean? Theory and survey evidence. *Journal of the European Economic Association*, 21(6): 2377–2412.
- Blanchflower, D. & Graham, C. (2021). The u shape of happiness: A response. *Perspectives on Psychological Science*, 16(6): 1435–1446.
- Bond, T. N. & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, 127(4): 1629–1640.
- Brickman, P., Coates, D. & Janoff-Bulman, R. (1978). Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology*, 36(8): 917–927.
- Chan, B. C. F., Craven, C. & Furlan, J. C. (2018). A scoping review on health economics in neurosurgery for acute spine trauma. *Neurosurgical Focus*, 44(5): e15.
- Clark, A., Diener, E., Georgellis, Y. & Lucas, R. E. (2008). Lags and leads in life satisfaction: A test of the baseline hypothesis. *The Economic Journal*, 118(529): F222–F243.
- Clark, A., Fleche, S., Layard, R., Powdthavee, N., & Ward, G. (2018). *The origins of happiness: The science of well-being over the life course*. Princeton University Press.
- Craig, A., Tran, Y. & Middleton, J. (2008). Psychological morbidity and spinal cord injury: A systematic review. *Spinal Cord*, 47(1): 108–114.
- Deaton, A. & Stone, A. (2016). Understanding context effects for a measure of life evaluation: How responses matter. *Oxford Economic Papers*, 68(4): 861–870.
- DeYoung, C. & Tiberius, V. (2022). Value fulfilment from a cybernetic perspective: A new psychological theory of well-being. *Personality and Social Psychology Review*, 27(1): 3–27.
- Diener, E. & Diener, C. (1996). Most people are happy. *Psychological Science*, 7(3): 181–185.
- Diener, E., Emmons, R. A., Larsen, R. J. & Griffin, S. (1985). Satisfaction with life scale. *Journal of Personality Assessment*, 49(1): 71–75.
- Diener, E., Inglehart, R. & Tay, L. (2013). Theory and validity of life satisfaction scales. *Social Indicators Research*, 112(3): 497–527.
- Emerson, S. D., Guhn, M. & Gadermann, A. M. (2017). Measurement invariance of the satisfaction with life scale: Reviewing three decades of research. *Quality of Life Research*, 26(9): 2251–2264.

- Fabian, M. (2021). Scale norming makes welfare analysis using life satisfaction data difficult. *Journal of Happiness Studies*, 23(4): 1509–1541.
- Fabian, M. (2022). *A theory of subjective wellbeing*. Oxford University Press.
- Ferrer-i-Carbonell, A. & Frijters, P. (2004). How important is methodology for the estimates of the determinants of happiness? *The Economic Journal*, 114(497): 641–659.
- Fleurbaey, M. & Blanchet, D. (2013). *Beyond GDP: Measuring welfare and assessing sustainability*. Oxford University Press.
- Frijters, P. & Krekel, C. (2021). *A handbook for wellbeing policy-making: History, theory, measurement, implementation, and examples*. Oxford University Press.
- Graham, C. (2017). *Happiness for all? Unequal hopes and lives in pursuit of the American dream*. Princeton University Press.
- Harris, E.C. & Barraclough, B. M. (1994). Suicide as an outcome for medical disorders. *Medicine*, 73(6): 281–296.
- Huppert, F. & So, T. C. (2013). Flourishing across Europe: Application of a new conceptual framework for defining well-being. *Social Indicators Research*, 110(3): 837–861.
- Kahneman, D., Krueger, A. B., Schkade, D., Schwartz, N. & Stone, A. (2006). Would you be happier if you were richer? A focusing illusion. *Science*, 312(5782): 1908–1910.
- Kaiser, C. (2022). Using memories to assess the intrapersonal comparability of wellbeing reports. *Journal of Economic Behavior and Organisation*, 193(1): 410–442.
- Kaiser, C. & Vendrik, M. (2023). How much can we learn from happiness data? *Preprint* <https://osf.io/preprints/socarxiv/gzt7a>.
- Kapteyn, A., Smith, J. P. & Van Soest, A. (2010). Life satisfaction. In *International Differences in Wellbeing*, edited by E. Diener, J. F. Helliwell & D. Kahneman. Oxford University Press.
- Killingsworth, M. A. (2021). Experienced well-being rises with income, even above \$75000 per year. *Science*, 118(4): e2016976118.
- Larroulet Philippi (2022). On measurement scales: Neither ordinal nor interval? Online first in *Philosophy of Science*.
- Liddell, T. M. & Kruschke, J. K. (2018). Analysing ordinal data with metric models: What could go wrong? *Journal of Experimental Social Psychology*, 79(1): 328–348.
- Lucas, R. (2007). Long-term disability is associated with lasting changes in subjective well-being: Evidence from two nationally representative longitudinal studies. *Journal of Personality and Social Psychology*, 92(4): 717–730.
- Martela, F., Lehmus-Sun, A., Parker, P. D., Pessi, A. B. & Ryan, R. (2022). Needs and well-being across Europe: Basic psychological needs are closely connected with wellbeing, meaning, and symptoms of depression in 27 European countries. *Social Psychological and Personality Science*, 14(5): 501–514.
- Montgomery, M. (2022). Reversing the gender gap in happiness. *Journal of Economic Behavior and Organisation*, 196(1): 65–78.

O'Donnell, G., Deaton, A., Durand, M., Halpern, D. & Layard, R. (2014). *Wellbeing and policy*. Legatum Institute. https://www.frontier-economics.com/media/l2havlsh/20151012_wellbeing-and-policy_frontier.pdf

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

Oswald, A. & Wu, S. (2009). Objective confirmation of subjective measures of human well-being: Evidence from the USA. *Science*, 327(5965): 576–579.

ONS (2019). Personal well-being in the UK: April 2018 to March 2019. UK Office for National Statistics. <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/april2018tomarch2019>

Sawatzky, R., Sajobi, T. T., Russell, L., Awosoga, O. A., Ademola, A., Bohnke, J. R. et al. (2024). Response shift results of quantitative research using patient-reported outcome measures: A descriptive systematic review. *Quality of Life Research*, 33(1): 293–315.

Schroder, C. & Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, 92(1): 337–358.

Schutte, L., Negri, L., Delle Fave, A. & Wissing, M. P. (2021). Rasch analysis of the satisfaction with life scale across countries: Findings from South Africa and Italy. *Current Psychology*, 40(10): 4908–4917.

Stevenson, B. & Wolfers, J. (2013). Subjective well-being and income: Is there any evidence of satiation? *American Economic Review*, 103(3): 598–604.

Appendix A0

Ordered probit models make the following assumptions³:

1. Latent satisfaction s_i is given by a linear equation of the form:

$$s_i = \beta x_i + \epsilon_i,$$

Where $\epsilon_i \sim N(0,1)$.

2. Observed responses r_i are generated by a rule of the form:

$$r_i = k \leftrightarrow \tau_{k-1} < s_i \leq \tau_k$$

With $\tau_0 = -\infty$ and $\tau_K = \infty$.

This yields a log-likelihood function of the form:

$$\mathcal{L} = \sum_i \sum_k 1(r_i = k) \ln[\Phi(\tau_k - \beta x_i) - \Phi(\tau_{k-1} - \beta x_i)]$$

We maximise $\mathcal{L}$ with respect to $\tau = \{\tau_1, \dots, \tau_k, \dots, \tau_{K-1}\}$ and β to obtain estimates.

For ease of exposition, suppose we estimated an ordered probit regression without any covariates. Doing so will only yield estimates of the response thresholds $\tau_0, \dots, \tau_k, \dots, \tau_{K-1}$. In this case, maximising $\mathcal{L}$ will yield response thresholds given by $\tau_k = \Phi^{-1}(s_k)$, where $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal cumulative distribution function and s_k denotes the cumulative response share up to and including the k^{th} response category.

Now suppose that the distribution of reported satisfaction is approximately normal. In that case, we have $s_k = \Phi(\frac{k}{K})$. Substituting this into the above expression for τ_k yields $\tau_k = \Phi^{-1}(s_k) = \Phi^{-1}\left(\Phi\left(\frac{k}{K}\right)\right) = \frac{k}{K}$. Hence, $\tau_k - \tau_{k-1} = \frac{k}{K} - \frac{k-1}{K} = \frac{1}{K}$ for every response category k . Therefore, in this case, all estimated response thresholds are equidistant from each other, implying linearity of scale use.

Now, in many cases, the distribution of reported satisfaction is not approximately normal in this way. Instead, as shown in the distribution of reported satisfaction from the 20xx UK Annual Population survey shown in panel (A) of Figure Axxx, responses tend to be left-skewed. Using these response distributions, we obtain thresholds shown in panel (B) of Figure Axxx. As expected, the size of differences in estimated thresholds tends to decline as we move towards higher response options.

Empirically, as shown in Table Axxx, adding covariates to the ordered probit model, has only a minimal effect on the location of the estimated response thresholds. This is likely driven by the fact that observed covariates tend to explain very little variance in satisfaction, implying that the conditional and the unconditional distribution of

³ An equivalent alternative set of assumptions is to allow for the variance σ of s_i to be estimated freely, while constraining two thresholds, e.g. τ_1 and τ_2 to some particular value. Given that these formulations are equivalent, the argument below would also in this case, but would require more notation. For a related argument to the one set out in this appendix, and further extensions of it, see Kaiser & Vendrik (2023).

satisfaction tend to be similar. Indeed, this fact was previously exploited by van Praag & Ferrer-I-Carbonell (2008) when proposing their ‘Probit Adapted Ordinary Least Squares’ estimator, which first estimates thresholds without covariates in a first step (as given above), then transforms the observed response data by these estimated thresholds, and finally estimates an OLS regression of the transformed response data. As also shown in Table Axxx, the results from OLS, ordered probit, and POLS regressions on a standard set of covariates are extremely similar to each other.

However, it is worth stressing again that even if these results were not similar to each other, it would still be the case that ordered probit models, with their assumption of a particular conditional distribution for latent satisfaction *imply* an untested assumption about respondents’ scale use. In this paper, we provide qualitative data on the validity of this assumption.

Works Cited:

van Praag, B. & Ferrer-i-Carbonell, A. (2008). *Happiness quantified: A satisfaction calculus approach*. Oxford University Press.

Appendix A1: interview Questions

Note: These are the instructions exactly as given to interviewers.

START THE RECORDING & confirm willingness to be recorded.

PREABLE: *I'm going to ask you a series of questions pertaining to life satisfaction. Please **think out loud** when answering these questions. There are no right or wrong answers. We are simply interested in what goes on emotionally, mentally, and in terms of language when people think about and answer these questions.*

If a question doesn't make sense, please ask for clarification. If a question just feels totally alien to you, like it is not at all how you think, then please say that. We value this information.

The first two-thirds of the interview will be mostly about your life, your satisfaction with it, and what determines that satisfaction. Then we shift more towards how you express this to us, the researchers, on a scale from 1–10.

1. All things considered, how satisfied are you with your life overall on a scale from 0–10?
 - a. Randomise with Cantril's ladder 50% of the time (see figure 1)
2. What did you think about while you were answering that life satisfaction question?

NB: A few quick, almost housekeeping questions...

3. Are you feeling any emotions right now? Do you think they had an effect on your answer?
4. Did you have to put aside your feelings to give an answer you felt was accurate, or did you use your feelings to arrive at your answer?
5. Has anything happened to you over the last few days that might have affected your answer? How much did it affect your answer?
6. What time frame did you use? Were you thinking about right now, or the past week/month/year?.
7. Were you thinking about how your family, friends, or society are doing when you answered the question or were you more focused on your own personal life?
8. Did you think about your life and how it could be better/worse, or did you think about all possible lives and then how your life fits in?

NB: Thank you, OK now some more substantive questions...

9. Did you compare how your life is going right now with how it has been in the past? [Or how you would like your life to be in the future?]
10. Did you compare how your life's going right now with how you think other people are doing?

11. To what extent is your life satisfaction determined by how well you are meeting certain goals or ambitions?

FLAG: Thank you so much for your answers thus far (be especially thankful to people who share raw or vulnerable information for their honesty). The questions will now start shifting more towards how you use the 1–10 scale to express yourself and what each number on that scale means to you. This might be a little bit weird or boring. Sorry about that. Again, if a question really doesn't resonate with you – if it's just not how you think about the life satisfaction scale – then please tell us that. Don't feel like you need to be helpful and give us an answer. Telling us that a question makes no sense is very helpful!

12. How did you interpret the word 'satisfied'? What does that mean to you?
13. Do you think that being satisfied with your life means that your life must be going well?
14. So you said you were an $x/10$, what do some of the other numbers on the scale mean to you?
- a. What sort of life would you need to be living to answer x ? What about y ?
 - b. **NOTE:** randomise the numbers used in this question
 - c. **NOTE:** try and use linear combinations e.g. 3 & 6 or 4 & 8, at least once
15. Did you have a particular life in mind for all the numbers on the scale or just the numbers around the number you answered?
16. When you answered the question, did you think about what sort of life you would need to be living to answer a 10 or 0 are?
17. What do 0 and 10 mean to you? Mid point/dissatisfied?
18. How much more satisfied would you need to be to go up 1 point in your scale? What sort of change in your life would make that happen?
19. Do you think that amount of satisfaction is the same for all points on your scale? Or would you need a really huge jump to go from say, a 9 to a 10 or a huge drop to go below a 4?
20. Staying with this theme, please have a look at the following picture (see figure 2). Do you think this is a reasonable way to represent how you use your scale? Do any of these correspond to the way you personally use your own scale?

NB: whether you ask the next few questions and in what order depends on what the respondent has already said, how much time is left in the interview, and what is most interesting to ask.

21. Do you think the points on your scale have meant roughly the same thing for the past few years or does that meaning change regularly?
 - a. When was the last time the meaning changed?
 - b. If I asked you about your life satisfaction again in a year, do you think you would consider the same issues as you did today?
22. Despite the meaning of the numbers changing, do you think if you said $x/10$ (where x is their answer today) 5 years ago, or 10 years ago, that it would correspond to the same level of *satisfaction* that you feel today?
23. Has there been a time in the past when you also would have said $x/10$? Do you prefer your life today or your life back then?
24. Stretch or shift?
25. After all the discussion we've had, are you thinking about your life differently to the start of the interview? Would you give a different number out of 10 now than at the start?
26. Is there anything you want to say about life satisfaction or these scales that my questions haven't given you a chance to express?

Figure 1: Cantril's ladder of life

Assume that this ladder is a way of picturing your life. The top of the ladder represents the best possible life for you. The bottom rung of the ladder represents the worst possible life for you.

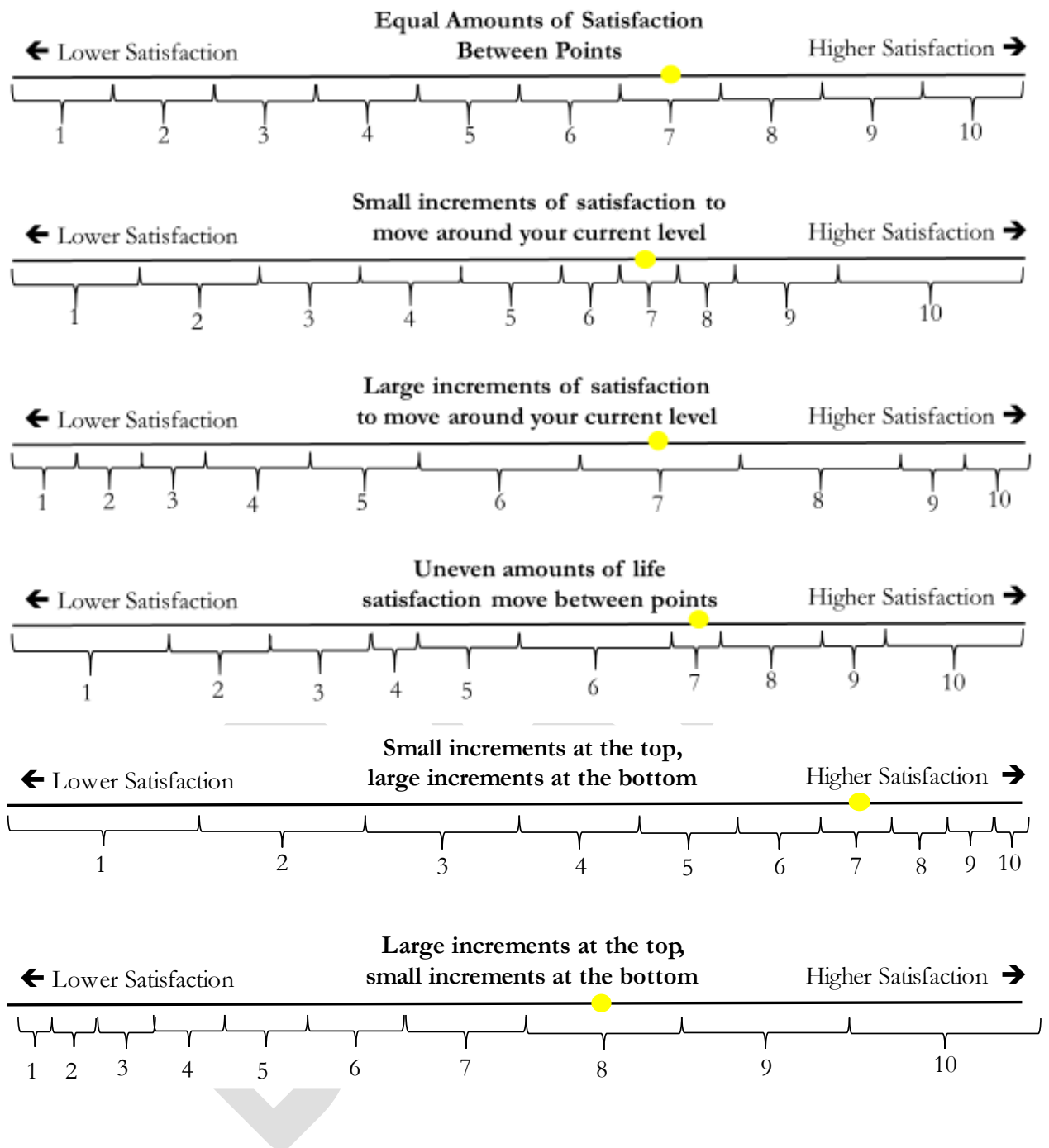
On which run of the ladder, 0–10, with 0 being the lowest and 10 the highest, do you currently stand?



Best possible life

Worst possible life

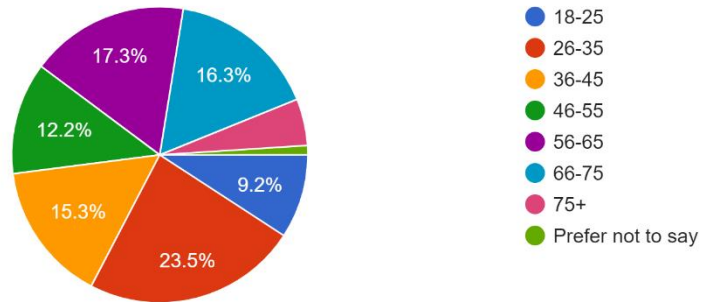
Figure 2: Reporting functions



Appendix A2: Descriptive statistics

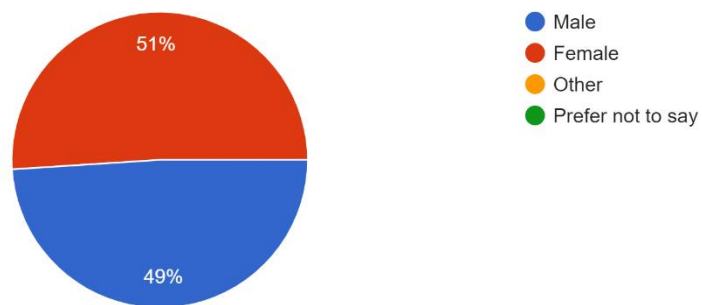
How old are you?

98 responses



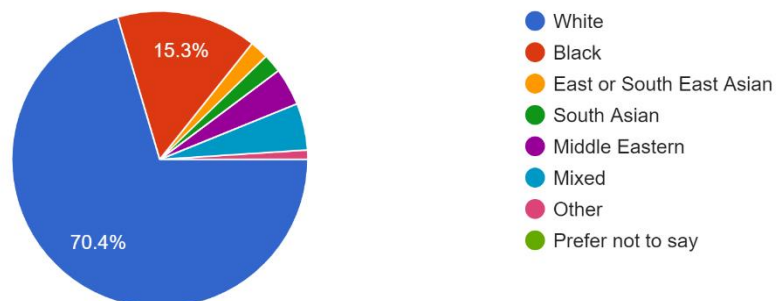
What is your gender?

98 responses



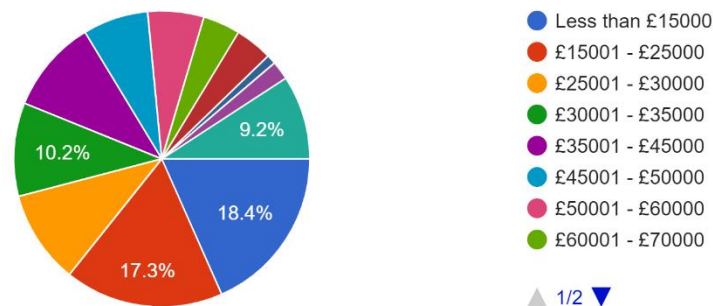
What is your ethnic identity?

98 responses



What is your gross annual income?

98 responses

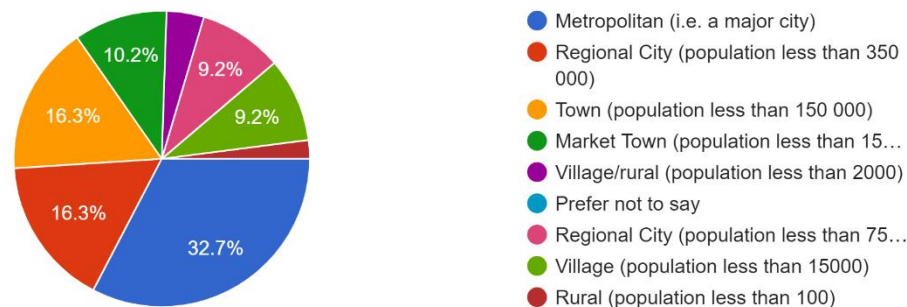


- Less than £15000
- £15001 - £25000
- £25001 - £30000
- £30001 - £35000
- £35001 - £45000
- £45001 - £50000
- £50001 - £60000
- £60001 - £70000

1/2

How would you describe the area in which you live?

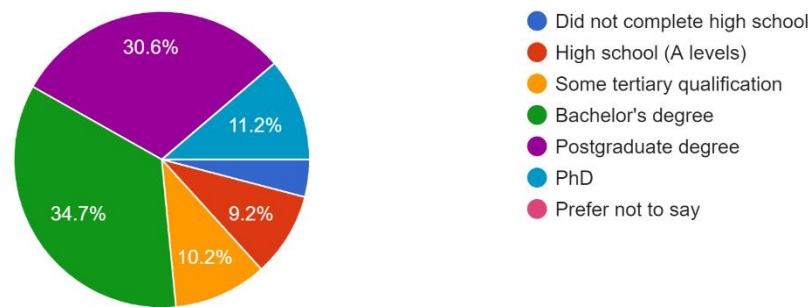
98 responses



- Metropolitan (i.e. a major city)
- Regional City (population less than 350 000)
- Town (population less than 150 000)
- Market Town (population less than 15 000)
- Village/rural (population less than 2000)
- Prefer not to say
- Regional City (population less than 75 000)
- Village (population less than 15000)
- Rural (population less than 100)

What is your highest level of education?

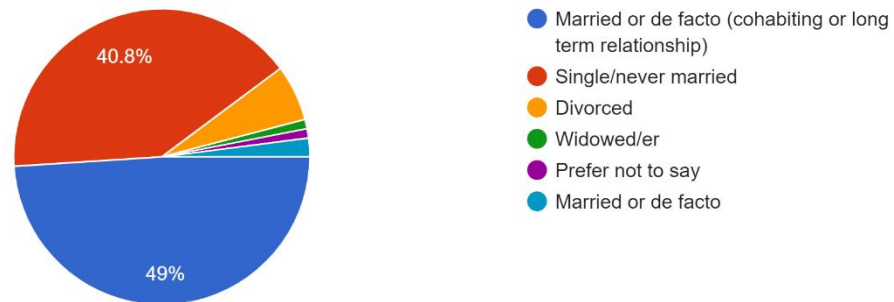
98 responses



- Did not complete high school
- High school (A levels)
- Some tertiary qualification
- Bachelor's degree
- Postgraduate degree
- PhD
- Prefer not to say

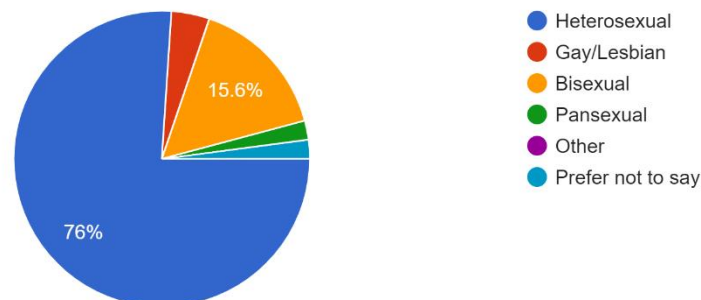
What is your marital status?

98 responses



What is your sexual orientation

96 responses



What is your occupation? (Abridged – duplicates removed)

University Lecturer, Software Developer, Retired academic, retired psychotherapist, Retired Air Traffic Controller, Grants Manager, Legal technical support and Live Events, retired teacher, Student, Retired Solicitor, Actress/Presenter/Foster Carer, Property Manager, Development Specialist, Actress and landlord, Healthcare, Research consultant, Gardener, Dispatch Rider, university research fellow, Education Programme Manager, Teacher, Stay at home Mum, Business Consultant, Designer, Almost retired but still running tiny online enterprise (subscription boxes), Volunteer Citizens Advice, University Administrator, assistant support worker, Tribunal member for General Medical Council, Social worker, Project Co-ordinator (Offshore Wind), Retired Chairman/CEO, Bid writer, Retired probation officer, Therapist, Nurse, Barman, Celebrant/Registrar, Sales Manager, Civil engineer, Dentist, Cleaner, Behavioural Insights Analyst, Consultant, Solicitor, Accountant, DevOps Engineer, Plant Supervisor, Retired Architect, UNEMPLOYED, Librarian, Academic medical practitioner, IT Consultant, Software Engineer, Volunteer coordinator, Development and Fundraising, Commercial Events Coordinator, Office Manager, Unpaid carer to family members, Unemployed/disabled, Delivery Driver, Admin, Self-employed soft furnisher, Painter and decorator