UNIVERSITY OF OXFORD

Wellbeing Research Centre

Working Paper Series

2401

A Happy Probability About Happiness (and Other) Scales: An Exploration and Tentative Defence of the Cardinality Assumption

Michael Plant

February 2024

A Happy Probability About Happiness (and Other) Scales: An Exploration and Tentative Defence of the Cardinality Assumption

0. Abstract

Numerical surveys of feelings, such as “How happy are you, on a scale of 0-10?”, are now ubiquitous and increasingly taken seriously by researchers, governments, companies, and others. The data are often treated as cardinal – that is, a difference between a 2 and a 3 for one person is the same as that of a 5 to a 6 of another. There is long-running scepticism about assuming cardinality; if we have been wrong to assume it, the existing conclusions in the literature may be in doubt. This paper investigates how reasonable it is for researchers to assume scale cardinality. It makes four contributions. First, I observe that cardinality is a matter of degree, so we must ask if plausible deviations from it are big enough to make a difference. Second, I offer a novel argument for why it is rational for respondents to interpret the scales as cardinal if they want to accurately convey their feelings. Third, I argue that uncertainty about how people interpret surveys does not push us away from assuming cardinality; if anything, the opposite is true. Fourth, I conduct what is, as far as I am aware, the first review of the evidence of the conditions underlying cardinality (linearity and comparability); from this, I conclude the deviations, if they exist, are small enough that few, if any, practical conclusions would need to be revised. Hence, it seems reasonable to assume cardinality for now, but further exploration should be done. I close by noting that detours from cardinality can, in theory, be corrected statistically, so worries about how people answer surveys need not prevent us from ever using survey data.

1. Introduction

In research and everyday life, we put numbers on our feelings. We rate, often on a 0-10 scale, our happiness, job satisfaction, relationship satisfaction, the movies we watch, the restaurants we eat at, and so on. Data on these ratings are now routinely collected and analysed not just by academics, but by national governments and companies. We would not do this if we thought these numbers were meaningless.

Yet, there is something paradoxical about *subjective scales* – that is, where we put numbers on feelings. They are simultaneously familiar and unfamiliar. When we ask someone how happy they are, and they say 2/10 rather than 9/10, we believe we understand them, at least roughly, and there is a big difference between the two. Yet, we might then ask, what *exactly* does a 2/10 *mean*? Kaiser and Oswald (2022) claim, “they are ‘made-up’ numbers on a scale that does not exist,” which raises questions about how to understand them. One oddity is that we are trying to put an apparently unbounded phenomenon on an apparently bounded scale: there’s no logical limit to how happy someone could be, so what does ‘10/10’ capture?

For just these sorts of reasons, there are doubts about how to interpret subjective data. Researchers typically treat the scales as *cardinal* – that is, the difference between a 6 and a 7 for any individual is the same as that between a 3 and a 4 for any other individual (Ferrer-i-Carbonell & Frijters, 2004; OECD, 2013: 189-90). Two conditions are individually necessary and jointly sufficient for cardinality: *linearity* and *comparability*. *Linearity* means the difference between each point represents the same change in magnitude: going from a 6 to a 7 is equivalent to going from a 2 to a 3. *Comparability* means there is a consistent meaning across people and times: your 7/10 is the same as my 7/10, etc. If happiness and other feelings scores are cardinal, we can treat them like those for objective measurable properties such as weight, height, and income: add them up, subtract them, take averages, and so on.

There is a longstanding scepticism about both the assumption of cardinality and, consequently, the scientific value of feelings data, particularly in economics (Bond & Lang, 2019; Ferrer-i-Carbonell & Frijters, 2004; Kaiser & Oswald, 2022b; Kristoffersen, 2010; Layard, 2003; Robbins, 1932; Schröder & Yitzhaki, 2017). After all, how confident can we be that people are using the scale in the same way? Is one person's 7/10 the same as another's? How would we even know?

An alternative assumption is that self-reports are *ordinal*: they represent a ranking, not a quantity. This means that, for each person, higher numbers indicate higher levels of feeling, but we don't know how much higher.

A key point of this essay is that an 'ordinal or cardinal?' framing is too narrow. What we should be asking is *how close* the measures are to cardinal; perhaps they are not exactly cardinal, but it makes no practical difference if we treat them as such.

If researchers have incorrectly assumed feelings data are cardinal, it's easy to see that much existing knowledge could be in doubt. For instance, the World Happiness Report ranks different countries by their average life satisfaction, and famously puts the Scandinavian countries at the top: the 2023 winner was Finland at 7.5/10 (Helliwell et al., 2023). It assumes cardinality: 10/10 means the same thing for a Finn as it does an American, and so on. If the numbers were ordinal, we could not aggregate them to conclude which country is most satisfied on average.

A recent prominent example of this scepticism is from Bond and Lang, who argued in *The sad truth about happiness scales* – the article this paper draws its name from – that we should not assume happiness scales are cardinal, and that under different assumptions, various 'canonical' results about happiness would no longer hold (Bond & Lang, 2018). (In §7.1, I explain the flaw in Bond and Lang's thesis.)

Kristoffersen (2011) traces the cardinality debate from the Victorian era to the present day, and notes that despite the latent uncertainty, there has been relatively little explicit discussion of the issue; she calls it "the elephant in the room". In the past few years, there has been a growing trickle of papers on some aspect of the topic, but this literature appears disconnected and incomplete.

I will mention some notable contributions now, although their relevance will be clearer after the clarifying remarks I make in §2. Ng (1995, 1997) argues feelings are cardinal in nature but does not offer arguments that the measures are cardinal. Kristofferson (2011), as noted, outlines the issues and offers a theoretical argue for linearity but does not provide supporting empirical evidence. Kristoffersen (2017) defends a linear interpretation by comparing life satisfaction reports to mental health scores; this uses one subjective scale to assess another and would not convince a sceptic – why trust either? Various authors have argued that 'noise' in measurement is not a concern for cardinality (Bertrand & Mullainathan, 2001; Bronsteen et al., 2012; Dolan & White, 2007); I agree, but explain (§2) we need to worry about another issue, bias. Ferrer-i-Carbonell & Frijters (2004) use different statistical tests which assume subjective data is either cardinal or ordinally comparable and find it makes little difference to the results. This is indicative of comparability in the population they examine, but different groups may attempt different meanings to their scales.[1]

The state of the literature is such that is difficult to get a clear sense of what the problems are, how serious they are, and whether background theory and current evidence better support optimism or pessimism about the cardinality assumption.

My aim in this paper is to investigate whether it is reasonable for researchers to suppose that subjective scales are cardinal. I begin from the position that scepticism of sort is appropriate, and examine the grounds there are to maintain. Ultimately, with some small caveats, I conclude it is reasonable for researcher to interpret subjective scales as cardinal.

This paper makes four contributions. First, I observe that cardinality is a matter of degree, so we must ask if plausible deviations from it are big enough to make a difference to the studied outcomes. Second, I offer a novel argument for why it is rational for respondents to interpret the scales as cardinal if they want to accurately convey their feelings. Third, I argue that uncertainty about how people interpret surveys does not push us away from assuming cardinality; if anything, the opposite seems true. Fourth, I conduct what is, as far as I am aware, the first review of the evidence of the conditions underlying cardinality (linearity and comparability); I conclude that the deviations, if they exist, are small enough that few, if any, practical conclusions would need to be revised. Of these, the second and fourth are the main contributions; the other two are largely elaborations or restatements of points made previously.

Here's how I proceed. §2 makes several preliminary comments to clarify the nature of the problem, including noting that there are three ways to interpret subjective scales – *cardinalism*, *ordinalism*, and *quasi-cardinalism* – which I explain there. §3 offers a novel 'Grice-Schelling' argument for why, if people want to accurately convey their

[1] Further, scale-use could be comparable but non-linear, so Ferrer-i-Carbonell & Frijters's (2004) analysis does not test for linearity.

feelings, it is rational for them to interpret subjective scales as cardinal, and explains how they could do this. This rationality-based account helps form our expectations (and makes subjective scales less mysterious) but it does not prove people genuinely do this.

§4 sets up the case that uncertainty about scale interpretation need not push away from assuming cardinality. §5 argues that the ordinalist interpretation implies an implausibly radical uncertainty about respondents' scale use. This leaves us with two options: cardinalism or quasi-cardinalism. §6 shows there are many different quasi-cardinal views and lays down an argumentative gauntlet: rejecting cardinality implies accepting some alternative belief about people's reporting behaviour. Whichever alterative one believes, there is no problem, in principle, in deriving cardinal data from it (Kristoffersen, 2011; Y. Ng, 1997): if we know how people's scales are 'distorted', we can apply mathematical transformations to 'correct' them.

§7 surveys the available evidence for the three assumptions relevant to cardinality (linearity, intertemporal comparability, interpersonal comparability); as far as I am aware, no such overall review has been attempted. I argue that (1) large deviations from cardinality for any of the three assumptions seem unlikely, (2) we cannot rule out small deviations, but (3) plausible deviations would be small enough that few, if any, of the current conclusions in the happiness literature would be revised.[2] Specifically, the worst-case revision on the available data is that we *might* now conclude that women are happier than men (current research suggests the opposite) and *slightly* alter the international happiness rankings of countries.

Hence, as it would seem to make little practical difference to assume cardinality, I conclude it is reasonable for researchers to treat subjective scales as cardinal, at least unless new evidence or analysis suggests otherwise. My analysis is speculative, theoretical, and uncertain, but I cannot apologise for this, as there is no alternative: necessarily, subjective properties cannot be measured by objective means, so empirically demonstrable certainty is not on offer. §8 makes some concluding remarks.

2. Preliminaries

One worry we might have is that feelings are not quantities: they are not properties that vary in degree. Of course, if feelings are not quantities – there is a lack of what we might *phenomenal cardinality* – it would clearly be mistaken to think we could *measure* them on a cardinal scale. That happiness is a quantity amounts to the claim that individuals can be more or less happy (*mutatis mutandis* for other feelings). Whilst some might doubt this, I do not and I doubt many others do either. For the sake of

[2] Ideally, we would ask whether abandoning cardinality in favour of a plausible alternative account of scale interpretation would change the priorities for governments and/or others. However, so little work has been done to determine what the wellbeing-based priorities are, even assuming cardinality, that such a comparison is not possible. Hence, I do the next best thing. On taking a wellbeing, or 'WELLBY', approach to cost-effectiveness, see Frijters & Krekel (2021).

space, I won't attempt to establish this claim here. For discussion, see Crisp (2006), Ng (1997), Schröder & Yitzhaki (2017).

The main worry seems to be *scale interpretation*: how people report their feelings when given numerical scales (Stone & Krueger, 2018). Perhaps people interpret the scales in very different ways. This raises the question of how people *would* need to interpret subjective scales to yield cardinal data. As stated above, there are two individually necessary and jointly sufficient conditions for cardinal scales: linearity and comparability. This claim does not require much explication: if your scale use is linear, that makes it cardinal for you at that time. If scales are also comparable across people (and/or times), that makes them cardinal across people (and/or times).

Of course, we may doubt that people are willing and/or able to answer happiness (and other) surveys in a linear, comparable way. If we think of subjective scales as measuring sticks for feelings, then people's 'sticks' could be crooked (non-linear) and/or different lengths (non-comparable). There is always some error in measurement, and measurement instruments don't need to be perfect to be useful. Hence, the important question to ask is: are measuring instruments so far from cardinal – so crooked and oddly lengthed – that we draw the wrong conclusions by assuming they are cardinal?

Following Kahneman et al. (2016) it is helpful to distinguish two types of measurement error here: *noise* and *bias*. *Noise* is the random variable of errors, whereas *bias* is a systematic deviation from the true answer. See Figure 1.

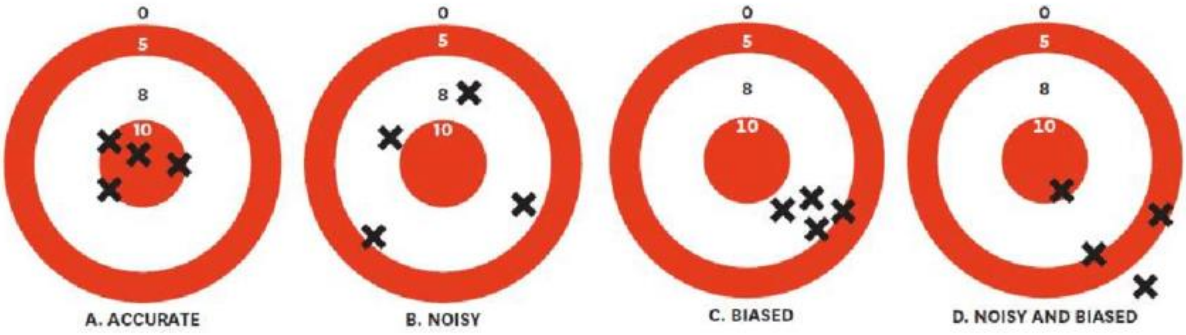


Figure 1. Distinguishing accuracy, noise, and bias. Figure from (Kahneman et al., 2016)

Random variation in individuals' scale use – noise – is not a threat for cardinal interpretation, as researchers have noted (Bronsteen et al., 2012; Dolan & White, 2007).[3] If surveyed populations are randomly selected and large enough, any deviations will 'wash out': those using a 7/10 rather than a 6 will be cancelled out by those using a 5 rather than a 6, so the average answer is accurate. Analogously, if you had many

[3] Although it will create a separate issue of attenuation bias, i.e., a bias towards zero in the results. This can be address via appropriate controls (Bertrand & Mullainathan, 2001).

poorly made thermometers whose errors were random, and took an average, this average would converge on the true answer the more thermometers used.

We do need to be alert to bias – non-random deviations – as these won't 'wash out' with more data: perhaps the Finns use their scales differently others, so they are not the most satisfied with life, despite what their numbers suggest. Similarly, if your thermometers were systematically wrong, you would not get the true answer even with infinite readings.

Hence, a natural framing of the debate – are subjective scales cardinal *or* ordinal? – is a false dichotomy (Frijters, 1999; Peart & Levy, 2005; van Praag, 1991). There is a third option: subjective scales are *quasi-cardinal*: they do capture some quantitative information, but there is *some* bias – exactly how much is a further question. Returning to the measuring sticks analogy, what I will call the *cardinalist* position is that people are (in aggregate) using straight measuring sticks of the same length – hence it is correct to treat the scales as cardinal. The *quasi-cardinalist* view is that our measuring sticks are bent and/or different lengths to *some* degree: we can't take the numbers at face value because there is a non-random error. This view does allow that we could get cardinal data: if we know what the problem is, and how bad it is, we could correct the scales ('straighten and adjust the sticks'). This would present a further methodological challenge, but is *not* a terminal problem for ever interpreting subjective data as cardinal. The *ordinalist* view is that we can't use measuring sticks for feelings at all. Those with this view will either believe feelings are not quantities (i.e., lack phenomenal cardinality), or alternatively, feelings are quantities, but we have no idea how linear or comparable the scales are (these beliefs are mutually exclusive: feelings cannot both be and not be quantities). Either way, subjective data cannot be interpreted as cardinal. This three-way distinction will be important later: while some may be tempted by ordinalism, I argue it is implausible (§5), which focuses our attention to the remaining two options.

In this essay, I focus on a particular kind of subjective scale, namely those that measure *subjective wellbeing* (SWB) – self-reported assessments of quality of life, such as happiness and life satisfaction (Dolan et al., 2011; Dolan & White, 2007; OECD, 2013). These are especially important because they plausibly capture some or all of our *wellbeing*, what ultimately makes our lives go well for us (Layard, 2005; K. Stiglitz et al., 2009). If there is, as I will argue, a common way of interpreting subjective scales (§4 and later), I imagine much of what I say will apply to other ratings of feelings, such as, "How good is this restaurant, on a scale of 1-5?"

Finally, I am only concerned with whether measures of feelings are cardinal. This is conceptually distinct from issues about what *wellbeing* is and whether cardinal interpersonal comparisons of welfare levels are possible (Broome, 2004; Crisp, 2008; Haybron, 2016). One might suppose wellbeing consists in preferences over states of affairs (what economists usually mean by 'utility') and that these ranking are ordinal in nature and thus not comparable on a cardinal scale (Hausman, 1995). One could hold this while still accepting that happiness and other feelings are quantities and could be measurable on cardinal scale and thus compared between people (Ng, 1997).

3. What is the rational way for respondents to interpret subjective scales? The Grice-Schelling account

Discussion about the nature of subjective scales – that is, the assumptions of linearity and comparability – tends to be quite mathematical. For instance, Ferrer-i-Carbonell & Frijters (2004) use different statistical tests suited to cardinally and ordinally comparable data, find these don't give different results, and conclude it is sensible to treat the data as cardinal.[4] While such work is helpful, it misses a human element: what's going on in people's heads? How do we try to answer these questions?

In this section, I propose an account of how it would be *rational* for people to use subjective scales, *assuming their aim is to accurately communicate their feelings*. Rationality here is simply understood instrumentally: choosing the right means for a given end. Economics often operates on the assumption that ascertaining what is rational for people to do is informative, though not decisive, for understanding behaviour. An appeal to rationality is particularly helpful here, given we cannot objectively measure feelings. Readers unhappy with the term 'irrational' could replace it with 'eccentric'.

Let's make the challenges of using subjective scales explicit.

Suppose I ask, "How happy are you, on a scale of 0-10?" This is an easy, familiar question. It does not require effortful thought, such as, "What is 15 x 15?" (Kahneman, 2011). Nor does it seem confused, such as, "How tall is the King of France?" (cf Russell (1905)). Notice we can easily and quickly give apparently meaningful answers about many properties, not just happiness. This is true even when the scales are vaguely labelled scales of objective dimensions ("How tall are you, 0-10?") or we've never rated the thing before ("How good are those clouds, 0-10?"). It is not difficult to answer when no scale is given: if I ask, "How good was the concert?", we might use a verbal label ("Pretty good") or a number ("Hmm, 7 out of 10"), and possibly an explanation.

Although these questions are intuitively easy to answer, doing so requires respondents to solve three problems, as Fleurbaey & Blanchet (2013) point out: (1) the *scope* problem: "What information is important for my evaluation?", (2) the *ranking* problem: "How do I rank the options based on the information in my scope?", and (3) the *calibration* problem: "How do I translate a position in this ranking into a numeric value

[4] See footnote 2.

on a finite scale?" Here, we are just concerned with the calibration problem.[5] This requires respondents to fill in two sets of details about the scale.

One is to decide what the endpoints of the scale mean: how happy do you have to be to be a 10/10 or a 0/10? And, how can this unbounded phenomenon be constrained to a bounded scale (I return to this shortly)? Even if I tell you that 10/10 means 'very happy' and 0/10 means 'very unhappy', you must still decide what *those* mean.

The other choice is your *reporting function*: the magnitude of difference between each point on the scale (Oswald, 2008). You might decide that the difference between each point on the scale is equal-interval – a linear reporting function. But you might do something else: you could treat your happiness scale a bit like the Richter scale for earthquakes, using a logarithmic function so that each 1-point scale increase represents a 10-fold increase in feeling.[6]

What would be rational for individuals to do?

An observation made by philosopher of language Paul Grice is that conversations are cooperative endeavours, where speakers and listeners rely on each other to think and act in certain ways in order to be understood (Grice, 1989). Grice proposed the *cooperative principle*: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged". This principle has several maxims, which are, roughly: to be truthful, to give no more and no less information than required, to be relevant, and to be clear. It's because of these background maxims that there are what Grice called *conversational implicatures*: conclusions that hearers can draw about what the speaker means without them being literally stated. For instance, if I turn to my colleague and say, "Get the door, will you?", they are likely to conclude I want them to close the door – perhaps because there's a draft – and not that I want them to take the door off its hinges and bring it to me, even though the latter is more literally implied.

We can take two relevant lessons from this. First, individuals often aim to be cooperative communicators. Second, they will use the available contextual information to determine how best to achieve this end. Hence, it seems reasonable to assume that, *when surveyed about their feelings*, their aim will be to convey those accurately, and use background information to consider how best to do that (Schwarz, 1995).

The difficulty for respondents is that subjective scales are vague and they cannot communicate with each other about which exact interpretation to use. Rational respondents would conclude that to successfully cooperate requires anticipating how

[5] We don't need to answer (1) and (2) here. That is what we hope to learn from the surveys: what it is that affects people's assessments of their happiness, for example.

[6] An interval scale is one where differences are meaningful. Both linear and logarithmic scales are interval scales, but the intervals are different in each scale. In this paper, I understand a cardinal scale as a linear interval scale.

other people interpret the scale and then attempt to interpret it in the *same* way so that their answers have the equivalent meaning. To illustrate, if I am confident my 6/10 represents a different amount of happiness from everyone else's 6/10, but I use it anyway, I am acting irrationally – or at least eccentrically – if my goal is to be understood. Similarly, if you point at a dog, and I know everyone else calls it a 'dog', but I insist it's a 'cat', I should expect my eccentric communication will cause others to misunderstand me.

Turning from philosophy to economics individuals are, in game theoretic terms, seeking a *Schelling point*, also known as a *focal point* – a default solution picked in the absence of communication (Schelling, 1960). The most famous illustration of the Schelling point is the New York question: if you are to meet a stranger in New York City, but you cannot communicate with the person, when and where will you choose to meet? Thomas Schelling, the economist after whom the term is named, asked a group of students this question, and found the most common answer was noon at the information booth at Grand Central Station. Although one could meet anywhere, certain options are, for whatever reason, more salient and more likely to lead to successful coordination.

So, on what we could call the 'Grice-Schelling' theory, individuals are trying to cooperate and make themselves understood, which they do by coordinating around focal points in their interpretation of subjective scales. Two quick points. First, note that, in contrast, there is no need to find focal points when using scales of objectively measurable properties: if I ask you your height in centimetres, there is no uncertainty about how to use the scale. Second, although I frame this as how someone might reason, starting from first principles, I am not claiming that people do engage in such reasoning when presented with 0-10 scales. They do not need to, because they absorb linguistic norms and thus can do so intuitively.

What, then, are the focal points for subjective scales? I suggest these are: (1) using a linear scale, and (2) taking the endpoints as the realistic limits of whatever is being measured – in other words, the top means 'most' and the bottom means 'least'. So, 10/10 for happiness is the most happiness you could have, 10/10 job candidate is the best job candidate you could expect, and so on.

Now, 'maximum realistic limit' is admittedly vague: Is this the happiest I have been? The happiest I could be? The happiest anyone has been? Is this vagueness a problem? Not necessarily. Respondents are not trying to guess the unique 'correct' answer: they are just trying to guess what *other* people will use. What's more, they only need to be in the 'ballpark': so long as differences in scale interpretation are random, they will 'wash out' and we can rely on the 'wisdom of crowds'. Note that in Galton's famous experiment of asking people to judge the weight of an ox at a country fair, the average guess is surprisingly accurate (Galton, 1907). Hence, we just need to worry about different groups having substantially different choices of endpoints.

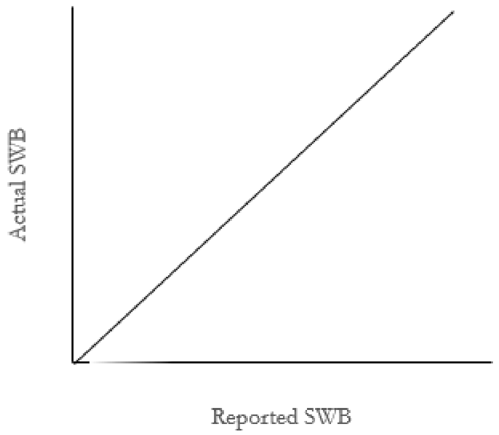
Regarding the reporting function, the obvious option is to use a linear scale. This is because we're used to using cardinal scales in ordinary life: we use them for height, weight, length, and so on (Ferrer-i-Carbonell & Frijters, 2004). Hence, they are the familiar, default option. For a comment on how the Weber-Fechner law is mistakenly taken as evidence for non-linear reporting, see footnote.[7]

I imagine some readers will assume, on reflection, these are the obvious choices and wonder both what the alternatives are and whether people might use them. I'll consider alternatives to the reporting function first, then return to the endpoints.

One alternative to the linear reporting function is the arc-tangential function proposed by Ng (2008). Each is displayed below in Figures 2 and 3. On the arc-tangential function, the ends of the scale represent larger differences than the middle. Ng's rationale is that, as there is no logical limit to happiness, the use of a linear function that covered the full *logical* range would make typical changes to happiness impractical to report. For instance, becoming unemployed would only register an interpretably tiny difference on the scale – say, a move from 5.1 to 5.100002. Ng supposes the advantage of the arc-tangent is that it makes the scale's middle comprehensive, while still allowing very high and low happiness scores to be represented at the top of the range.

The challenge, if you want to make yourself understood, is that the other person could not possibly guess what *specific* deviation from linearity you are using – how bendy your measuring stick is. Whilst you may know what *you* mean by 8/10, the surveyor will not, and will instead assume you mean the same as everyone else. *Pace* Ng, you are acting irrationally because this makes your scale use incomprehensible to others – like expecting someone playing Schelling's New York game to meet you at your favourite deli. *Mutatis mutandis*, this issue of incomprehensibility applies to all non-linear reporting functions, such as logarithmic reporting functions.

[7] Support for a logarithmic function for wellbeing is sometimes derived from the Weber-Fechner law in psychophysics: it takes (roughly) a doubling in some objectively measured property – e.g., light or sound – for people to feel a 1-unit difference in subjective perception of intensity (Portugal & Svaiter, 2011). That law refers to the presumed relationship between changes in (1) objective properties and (2) actual experience. We are concerned here with a different relationship, that between (2) actual experience and (3) reported experience. Yet, to conclude there is the logarithmic relationship between (1) and (2), given the data, one must *assume* a linear relationship between (2) and (3). (If the relationship between (2) and (3) was, instead, logarithmic, we would observe a linear relationship between (1) and (2)). Hence, the Weber-Fechner law cannot be evidence against linear reporting in wellbeing when it assumes linear reporting(!). Examples of this confusion are (Gómez-Emilsson, 2019; Y.-K. Ng, 2008; Wodak, 2019).



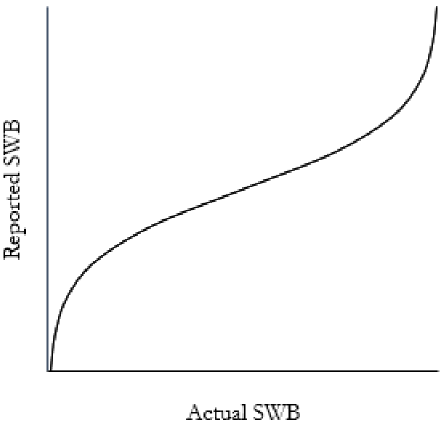


Figure 2. Linear relationship Figure 3. Arc-tangential relationship

Why is the rational choice to use the realistic limits for the scale endpoints?

If the endpoints don't cover the full range of intensity levels, we are going to run out of room on our scales. If I rate myself as 10/10 today, when I know I could be happier tomorrow, how can I distinguish those answers? You must either change your scale, which the surveyor won't know you'd done, or use a compressed, ambiguous scale where 10 can mean different things; that reduces your communicative accuracy. Indeed, it is *only* possible to use a linear reporting function if your scale covers the full range of the values. As Figure 4 demonstrates, if your scale were linear in the middle, but whose range did not contain the limits, the end categories would effectively expand to fill the remaining space. I've already said non-linear scale use is irrational.

Why not use a longer-than-actual scale? What would we choose? There's no logical limit to happiness, so there's no sensible way to report that on such a scale (see Ng's point above). Similarly impractical would be the *nomological* limits, the maximum within the laws of nature – where are those? You could use something arbitrarily larger than the real limits (for instance, two times the maximum) but you couldn't expect others to match your arbitrary choice.

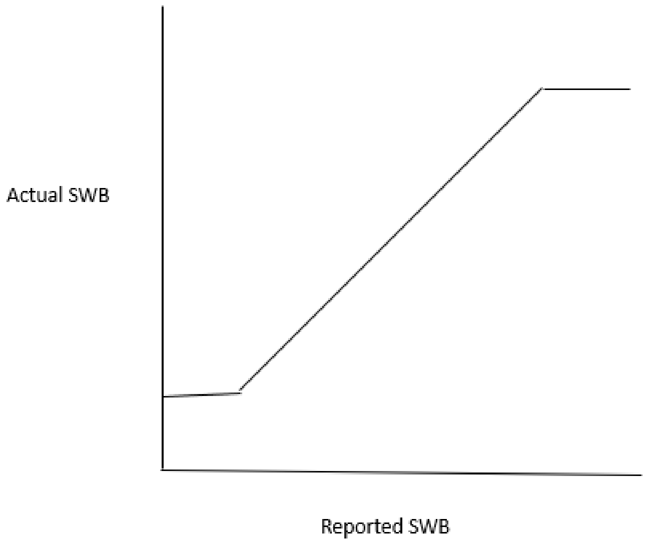


Figure 4. Sub-actual SWB scale: scale is necessarily (partially) non-linear unless it covers the actual range.

What makes the actual limits the sensible Schelling point, in addition to the above, is that you will have some idea of where they are, and so will others, and you know this. For instance, you observe many of states of happiness, and we communicate about this a lot, so you will have a notion of what maximum happiness is. An implication of this is that endpoints for different groups will converge to the extent that each group intuitively uses a reference class with the same maximum and minimum limits. This is more believable for some subjective scales than others. To illustrate, we might expect that a Dutchman (average male height, 6 ft ½ in) and an Albanian (average male height, 5 ft 9 in) will take 'very tall' or '10/10 tall' to refer to different objective heights in virtue of using their own relevantly different national reference classes. It does, however, seem plausible that the upper and lower bounds of feelings are much the same and observable in every society: we share the same biology and we all see or experience what seem to be the equivalently high highs (falling in love, clinching victory in a tournament) and low lows (destitution, bereavement).[8] Some suggestive, non-definitive evidence of this will be presented in §7.2.

Hence, on this Grice-Schelling story, the rational choice is using a linear scale and the real endpoints. If scales are linear and have the same endpoints, they are also comparable, as each number on the scale has a consistent meaning across people.

[8] Research suggests that people's concept of 'good health' is rather mutable and changes, for instance, with age (Salomon et al., 2004). Health seems relevantly different from happiness: plausibly, when a 20- and 70-year-old talk about their health, being cooperative communications requires they use a different reference class – their own age – precisely because health functioning varies so much with age. It does not seem that happiness is so age- or context-variable. To illustrate, note that 'healthy for a 20-year-old' and 'healthy for a 70-year-old' do appear to mean different things, while 'happy for a 20-year-old' and 'happy for a 70-year-old' do not.

A linear, comparable scale is a cardinal scale – so, if people do interpret subjective scales in the way outlined, subjective scales will be cardinal.

We can now turn to the other paradoxes.

We worried about having a bounded scale to measure something unbounded. My suggestion is we naturally stretch our scales so that they include all the actual possibilities.[9] In this case, although the scale is bounded in theory, it is not in practice, because it captures all the cases we care about – namely, the actual ones. This is a bit like a kettle measuring water temperature up to 100 degrees centigrade: H_2O can be hotter, but if want to measure water, it's not important for the kettle's scale to detect this.

What about the claim that numerical scores of feelings being "are 'made-up' numbers on a scale that does not exist" (Kaiser & Oswald, 2022a)? Feelings and judgements are certainly real. The numbers we use indicate an intensity of feeling on a scale where the endpoints represent the limits of intensity. "7/10" happy is then just "70% of the way from minimum to maximum happiness". Although putting numbers on feelings may seem odd, is this any stranger than using words to describe our strength of feeling ("quite happy", "very happy"), something we do not consider problematic (see §7.3)? Hence, a sceptic mind to deny that numerical measures of feeling are meaningful is in danger of 'proving too much': if numerical description of feelings are meaningless, given they seem interchangeable with verbal ones, that implies, implausibly, that verbal descriptions are also meaningless.

To summarise: I have argued that, on the Grice-Schelling account of scale interpretation, it is rational for respondents to use a cardinally comparable scale if they want to be understood. It is difficult see how we could expect to accurately communicate our feels except by using a linear, comparable scale.

4. What if we're uncertain about how people interpret subjective scales?

The last consisted in reasoned speculation, but speculation nonetheless. We should be uncertain whether the cardinality assumption is true. Should uncertainty lead us to reject cardinality, perhaps as a precautionary move? I now argue it should not. My argument here is not novel, but rather an elaboration of the 'washing out' hypothesis given earlier.

In §1, I mentioned three views on the nature of subjective scales: cardinalism, quasi-cardinalism, and ordinalism. To bring out the difference, suppose X reports being 1 point higher than Y on some 0-10 scale. X and Y could refer either to different

[9] Some people say it's not possible to be 10/10 happy. One way to explain this is people intuiting you need a scale big enough you can always answer within its bounds.

individuals or groups, or to individuals or groups at different times, or both. What would each of the positions conclude is the true difference in magnitude between X and Y?

The cardinalist holds X is 1 higher than Y (in expectation). This is because cardinalists assume linearity and comparability. We might ask what counts as cardinalism: must cardinalists believe X is <1.1 higher? What about <1.0001? I won't offer a mechanical definition here because, as noted (§1), the important test is a practical one: does assuming cardinality yield the wrong results? More on this in §6-10.

The ordinalist, as I use the term, thinks we cannot conclude anything about the differences: we are entirely in the dark. They believe we can assume nothing about the respondents' choice of reporting function or endpoints.

The quasi-cardinalist believes we can determine the quantitative difference between X and Y (unlike the ordinalist) but that it is not 1 higher (unlike the cardinalist). Because quasi-cardinalists believe linearity and/or comparability are violated to some degree, it represents not a single view, but a spectrum of positions (see §5). So, perhaps one quasi-cardinalist believes X is truly 0.7 higher than Y, another that X is 0.1 lower, and so on. Note that a quasi-cardinalist who insists we know nothing about the degree of violation has become an ordinalist.

What if we're uncertain about the difference? We can represent this uncertainty in terms of a probability distribution. In Figure 5 there are two different cardinalists: the *confident* cardinalist thinks there is very little chance X is more or less than 1 greater than Y, whereas the *inconfident* cardinalist has greater doubts. But both remain cardinalists: although both think subjective scale data are noisy, neither thinks the data are biased.

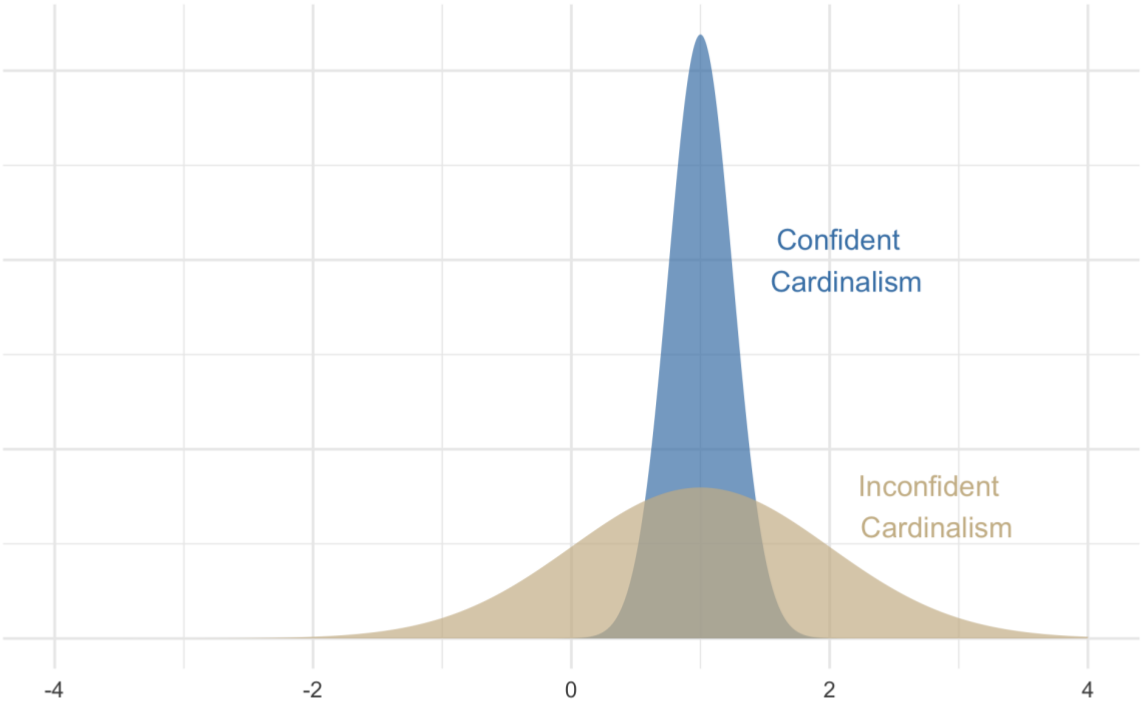


Figure 5. Comparing the uncertainty of a confident and an inconfident cardinalist

How would we represent the other views? The ordinalist thinks we can say nothing, which means they are equally spread among the possibilities and are maximally uncertain.

The quasi-cardinalist, to be distinct from the other two, must have a view about what, in expectation, the difference is: they must believe there is bias, and they may believe there is noise. Again, we could contrast a *confident* quasi-cardinalist from an *inconfident* quasi-cardinalist. The confidence the quasi-cardinalist has is logically distinct from what, in expectation, they believe the difference to be. To illustrate this, the confident quasi-cardinalist in Figure 6 below believes the expected difference is 0.7 greater, while the inconfident one believes the difference is 0.1 lower.

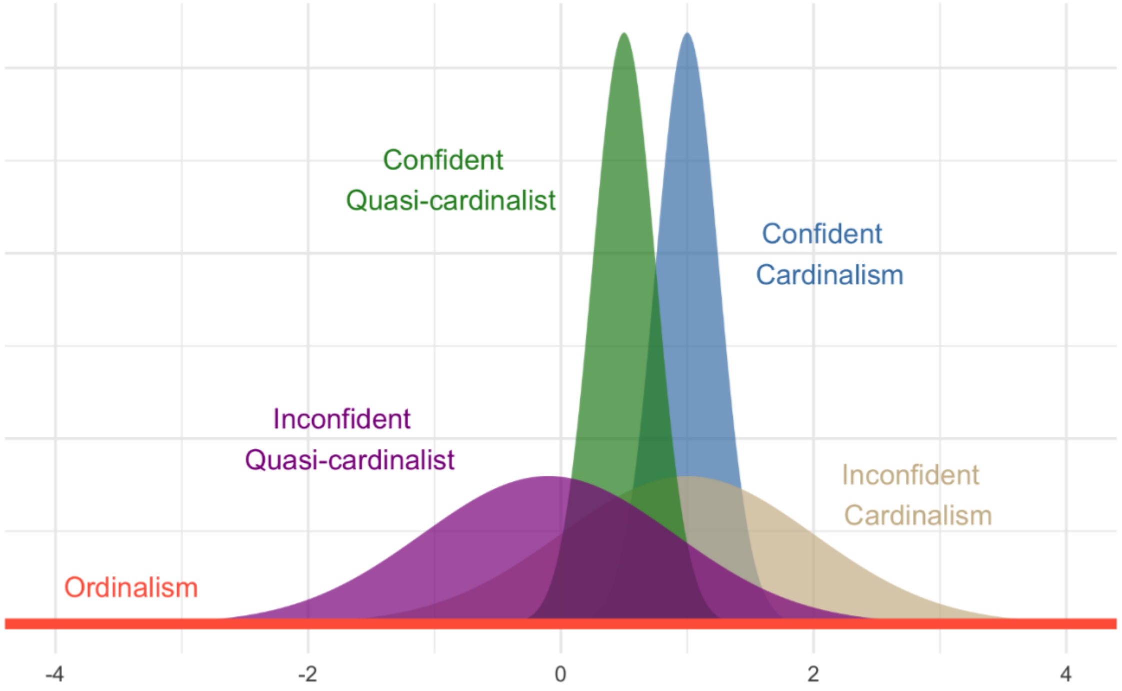


Figure 6. Comparing the uncertainty of different views

From this we can see that having some uncertainty does not draw us from cardinalism. We need extreme uncertainty to get to ordinalism, and quasi-cardinalism requires having some beliefs about how subjective scales detour from cardinality. Should we accept ordinalism, then? In §5, we'll see that ordinalism's level of uncertainty and should be ruled out. This leaves us to choose between the remaining options: cardinalism or quasi-cardinalism.

5. *Contra* ordinalism

Ordinalism, as I've defined it, implies we can make no quantitative comparison of people's feelings. I am not sure if this position has any true believers, but some appear to think we should honour something like it in practice (T. Bond & Lang, 2018; Robbins, 1932; Schröder & Yitzhaki, 2017). For my purposes, it does not matter if I am attacking a straw man: my aim is to show the view is implausible so that I can go on to say other things; if no one believes it, that is all the better for my argument.

Let's return to the case from the previous section: X says they are 7/10 happy, Y says they are 6/10. An ordinalist would say that we have absolutely no idea who is happier – not that we are *a bit* uncertain, but we are so uncertain we must assign equal probability to each. Presumably no one really believes this; if this example is not sufficient, consider what you would think if X said 9/10 and Y said 2/10. Anyone who does not assign equal probabilities to each being happier is *not* an ordinalist.

This indicates ordinalism is implausible, but it doesn't explain why it is.

The first point, almost too obvious to state, is that we are able to use language to communicate; that indeed is its point. If we can communicate in general, then it follows we are also able to communicate specifically about how we feel. Putting numbers on a scale is a way of doing that. It would be puzzling if our conversations were mostly intelligible, but those about our feelings were entirely unintelligible.

We can also ask what we would expect to see in the subjective data if ordinalism were true. If the numbers had no shared meaning, respondents would either not answer the question, or their answers would be random. Yet, we find neither. In household surveys, response rates to SWB questions are around 96-99%, indicating people have no difficulty answering such questions (Bonikowska et al., 2014). What's more, we find all sorts of patterns in SWB: higher SWB is associated with being richer, being employed, being in a relationship, and so on (Alexandrova & Haybron, 2016; Diener et al., 2013, 2018; Dolan et al., 2008). Hence, people cannot be answering at random.

This seems sufficient to rule out ordinalism. It shows we are not so uncertain about subjective scales we think they tell us nothing. If we believe self-reports capture some quantitative information, we must be either cardinalists or quasi-cardinalists.

6. A two-horse race? Cardinalism and quasi-cardinalism

It might seem we're now down to two options – cardinalism or quasi-cardinalism – and those sceptical of the former should accept the latter. Matters are not so simple for the aspirant quasi-cardinalist. Quasi-cardinalism is not a single view but refers to a spectrum of options. Given the two assumptions for cardinality are linearity and comparability, one becomes a quasi-cardinalist if one rejects one or both of those to some degree.

To illustrate this, we can represent the space of options as shown in Figure 5 A-F below. The two relevant axes are that subjective scales are more or less linear and more or less comparable. In the top right, we have cardinal scales (where we assume linearity and comparability); in the bottom left, we have ordinal scales (with no information about linearity and comparability). In between these two extremes are the quasi-cardinal views.

Figure 5A The possibility space of the interpretation of subjective scales

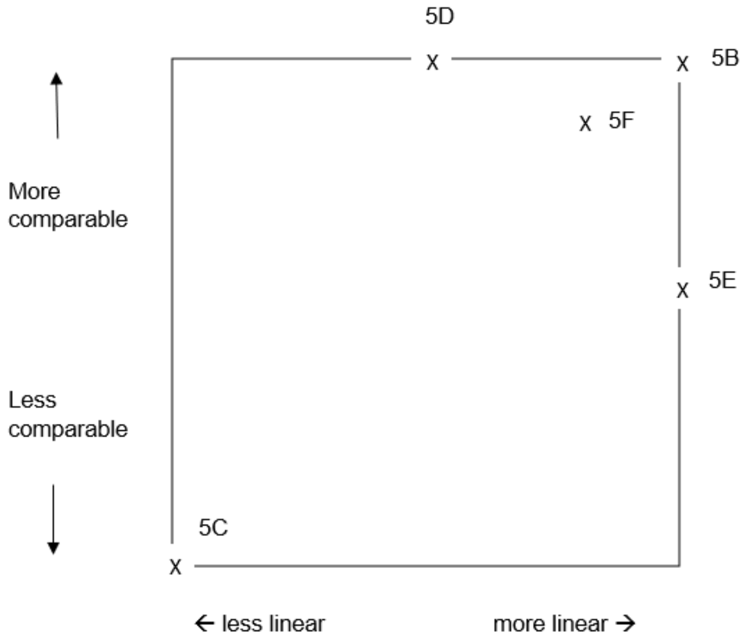


Fig 5B. Cardinal scale (linear and comparable)

Fig 5C. Ordinal scale (relationship unknown)

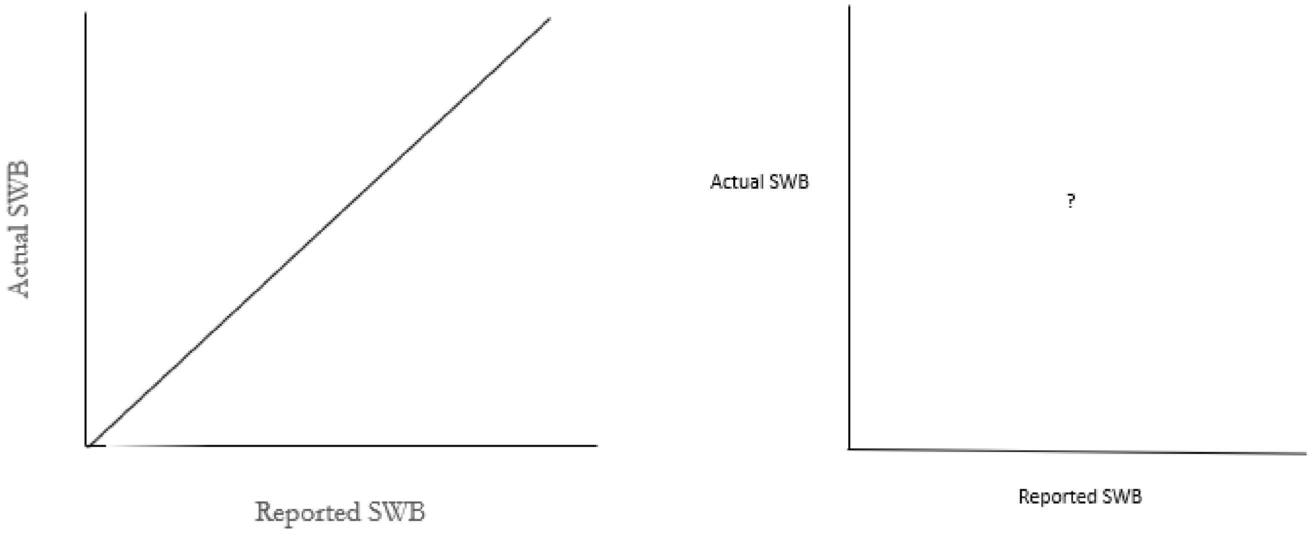


Fig 5D. Arc-tangential scale (non-linear, comparable)

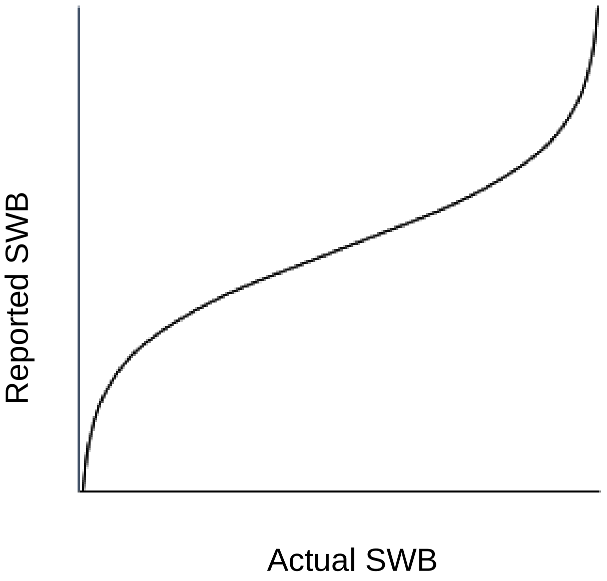


Fig 5E. Two linear, non-comparable scales

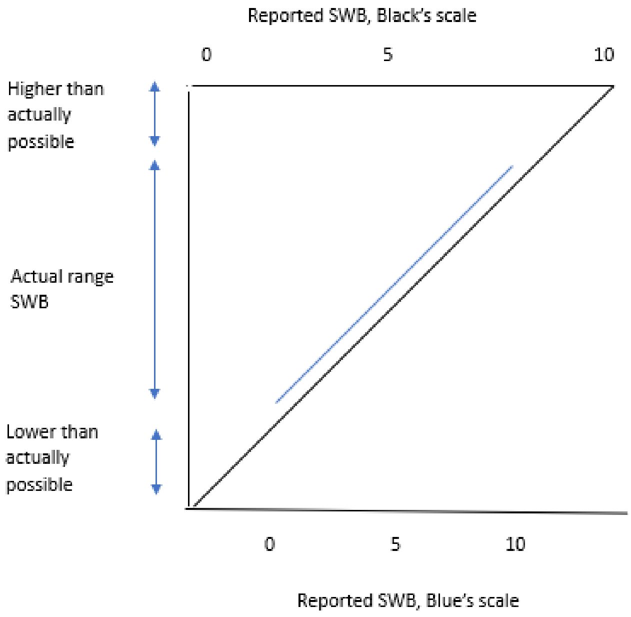
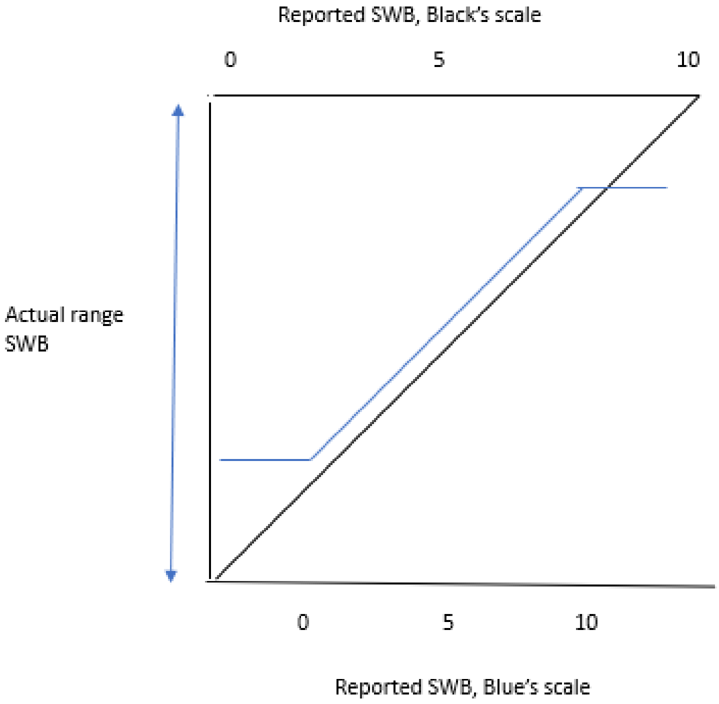


Fig 5F. Two partially linear, non-comparable scales



The important point here is that there is not a single quasi-cardinalist position. For example, advocates of the scales in 5D-F agree that subjective scales are not cardinal but disagree about what the problem is. Hence, sceptics of cardinality cannot simply reject cardinalism: they must opt for an alternative – and, ideally, justify that alternative.

What is the most plausible version of quasi-cardinality, given the evidence, and how strongly quasi-cardinal is it? By ‘strong’ here, I mean in the sense of how large its deviation is from cardinality.

7. Reviewing the evidence for deviations from linearity and comparability

In the following three sub-sections, I consider, in turn, some empirical evidence relevant for assessing linearity, intertemporal comparability (whether each individual changes their scale use over time) and interpersonal comparability (whether different individual use different scales at a time). In each case, the story is similar: using the principle of inference the best explanation, the evidence is consistent with the deviations from linearity and comparability being small or non-existent, but not with large deviations. These are ‘small’ in the sense that, even on the interpretations sceptical of cardinality, for the plausible amounts of bias, few, if any, of the standard conclusions in the literature would change (compared to those reached assuming cardinality). If one wants to be a quasi-cardinalist, the most – perhaps only – credible possibility is one based interpersonal differences; I expand on these details in the text below.

I stress that my analysis here is necessarily brief and non-exhaustive: while there is not yet much research to draw on, discussion of each topic could fill a paper. This is an opening salvo to spur discussion, not the final word: it is an open question whether there are other and better tests, and what new research would find.

It would be repetitive to say this in each case, so I’ll make the point once here: if the Grice-Schelling analysis is correct, we should expect people to use linear, comparable scales because that is the rational response.

7. 1. Are subjective scales linear, i.e., equal-interval?

A number of different pieces of evidence point to linear scale use. I am not aware of any evidence that suggests substantial use of non-linear scales.

The first comes from Oswald (2008) who asked respondents to rate their own height relative to their gender, on a horizontal line labelled “very short” on the far left and “very tall” on the far right. Ten small equidistant vertical dashes were marked as a visual aid. The objective height of the participants was also measured. The correlation between subjective and objective height was very high (0.8) and regression equations found the relationship between subjective and objective height was effectively linear. This indicates individuals treat numerical, bounded scales of objectively measurable properties – in this case, height – as linear.

In a second study, van Praag (1991) gave subjects ordered evaluative verbal labels ("very bad", "bad", "not bad", "not good", "good", "very good") and asked subjects to place these on a cardinal numerical scale labelled with endpoints "1" and "1000". The general pattern across individuals was to place the labels so they were roughly equal distances apart on the scale; in other words, individuals constructed a cardinal scale with the ordered subjective data.

A final compelling, but indirect, argument emerges from the *homoscedasticity* of errors in subjective reports. Krueger & Schkade (2007) conducted a test-retest of net affect – individuals are asked how happy they are one day, asked again a week later, and the results are compared. Intuitively, individuals' happiness varies by about the same amount from week to week, regardless of their *level* of happiness – we don't observe those who are very happy to have wild swings in their moods whilst those who are unhappy have small changes, or *vice versa*. Because *actual* differences in happiness should vary by the same amount regardless of level (if the scale is linear), what we would *observe* is the *reported* differences vary by the same amount at different parts of the scale. With a linear scale we expect, technically, *homoscedasticity*: for the error in the regression model to be constant as the value of the predictor variable changes. Krueger and Schkade find that the test-retest differences for *reported* levels of net affect are close to homoscedastic. To see why this indicates linearity, imagine what we would find if we had a logarithmic scale and those reporting 10/10 were 1,000 times happier than those who were 7/10: we'd expect the *reported* test-retest differences for the 10/10s to be 1,000 times *smaller* than the 7/10s because, even though the actual average change per person is the same, the scale is so much larger at the top end.

The strongest apparent argument against linearity comes from Bond & Lang (2018). Bond and Lang point out that happiness scales are logically unbounded but individuals have only limited numbers of labels; therefore, reports in the top or bottom categories could potentially be infinitely large or small. Hence, an individual who reports being in the top category – say 10/10 – may have an actual level of happiness that is hundreds or thousands of times higher than other individuals also in that top category, or the category directly below. Under these conditions, it is possible to reverse results that are found assuming a linear scale – for example, that greater income is associated with lower wellbeing, rather than higher.

There are two issues with this argument. First, Bond and Lang should be understood as making a hypothetical argument about what may follow *if* scales are non-linear. Whilst their hypothetical argument may be correct, Bond and Lang do not provide evidence to support their claim that individuals do use a (strongly) non-linear reporting function. On the contrary, as we've just seen, it seems individuals do use a linear reporting function.

Second, Kaiser & Vendrik (2020), who replicate and modify Bond and Lang's approach, argue that the reporting function would need to be strongly non-linear and that reversals are "impossible or implausible for almost all variables of interest".

Hence, it seems reasonable to treat subjective scales as linear.

7. 2. Are subjective scales comparable over time?

Ng (2008) observes that happiness researchers seem not to have noticed that individuals can *rescale* – alter what a scale's endpoints represent – over their lives. Although substantial rescaling seems irrational, given the objective of accurate communication – because it changes the meaning of one's self-reports – it may happen anyway. For example, if you encountered something that is better or worse than you had previously experienced or imagined, you would adjust the limits of your scale accordingly.

There seem to be two sorts of evidence here: life shocks and memory data. Neither provides conclusive evidence that people rescale. Furthermore, if rescaling over time does occur, it does not seem to be large enough to affect the results.

There is literature on how people's subjective wellbeing changes in response to major life events such as marriage, getting into a relationship, bereavement, becoming unemployed, or becoming disabled (Clark et al., 2008, 2016, 2018; Luhmann et al., 2012). In general, people's *reported* wellbeing returns *towards* its pre-event level in the years following.

There are two possible explanations: rescaling (as mentioned above) and *hedonic adaptation* – where the subjective experience of goodness or badness reduces over time. To illustrate the difference, suppose Sam reports he is 8/10 happy. He then has an accident and, two years later, reports 8/10 again. It could be Sam has *adapted* and is genuinely as happy as he was before. Or, it could be that Sam is less happy but has *rescaled* – specifically, he has shrunk his scale, lowering the level of happiness a 10/10 represents. Some combination of both of these effects could be at play. Unlike rescaling, adaptation poses no threat to intertemporal cardinality, as the same numbers still represent the same intensities of experiences.

If reported adaptation to life shocks can be explained as hedonic adaptation, we lack reason to assume rescaling occurs. How could we tell if the evidence on life events supports rescaling or genuine adaptation? The key point seems to be that while people *fully* adapt to some life events, such as bereavement and getting married, they only *partially* adapt to others, namely being in a relationship, becoming disabled, or becoming unemployed (Clark et al., 2018). If rescaling occurs as a result of cognitive process, we might expect to see it everywhere, but we don't. What's more, we can appeal to our wider intuitive understanding of human lives to explain the difference. For instance, it makes sense that getting married (as distinct from being in a relationship) has a short-term effect – the 'honeymoon phase' wears off and normality resumes – whereas being unemployed continues to feel bad. Hence, applying the principle of Occam's razor, we do have reason to assume hedonic adaptation occurs, but not that rescaling does.

The other test of rescaling comes from utilising memories. Prati & Senik (2020) compare *remembered* SWB – how satisfied individuals recall being in the past – with *observed past* SWB – how satisfied individuals said they were at the. Specifically, they analyse data from respondents of a German panel, who had been asked about their life satisfaction for years. Respondents were given nine different pictures of changes in life satisfaction over time (see Figure 6) and asked to pick the one that best represented their own life.

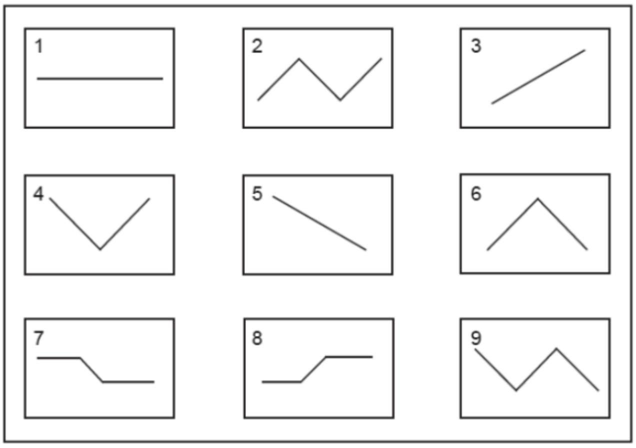


Figure 6. Potential patterns of recalled life satisfaction

Figure 7 displays, for each group that picked a schematic pattern, what their average *observed* life satisfaction was. It's worth stressing that this is an extremely cognitively demanding task, and the individuals were only given a limited range of options to pick from. The match between the patterns of recalled and observed past satisfaction is thus impressive.

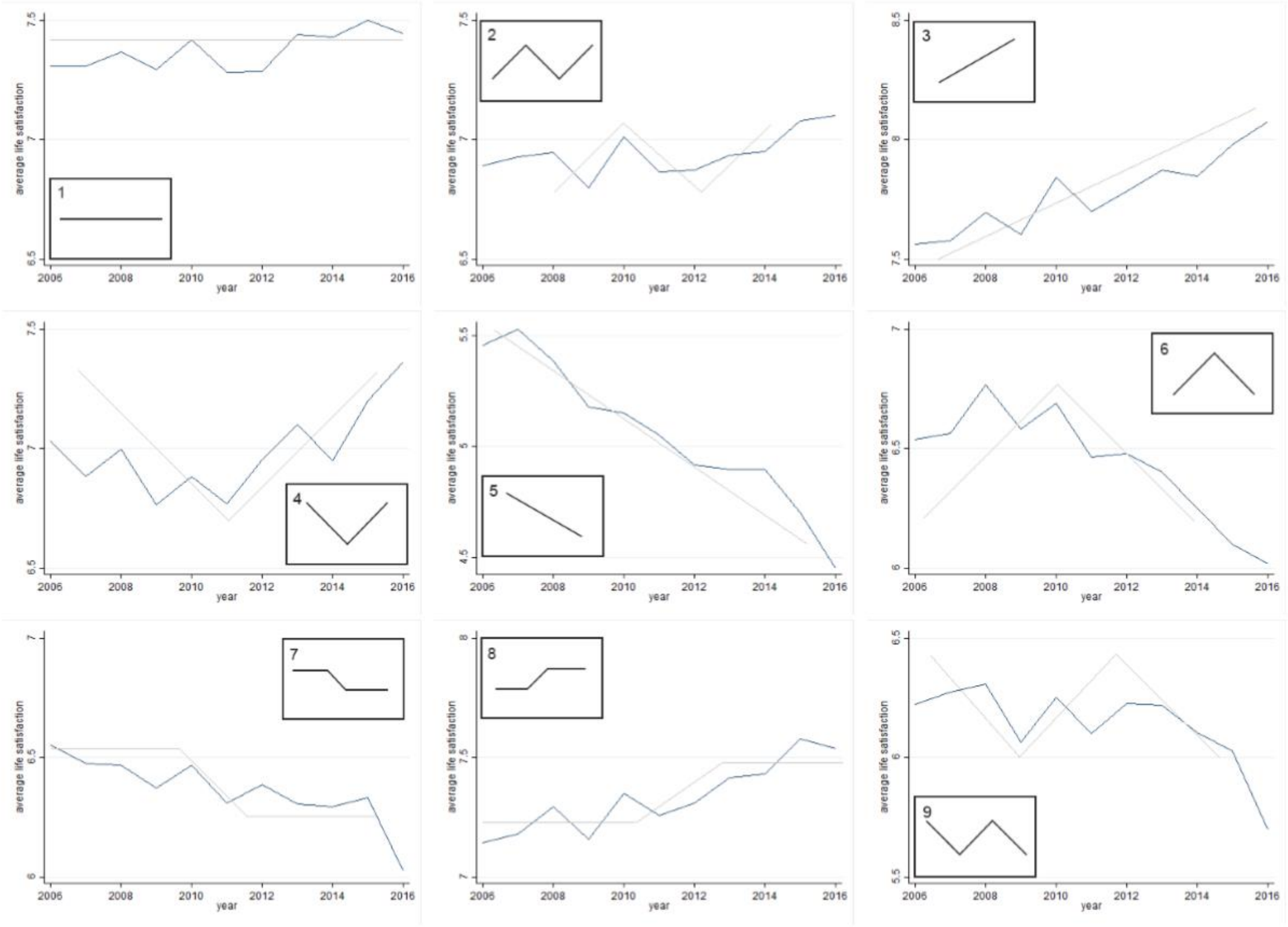


Figure 7. Observed past satisfaction, conditional on chosen pattern. Reproduced from Prati and Senik (2020)

It seems there is approximate consistency between remembered and observed past satisfaction. Let's assume for now that there is consistency, and ask: what would this imply about if and how rescaling occurs? In short, it suggests that it is possible, but unlikely. Let's illustrate with the conceptually simplest case, where reported and past observed wellbeing remain constant in numerical terms (i.e., the top-left image in figure 7, which admittedly do not show extreme consistency) and explain how this match could occur. What we need to be true for us to observe that the *self-reported numbers* for remembered and observed past wellbeing were flat? There seem to be three ways this could happen.

The first is that people have good memories and use the same scale over time. We observe *reports* of remembered and observed past wellbeing are flat because actual wellbeing was flat.

The second is that there is a rescaling, but the person misremembers their change in wellbeing. To elaborate, suppose the person's scale shifts up over time, so the same scores now represent higher wellbeing. This means their wellbeing has truly gone up. For them to report their remembered wellbeing as flat, they must have misremembered it as flat (assuming they want to accurately convey changes). Note that if they misremembered their wellbeing as having gone up or down, there would be inconsistency.

The third is that there is a rescaling, the person correctly remembers their change in wellbeing, then retrospectively adjusts their remembered wellbeing by using the scale they had at the time (rather than their current scale). Again, if the scale has shifted up, past observed wellbeing that appears flat indicates increased *actual* wellbeing over time. To then *report* constant past observed wellbeing, they would need to mentally adjust their remembered wellbeing numbers by the scale they used at the time. The surveyor would have no idea the person is changing the meaning of their scale over time. This means the person is engaged in consciously inaccurate reporting: they scores won't convey that their wellbeing has increased, even though it truly has.

Thus, rescaling occurs only in the second and third cases. People either have bad memories but are lucky (in the sense the remembered and past observed numerical scores match), or they have excellent memories, but choose to report it in ways the surveyor will misunderstand. Note these explanations are directly in tension: people cannot generally have both excellent and terrible memories. Hence, the simplest, most probable explanation is that there is little rescaling.

The analysis above assumed that past observed wellbeing and remembered wellbeing is consistent. We might reject this assumption – in which case, are we back to square one, given that inconsistencies could be due to either bad memories or rescaling? Not necessarily.

A helpful test of the scale of the problem is provided by Kaiser (2022) who analyses a dataset where people reported their wellbeing today, how they were last year, and whether they are better or worse than last year. A subset of people (about 6,000 of 75,000) say their life is better or worse than last year, but their reports show the opposite. This could be due to misremembering or rescaling. Kaiser supposes, for the sake of analysis, that inconsistencies are entirely due to rescaling. If they were, would the results change? Kaiser finds they would not: none of the variables would reverse their coefficients – e.g., increased income still has a positive associated with wellbeing, unemployment still has a negative associated, etc. This suggests that, even if all inconsistencies are due to rescaling, this rescaling is not large enough to make a practical difference to the data interpretation either.

6. 3. Are subjective scales comparable between different people?

How worried should we be about different people interpretation the endpoints of their scales differently? I suggest we can rule out large differences, but there may be small or no differences.

Research finds many things are associated with different subjective wellbeing scores: age, gender, income, employment, proximity to green space, and so on (Dolan et al., 2008). In every case, one explanation is that reported differences are due to different scale use (e.g., the rich aren't happier than the poor, they just use scales differently), and another is that the differences between people are genuine (i.e., the rich really are happier). And of course, in some cases, both factors could be at play.

It's not plausible *all* these differences are *entirely* due to scale use: that would imply, unintuitively, that everyone had the same underlying level of happiness. Hence, if one wanted to vindicate quasi-cardinalism, the approach should presumably be to identify the specific characteristics that result in differential scale use. Importantly, concerns about differences due to specific characteristics (e.g., gender, nationality, etc.) would only motivate quasi-comparability *for those characteristics*, not in general: if we believe that nationality alters scale use, we cannot extrapolate from that that all other characteristics do too.

At present, it is an open question of how best to test and adjust for this, and relatively little investigation has been done in specific characteristics. The workhorse method for assessing interpersonal comparability is *vignettes*: survey participants are given a description of someone's life – the vignette – and then asked to rate how satisfied that person is. Here is an example vignette taken from Angelini et al. (2014, 15):

> John is 63 years old. His wife died two years ago and he still spends a lot of time thinking about her. He has four children and ten grandchildren who visit him regularly. John can make ends meet but has no money for extras such as expensive gifts to his grandchildren. He has had to stop working recently due to heart problems. He gets tired easily. Otherwise, he has no serious health conditions. How satisfied with his life do you think John is?

If we assume *vignette equivalence* – that all individuals think the person in the vignette has the same underlying experience of life – then differences in reports will be due to differential scale use, meaning vignettes can be used to make scales comparable; it's not essential, for our purpose, to get into those details here (see Angelini et al., 2014)

The only characteristics I am aware of having been investigated with vignettes are nationality and gender (Angelini et al., 2014; Montgomery, 2022). Montgomery (2022) finds that, although women report higher average life satisfaction worldwide, after a vignette adjustment, we should conclude they are less satisfied. However, both the initially difference and the change are very small: the raw scores are 2.95 for women and 2.93 men, and after a proposed alteration (using men's reporting function), they are 2.91 for women and 2.93 for men.[10] The reversal occurs because the initial difference is tiny, and there is a similarly minor adjustment. What's more, Montgomery

[10] From the bottom of table 6 in Montgomery (2022). I reported these to two decimal places (from three) to improve readability.

(2022) also observes that other commonly studied characteristics are not reversed after adjusting for gender-based differences in reporting – e.g., income and marital status still have positive effects.

Similarly, Angelini et al. (2014) find that applying vignettes to various European countries changes the country ordering (e.g., Denmark drops from 1st to 3rd); yet, this adjustment results only in an average cardinal change for each country of about 0.2 on a 5-point scale.[11] Given developed European countries report around a 7/10 (on a 1-10 scale) in the World Happiness Report, whereas places like Lesotho and Rwanda report about 3/10 (see figure 9), this adjustment is sufficient only to shuffle European countries around within the pack, not to conclude countries we had thought satisfied were very unsatisfied

Taking these at face value provides a worst-case assessment of how results change if we are wrong to assume cardinality. Even on this worst-case assessment, the practical upshot is rather modest.

It's unclear we should take these at face value, though. The assumption of vignette equivalence is questionable. For instance, regarding John, about 30% of Germans rated him as very satisfied or satisfied, whereas 20% rated him dissatisfied or very dissatisfied, indicating genuine disagreement (Angelini et al., 2014). Different answers to vignettes could be due to differences in scale use or in their *evaluative standards* (what respondents believe matters) – and disentangling these is not straightforward. The concern in this paper is whether the numbers people use in subjective scales refer to comparable quantities of feelings; it is no threat to that if people have different reasons for why they have – or believe others would have – that level of feeling. Happiness research must be sensitive to different things mattering for different people.

An assumption that vignette equivalence relies on is *ordering consistency* – that is, different individuals agree on how to rank the vignettes. If there isn't an agreement on ranking, necessarily individuals cannot agree on levels of wellbeing. Standardly, vignettes contain only objective descriptions of life characteristics. A pilot study found that introducing subjective information about how the vignette characters feel during and about their own life (e.g., "She is generally happy") improved ranking consistency compared to standard vignettes (Samuelsson et al., 2023). If including subjective information improves the validity of the vignettes, that suggests a more informative test may come from directly focusing on description of feelings.

Another approach, then – and one I do not think has been articulated before – is to assume *semantic equivalence*: that is, individuals agree on the meaning of words. We can then test for *semantic-numerical* consistency: whether individuals agree on the numerical ratings that should be given to different verbal labels of feeling intensity. If there is agreement, that indicates interpersonal comparability in numerical scale use.

[11] Estimated from observation on figure 3 in Angelini et al. (2014).

Semantic equivalence (that we use words in the same way) seems substantially more plausible than vignette equivalence (that we agree how someone else's life is going).[12]

The only data I am aware of which bears on semantic-numerical consistency involved giving respondents a 0-10 scale, with 0 as 'very negative' and 10 as 'very positive', and asking them to describe how positive or negative various adjectives are, such as 'bad', 'average', 'fantastic', and so on (YouGov, 2018). This found that people tend to give broadly overlapping answers to each other, as shown in Figure 10. This is reassuring, indicating a general consistency. It's hardly surprising, given it is intelligible to answer questions such as "How happy are you?" with either words ("I'm great") or a number ("I'm an 8 out of 10"). It is not decisive, however, as it uses aggregate data: some individuals may have different number patterns – e.g., if someone assigns all words a score closer to the middle, and none at the extreme ends, that suggest differential scale use. This requires further exploration.

Even if we suppose there is interpersonal scale consistency within a linguistic group, such as a nation, that leaves open differential use between different groups. Let's consider international comparisons again. There are mean-level differences in SWB across countries (Helliwell et al., 2023). As before, these could be due to scale differences, be genuine, or be some mix of the two. Now, let's appeal to ordering consistency: if the (unadjusted) country average scores are in what appears to be the right order, that suggests respondents from different countries are using the same scale.

It should suffice to pick out a few countries. The Scandinavians famously lead the World Happiness Report ranking with a self-reported life satisfaction of around 7.5/10 (Helliwell et al., 2023). Less famous is the bottom of the pack: last is Afghanistan (2.4/10); Zimbabwe is third last (3/10). In the middle are countries like China (5.5) and Peru (5.5).

This seems a credible ordering, with safe, developed countries at the top and unsafe, poor ones at the bottom. It suggests that some general global comparison is being made. We should be worried if we saw instead, for example, that Afghans said they were as satisfied as the Finns; that would indicate Afghans were using their scales in a very different way. We can supplement this by noting that the authors of the World Happiness Report argue that half a dozen factors can explain a large part of the differences in the scores (GDP, social support, life expectancy, freedom to make choices, generosity, and perceptions of corruption – see Figure 11); further details are not important for our purposes (Helliwell et al., 2023).

We can make three observations. First, it indicates (*pace* vignettes) that assuming there are inter-country differences is *prima facie* inelegant and unmotivated, when we

[12] Cf. Wittgenstein's (1953) 'language game' argument that words have meaning in virtue of having a common use.

look at the ordering; if there is a general pattern, we need a further explanation as to why some given country bucks the trend. Second, it shows that differences, if they exist, are not large. The Angelini et al. (2014) vignette study could be taken as an upper bound. The largest change there was Denmark, which moved down about 0.5 points on a 5-point scale, i.e., about 1 point on a 1-10 scale. Even with that drop, the Danes would still be in the pack of happy, highly developed European states: this is a modest revision of who’s happy. Third, that there is a plausible ordering in general does not prove country’s scales are exactly comparable.

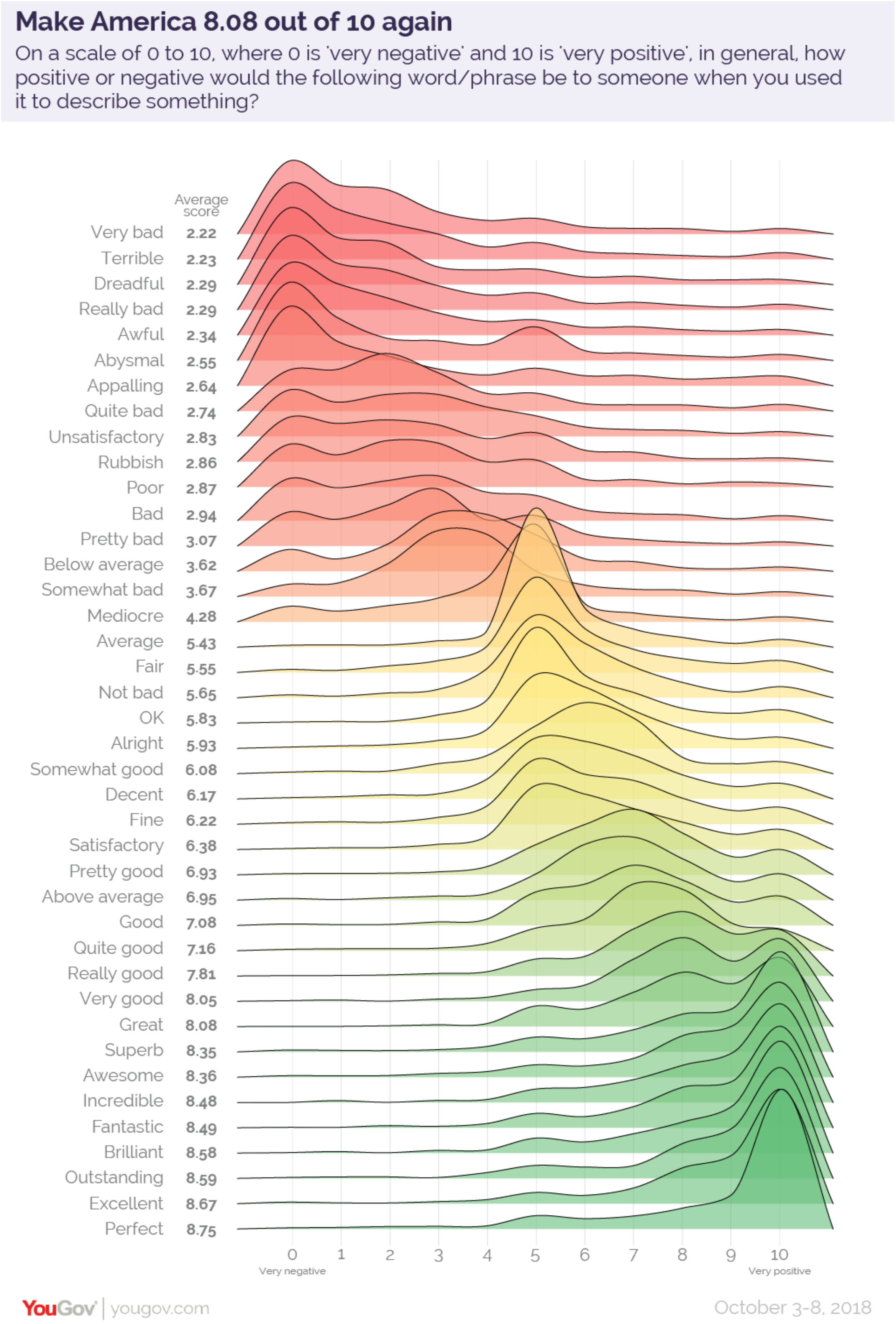


Figure 8. A frequency distribution of numerical scores attached to verbal labels. Figure from YouGov (2018)

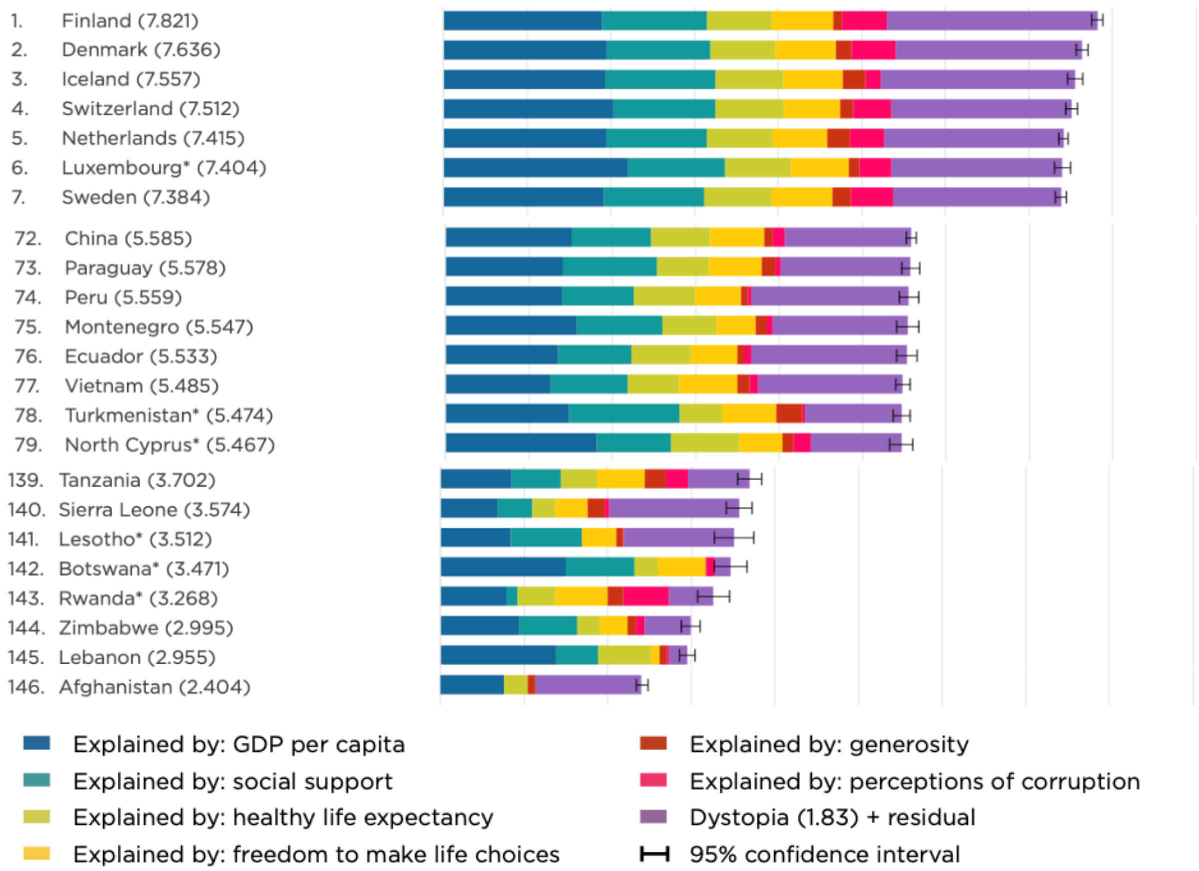


Figure 9. Life satisfaction scores for different countries including how between-country variation is explain various factors. Figure from World Happiness Report (Helliwell et al., 2023)

Let’s take stock of this section, then conclude more generally. The evidence for non-comparability came from the existing vignettes, which I suggested lie on shaky foundations. Even taking these at face value would not seem to radically threaten comparability. I suggested some alternative assumptions we might draw on and showed how these pointed *in favour of* interpersonal comparability. These assumptions do not allow us to conclude subjective scales are exactly interpersonally comparable. I cannot think of any tests that would decisively determine this; I hope further work can identify them.

7. Concluding remarks

Suppose we started with a healthy dose of scepticism about interpreting data from subjective scales as cardinal; after all, it may seem too good to be true that we can have quantitative, comparable measures of feelings by taking people’s answers literally. My conclusion is that serious scepticism is far harder to sustain that we might have expected.

I started with the observation that subjective scales can be closer or further from cardinality, and what ultimately matters is not whether are perfectly cardinal, but how far they are from it. I argued it is rational for respondents to interpret surveys as cardinal and showed how they could achieve this. I then argued that some uncertainty about scale use would not justify rejecting cardinality: we need either implausibly high uncertainty (to get to ordinalism), or specific beliefs about the type, sign, and magnitude of bias (to get to a specific version of quasi-cardinalism). Reviewing the evidence, it is difficult to find a strong case for any type of quasi-cardinalism; at most, deviations from cardinality are, practically speaking, small. Hence, until and unless new evidence comes to light, it does seem reasonable for researchers to treat subjective scales as cardinal. Or, to put the same thing differently, it seems unreasonable to reject cardinality. If you and I say we are ‘7/10 happy’, the safest guess is that we are about as happy as each other.

Presumably, some will remain sceptical. Maybe new evidence will vindicate this scepticism. What should people do if they do not believe that, practically speaking, subjective scales are cardinal? My final plea is not to give up on feelings data altogether. So long as we reject ordinalism (which I argued we should) then we are quasi-cardinalists, which means we have beliefs about how bent and oddly lengthed our ‘measuring sticks’ are. Hence it is possible, in principle, for quasi-cardinalists to apply the appropriate size and type of correction, whatever they believe it is, to yield answers that are cardinally comparable. Suppose we thought our reporting function was logarithmic; we simply need to apply a mathematical transform to get self-reported numbers on a linear, interval scale (Kristoffersen, 2011; Ng, 1997). This sort of adjustment may be onerous in practice, but it means that concerns about differences in how people report their happiness is not a terminal barrier to understanding how they could be happier.

Bibliography

Alexandrova, A., & Haybron, D. M. (2016). Is Construct Validation Valid? *Philosophy of Science*, *83*(5), 1098–1109. https://doi.org/10.1086/687941

Angelini, V., Cavapozzi, D., Corazzini, L., & 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. https://doi.org/10.1111/obes.12039

Bertrand, M., & Mullainathan, S. (2001). Do People Mean What They Say? Implications for Subjective Survey Data. *American Economic Review*, *91*(2), 67–72. https://doi.org/10.1257/AER.91.2.67

Bond, T., & Lang, K. (2018). *The Sad Truth About Happiness Scales: Empirical Results*. https://doi.org/10.3386/w24853

Bond, T. N., & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, *127*(4), 1629–1640. https://doi.org/10.1086/701679

Bonikowska, A., Helliwell, J. F., Hou, F., & Schellenberg, G. (2014). An Assessment of Life Satisfaction Responses on Recent Statistics Canada Surveys. *Social Indicators Research*, *118*(2), 617–643. https://doi.org/10.1007/S11205-013-0437-1

Bronsteen, J., Buccafusco, C. J., & Masur, J. S. (2012). Well-Being Analysis vs. Cost-Benefit Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1989202

Broome, J. (2004). *Weighing Lives*. Oxford University Press. https://doi.org/10.1093/019924376X.001.0001

Clark, A. E., D'Ambrosio, C., & Ghislandi, S. (2016). Adaptation to poverty in long-run panel data. *Review of Economics and Statistics*, *98*(3), 591–600. https://doi.org/10.1162/REST_a_00544

Clark, A. E., 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), F243.

Clark, A. E., Powdthavee, N., Flèche, S., Layard, R., & Ward, G. (2018). *The origins of happiness : the science of well-being over the life course*.

Crisp, R. (2006). Hedonism reconsidered. *Philosophy and Phenomenological Research*, *73*(3), 619–645.

Crisp, R. (2008). Well-being. *Stanford Encyclopedia of Philosophy*.

Diener, E., Inglehart, R., & Tay, L. (2013). Theory and Validity of Life Satisfaction Scales. *Social Indicators Research*, *112*(3), 497–527. https://doi.org/10.1007/s11205-012-0076-y

Diener, E., Lucas, R. E., & Oishi, S. (2018). Advances and Open Questions in the Science of Subjective Well-Being. *Collabra: Psychology*, *4*(1), 15. https://doi.org/10.1525/collabra.115

Dolan, P., Layard, R., & Metcalfe, R. (2011). Measuring Subjective Wellbeing for Public Policy: Recommendations on Measures. *CEP Special Papers*.

Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being.

Journal of Economic Psychology, *29*(1), 94–122. https://doi.org/10.1016/j.joep.2007.09.001

Dolan, P., & White, M. P. (2007). How Can Measures of Subjective Well-Being Be Used to Inform Public Policy? *Perspectives on Psychological Science*, *2*(1), 71–85. https://doi.org/10.1111/j.1745-6916.2007.00030.x

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. https://doi.org/10.1111/j.1468-0297.2004.00235.x

Ferrer-i-Carbonell, A., & Frijters, P. (2004). How important is methodology for the estimates of the determinants of happiness? *The Economic Journal*. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0297.2004.00235.x/full

Fleurbaey, M., & Blanchet, D. (2013). *Beyond GDP: Measuring Welfare and Assessing Sustainability*. https://doi.org/10.1093/ACPROF:OSO/9780199767199.001.0001

Frijters, P. (1999). *Explorations of welfare and well-being*. Thela Thesis Amsterdam.

Frijters, Paul., & Krekel, Christian. (2021). *A Handbook for Wellbeing Policy-Making: History, Theory, Measurement, Implementation, and Examples.* OUP.

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451. https://doi.org/10.1038/075450a0

Gómez-Emilsson, A. (2019). *Logarithmic Scales of Pleasure and Pain: Rating, Ranking, and Comparing Peak Experiences Suggest the Existence of Long Tails for Bliss and Suffering - EA Forum*. https://qualiacomputing.com/2019/08/10/logarithmic-scales-of-pleasure-and-pain-rating-ranking-and-comparing-peak-experiences-suggest-the-existence-of-long-tails-for-bliss-and-suffering/

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Hausman, D. M. (1995). The impossibility of interpersonal utility comparisons. *Mind*, *104*(415), 473–490.

Haybron, D. M. (2016). Mental State Approaches to Well-Being. In M. D. Adler & M. Fleurbaey (Eds.), *The Oxford Handbook of Well-Being and Public Policy* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199325818.013.11

Helliwell, J. F., Layard, R., Sachs, J. D., Neve, J.-E. De, Aknin, L. B., & Wang, S. (2023). *World Happiness Report 2023*. https://worldhappiness.report/ed/2023/

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., Rosenfield, A., Gandhi, L., & Blaser, T. (2016, October). *Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making*. Harvard Business Review. https://hbr.org/2016/10/noise

Kaiser, C. (2022). Using memories to assess the intrapersonal comparability of wellbeing reports. *Journal of Economic Behavior & Organization*, *193*, 410–442. https://doi.org/10.1016/J.JEBO.2021.11.009

Kaiser, C., & Oswald, A. J. (2022a). The scientific value of numerical measures of human feelings. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(42). https://doi.org/10.1073/PNAS.2210412119/-/DCSUPPLEMENTAL

Kaiser, C., & Oswald, A. J. (2022b). The scientific value of numerical measures of human feelings. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(42), e2210412119. https://doi.org/10.1073/PNAS.2210412119/SUPPL_FILE/PNAS.2210412119.SAPP.PDF

Kaiser, C., & Vendrik, M. (2020). *How threatening are transformations of happiness scales…* (2020–19). https://www.inet.ox.ac.uk/publications/no-2020-19-how-threatening-are-transformations-of-happiness-scales-to-subjective-wellbeing-research/

Kristoffersen, I. (2010). The Metrics of Subjective Wellbeing: Cardinality, Neutrality and Additivity*. *Economic Record*, *86*(272), 98–123. https://doi.org/10.1111/J.1475-4932.2009.00598.X

Kristoffersen, I. (2011). The Subjective Wellbeing Scale: How Reasonable is the Cardinality Assumption? In *Economics Discussion / Working Papers* (11-15). The University of Western Australia, Department of Economics.

Kristoffersen, I. (2017). The Metrics of Subjective Wellbeing Data: An Empirical Evaluation of the Ordinal and Cardinal Comparability of Life Satisfaction Scores. *Social Indicators Research*, *130*(2), 845–865. https://doi.org/10.1007/s11205-015-1200-6

Krueger, A., & Schkade, D. (2007). *The Reliability of Subjective Well-Being Measures*. https://doi.org/10.3386/w13027

Layard, R. (2003). Happiness: has social science a clue? Lecture 1: what is happiness? Are we getting happier? *Lionel Robbins Memorial Lecture Series*.

Layard, R. (2005). *Happiness : lessons from a new science*. Allen Lane.

Luhmann, M., Hofmann, W., Eid, M., & Lucas, R. E. (2012). Subjective well-being and adaptation to life events: A meta-analysis. *Journal of Personality and Social Psychology*, *102*(3), 592–615. https://doi.org/10.1037/a0025948

Montgomery, M. (2022). Reversing the gender gap in happiness. *Journal of Economic Behavior and Organization*, *196*, 65–78. https://doi.org/10.1016/J.JEBO.2022.01.006

Ng, Y. (1997). A case for happiness, cardinalism, and interpersonal comparability. *The Economic Journal*, *107*(445), 1848–1858.

Ng, Y.-K. (1995). Towards welfare biology: Evolutionary economics of animal consciousness and suffering. *Biology and Philosophy*, *10*(3), 255–285. https://doi.org/10.1007/BF00852469

Ng, Y.-K. (2008). Happiness studies: Ways to improve comparability and some public policy implications. *Economic Record*, *84*(265), 253–266. https://doi.org/10.1111/j.1475-4932.2008.00466.x

OECD. (2013). *Guidelines on Measuring Subjective Well-being*. OECD Publishing. https://doi.org/10.1787/9789264191655-en

Oswald, A. J. (2008). On the curvature of the reporting function from objective reality to subjective feelings. *Economics Letters*, *100*(3), 369–372. https://doi.org/10.1016/j.econlet.2008.02.032

Peart, S. J., & Levy, D. M. (2005). From Cardinal to Ordinal Utility Theory. Darwin and Differential Capacity for Happiness. *American Journal of Economics and Sociology*, *64*(3), 851–879. https://doi.org/10.1111/j.1536-7150.2005.00394.x

Portugal, R. D., & Svaiter, B. F. (2011). Weber-Fechner law and the optimality of the logarithmic scale. *Minds and Machines*, *21*(1), 73–81. https://doi.org/10.1007/s11023-010-9221-z

Prati, A., & Senik, C. (2020). Feeling good or feeling better? In *Working Papers* (13166; DP). HAL.

Robbins, L. (1932). *An essay on the nature and significance of economic science,*. Macmillan. http://www.worldcat.org/title/essay-on-the-nature-significance-of-economic-science/oclc/838285

Russell, B. (1905). ON DENOTING. *Mind*, *XIV*(4), 479–493. https://doi.org/10.1093/MIND/XIV.4.479

Salomon, J. A., Tandon, A., & Murray, C. J. L. (2004). Comparability of self rated health: Cross sectional multi-country survey using anchoring vignettes. *British Medical Journal*, *328*(7434), 258–261. https://doi.org/10.1136/bmj.37963.691632.44

Samuelsson, C., Dupret, S., Plant, M., & Kaiser, C. (2023). *Can we trust wellbeing surveys? A pilot study of comparability, linearity, and neutrality*. https://www.happierlivesinstitute.org/report/can-we-trust-wellbeing-surveys-a-pilot-study-of-comparability-linearity-and-neutrality/

Schelling, T. C. ,. (1960). *The strategy of conflict*. Harvard University Press.

Schröder, C., & Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, *92*, 337–358. https://doi.org/10.1016/J.EUROECOREV.2016.12.011

Schwarz, N. (1995). What Respondents Learn from Questionnaires: The Survey Interview and the Logic of Conversation. *International Statistical Review / Revue Internationale de Statistique*, *63*(2), 153. https://doi.org/10.2307/1403610

Stiglitz, K., Sen, A., & Fitoussi, J. (2009). *Report by the commission on the measurement of economic performance and social progress*.

Stone, A., & Krueger, A. (2018). Understanding subjective well-being. In J. E. Stiglitz, J.-P. Fitoussi, & M. Durand (Eds.), *For Good Measure: Advancing Research on Well-being Metrics Beyond GDP. OECD*. OECD. https://doi.org/10.1787/9789264307278-en

van Praag, B. M. S. (1991). Ordinal and cardinal utility. An integration of the two dimensions of the welfare concept. *Journal of Econometrics*, *50*(1–2), 69–89. https://doi.org/10.1016/0304-4076(91)90090-Z

Wittgenstein, L. (1953). *Philosophical investigations* (G. Anscombe & R. Rhees, Eds.). Blackwell.

Wodak, D. (2019). What If Well-Being Measurements Are Non-Linear? *Australasian Journal of Philosophy*, *97*(1), 29–45. https://doi.org/10.1080/00048402.2018.1454483

YouGov. (2018, November). *How good is "good"?* https://today.yougov.com/topics/lifestyle/articles-reports/2018/10/11/how-good-good