A Happy Possibility About Happiness (And Other Subjective) Scales: An Investigation and Tentative Defence of the Cardinality Thesis

Michael Plant[1, 2]

[1] Happier Lives Institute
[2] Wellbeing Research Centre, University of Oxford

Working paper | November 2020

Happier Lives Institute

A Happy Possibility About Happiness (And Other Subjective) Scales: An Investigation and Tentative Defence of the Cardinality Thesis[1,2]

Michael Plant[3]

This version: November 2020

0. Abstract

There are long-standing doubts about whether data from subjective scales—for instance, self-reports of happiness—are cardinally comparable. It is unclear how to assess whether these doubts are justified without first addressing two unresolved theoretical questions: how do people interpret subjective scales? Which assumptions are required for cardinal comparability? This paper offers answers to both. It proposes an explanation for scale interpretation derived from philosophy of language and game theory. In short: conversation is a cooperative endeavour governed by various maxims (Grice 1989); because subjective scales are vague and individuals want to make themselves understood, scale interpretation is a search for a *focal point* (Schelling 1960). A specific focal point it hypothesised; if this hypothesis is correct, subjective data will be cardinally comparable. Four individually necessary and jointly sufficient conditions for cardinal comparability are specified. The paper then argues this hypothesis can be empirically be tested, makes an initial attempt to do so using subjective well-being data, and concludes it is supported. Numerous areas for further research are identified including, at the end of the paper, how certain tests could be used to 'correct' subjective data if they are not cardinal.

1. Introduction

Individuals regularly use *subjective scales*, numerical ratings of subjective phenomena. For instance, they rate, on a (say) 0-10 scale, their happiness, life satisfaction, job satisfaction, health, pain, the sexual attractiveness of others, the movies they watch, the products they buy, and so on.

There are long-standing doubts over whether answers from subjective scales are *cardinally* comparable: does a one-point change, on a given scale, represent the same size change in subjective experience for different people and at different times (on average).[4] For instance, if I say my

[1] I would like to thank Patrick Kaczmarek, Caspar Kaiser, Clare Donaldson, Joel McGuire, and Andrew Oswald for reading various drafts of this document and providing comments. I am also grateful to the audiences of the Well-being Seminar of Centre for Economic Performance at the London School of Economics and of the Wellbeing Research Centre, Oxford. This research was made possible by funding from the Wellbeing Research Centre, Oxford and the Happier Lives Institute. This essay began as a chapter in my D. Phil. thesis at Oxford which was supervised by Peter Singer and Hilary Greaves.

[2] This working paper has been submitted for publication at *Economics and Philosophy*.

[3] Wellbeing Research Centre, University of Oxford; Happier Lives Institute

[4] This is a specification for what I call 'minimal cardinality' – cardinality of changes. 'Maximal cardinality' involves, in additional, cardinality of levels. I return to this later.

happiness has gone from a 4 to a 5 out of 10, and you say your happiness has gone from a 3 to 4, can we conclude we each had the same size increase in happiness?[5]

One concern is that subjective phenomena are not perceived in units of intensity—they lack *phenomenal cardinality*. If, for instance, pain is not experienced in units, then it would be confused to say, "this hurts me *as much as* it hurts you". While I discuss it later, I take it this is not the main concern, which is whether groups of people *interpret* subjective scales in consistently different ways such that numerical *reports* of internal states are not cardinally comparable between people ('interpersonally') and/or over time ('intertemporally').

If subjective scales turned out to be merely *ordinal*—that is, the numbers represent a ranking but contain no information on the relative magnitudes of differences—that would be a serious inconvenience. Ordinal data do not allow us to express unit changes; this means we could not, for example, say how much different outcomes increase total happiness. Ordinal scales cannot be meaningfully averaged; therefore it would be incoherent to say things such as "the average life satisfaction of Germany is 8.2/10, 1 point higher than France" or "customer satisfaction is up by 20% this month", things we regularly do say.

Ferrer-i-Carbonell and Frijters (2004) observe that beliefs about the *Cardinality Thesis* (*CT*)—the assumption of cardinal comparability for subjective scales—divide on disciplinary lines: speaking broadly, psychologists tend to find it unproblematic, whereas economists consider it highly suspicious. In economics, this suspicion can be traced back to Lionel Robbins in the early 20th century who thought cardinal measures of people's feelings were not only impossible but unnecessary for the discipline (Layard, 2003; Hausman, 1995). Robbins supposed the aim of economics was to determine how people act under conditions of scarcity, all that is needed to do that is the assumption individuals have a stable ordering of preferences. A prominent and recent example economists' scepticism is Bond and Lang (2018), who argue in a descriptively named paper—*The Sad Truth About Happiness Scales*—that various 'key' results in the so-called 'happiness literature' rely on *CT* and these results can be reversed if one makes different assumptions about how individuals report their happiness. I briefly discuss this paper later and argue it relies on assumptions there are good reasons to reject.

Although the nature of subjective data is both fundamental and disagreed upon, there seems to have been little explicit discussion of the cardinality thesis in the literature, excepting Ferrer-i-Carbonell and Frijters (2004), Kristoffersen (2011, 2017), Adler (2013), Ng (1997, 2008), and Van De Deijl (2017). This paucity of discussion is perhaps, on reflection, unsurprising given these issues inhabit an apparent academic no-man's land. The topic is theoretical by the standards of the empirical researchers who collect and analyse subjective data—mostly economists and psychologists. We might expect it, then it would be the nature terrain of philosophers. However, by the standards of philosophy, this question is highly empirical. While philosophers have discussed social scientists' attempts to measure subjective well-being—see e.g. Alexandrova (2012, 2016), Angner (2013), Haybron (2008, 2016)—I am unaware of any discussion by philosophers of the cardinality of these measures.

[5] An entirely separate question is whether, given we have had the same increase in happiness, our increases have the same value. This essay is solely concerned with descriptive questions about quantities (or intensities) of subjective states and not about the value of distributions of these states.

Kristoffersen (2011) notes that researchers often treat their data as ordinal or cardinal (and so use different statistical tests) without clearly articulating their reasons. Stone and Krueger (2018, p189) in an article summarising the recent research into subjective well-being (SWB)—ratings of happiness, life satisfaction, and purpose—write that:

> one of the most important issues inadequately addressed by current [SWB] research is that of systematic differences in question interpretation and response styles between population groups. Is there conclusive evidence that this is a problem? And, if so, are there ways to adjust for it? Information is needed about which types of group comparisons are affected, about the magnitude of the problem, and about the psychological mechanisms underlying these systematic differences.

Why hasn't this issue been addressed? Perhaps because researchers have held that this problem can often simply be ignored. Discussions of *CT* often start and end with the *Washing Out* argument: see e.g. Dolan and White (2007) and Bronsteen, Buccafusco and Masur (2012). The thought is that variations in individuals' interpretations of subjective scales, their capacities for subjective experience, etc. are random. Therefore, so long as the surveyed populations are randomly constructed and large enough, any differences will statistically 'wash out' as noise and can be ignored. Researchers might then confidently suppose that, in a given country, any differences will be random, hence this issue can be ignored for matters of national and sub-national policy.

However, the occurrence of washing out is necessary but not sufficient for the cardinality thesis to hold; cardinal comparability also requires that individuals must, on average, use a linear reporting function, where each 1-unit change on the scale represents the same magnitude change. If reporting is non-linear, e.g. each 1-unit reported increase represents twice as big an increase as the one before, subjective scales will not be cardinal whether or not deviations from this reporting pattern are random.

Even if we set aside this concern, we cannot confidently assume washing out will always occur: we might anticipate that some group differences, for instance those due to culture, will lead to systematic differences in interpretation, and/or that individuals will change their interpretations over time. Hence, it is necessary to understand and then evaluate the cardinality thesis.

Progress on this issue is currently hindered—if not halted—by a lack of theory. Quoting Stone and Krueger (2018, p175) again:

> In order to have more concrete ideas about the extent to which this may be a problem, we should have a better idea of why such differences might exist in the first place, and have some theoretical justification for a concern with systematic differences in how subjective well-being questions are interpreted and answered.

For clarity, these seem to be the key theoretical questions here: (1) how (and why) do people interpret subjective scales? (2) does this interpretation generate cardinally comparable answers—and, if not, what does it lead to? (3) exactly which assumptions are needed for cardinal comparability? (4) can these assumptions be tested and, if so, how? Unless and until we answer these questions we are unable to address the 'downstream' practical concerns, namely: (5) is it reasonable to assume cardinal comparability on the basis of the current evidence? (6) can we adjust

subjective data if they are not cardinal and, if so, how? To the best of my knowledge, proposed answers to these questions are mostly limited or non-existent.

To make progress, I do the following in this paper. First, I propose a theoretical explanation for how individuals interpret subjective scales and explain how, if this hypothesis is true, we would expect subjective data of all types to be cardinal. Second, I decompose cardinal comparability into four individually necessary and jointly sufficient conditions. Third, I note these conditions are empirically testable, even though they concern subjective states. Fourth, I survey the small existing evidence for each condition; I find this broadly supports the Cardinality Thesis and gives no strong reason to reject it. Fifth, I set out some testable predictions of the theory and explain how, if subjective data turned out not to be cardinal, we could adjust the data by using such tests. Areas for further research are identified throughout.

Before we proceed, four further comments on the nature and scope of this problem.

First, it is quite common to express scepticism specifically about the cardinal comparability of subjective well-being; Davies (2015) for a book length critique. See OECD (2013) for an overview of measures of SWB. However, if the concern is about differential interpretation of survey questions, then this will, presumably, apply to numerical ratings of subjective phenomenon *in general*.. Hence, someone who objects to treating happiness surveys, but not other subjective data—health, hotel quality, employee satisfaction, etc.—as cardinally comparable needs to justify this differential treatment.

Second, what is puzzling about the professional scepticism of some academics is that it does not seem to be shared by the public. Non-response rates for happiness and life satisfaction questions, which are often asked on a 0-10 scale, are very low and about the same as those for educational attainment and marital status (OECD, 2013). Suppose respondents sincerely believed these scales were ordinal—either because the subjective states lacked phenomenal cardinality, or the states were cardinal but the scales ordinal. How would they answer? If you believe the scale is ordinal, the numbers contain no information about magnitudes, so your choice of number is arbitrary. If the choice seemed arbitrary to respondent, we would expect both that many not to answer and for the answers that were given to show no pattern. As it is, not only do people find it no more difficult to rate their subjective experiences on an apparently cardinal scale than to know whether they are married or not, but we also do see all sorts of patterns: greater subjective well-being is associated with higher income, being partnered, and so on (Dolan, Peasgood and White, 2008).

Third, the question of whether reports of subjective data are cardinal is conceptually distinct from that of whether *well-being* (what ultimately makes life go well for us) is cardinal; we are only concerned with the former here. For discussions of the cardinality of well-being, see e.g. Broome (2004, ch. 5), Hausman (1995). One might unproblematically maintain that well-being consists in a ranking of preferences (economists usually call this conception of well-being 'utility'), and thus that well-being is only ordinally comparable, whilst nevertheless maintaining that reports of happiness (defined as a psychological state consisting in a positive balance of pleasure over displeasure), movie quality, etc. are cardinally comparable.

Fourth, for ease of exposition, I focus on subjective well-being data but, as noted, I expect this analysis to apply to subjective scales in general.[6] Whether it does is a topic for further research.

Before we turn to the Schelling point story about scale use, section two provides a brief primer on units of measurement.

2. Units of measurement

Units of measurement are typically grouped according to their quantitative properties. The standard four-fold division is as follows (Edwards, 1964).

Nominal scales are used for labelling variables without quantitative information, for instance, gender, or hair colour.

Ordinal scales contain variables which have a relative magnitude, such as the order that runners finish in a race—1st, 2nd, 3rd, etc.—but lack information about the relative difference between those magnitudes.

Interval scales contain variables which can not only be ordered but where the differences between measurements on the scale are *equal-interval*—this is the condition for *cardinality*. Celsius temperature is the classic example—the difference between each one degree of temperature is the same in terms of the change in thermal motion. What interval scales lack is a non-arbitrary zero-point on the scale—a location where there is no underlying quantity of what the scale measures: there is still thermal motion at zero degrees Celsius.

Ratio scales have the same properties of interval scales with the additional of having a non-arbitrary zero point. Examples of this include mass, time, distance, and temperature when measured in Kelvins. Ratios are meaningful, e.g. 10 minutes is twice as long as 5 minutes.

Here we are just concerned with cardinality, the property which is sufficient for an interval scale but not for a ratio scale.

What does it mean for to say two scales are 'cardinally comparable'? There is some ambiguity here which I have not seen stated. Let's say *minimal cardinality* obtains where a one-point change, on a given pair of scales, represents the same size change in the underlying property. If, as it common with subjective scales, there is a fixed number of units on the scale, e.g. an 11-unit 0-10 scale, that means each scale has the same *range*. However, this still allows the scales cover different *levels*. An example of this type is displayed in figure 1: A and B, which might represent (say) two countries, have SWB scales with the same range—6 units of actual SWB—but span different levels.

[6] For a substantial discussion of SWB and its measurement, see OECD (2013).

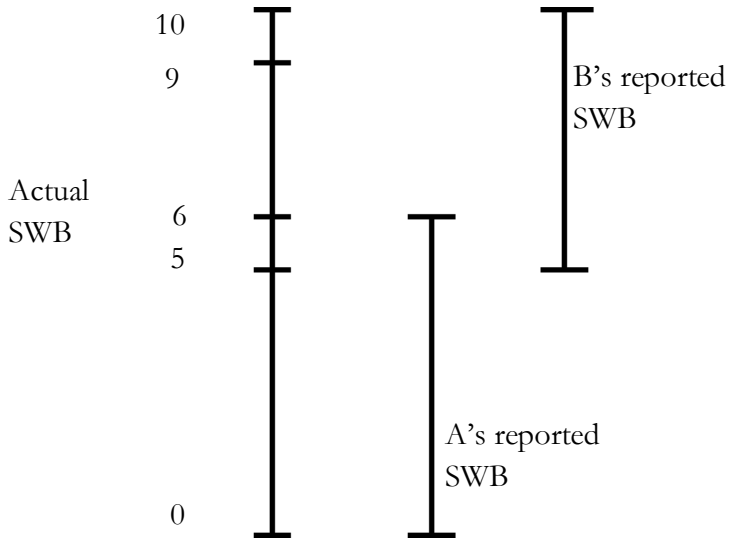

Figure 1. Minimal cardinally comparable scales with different levels

In contrast, let's say a pair of scales has *maximal cardinality*, if, in addition to having minimal cardinality, the levels are the same, i.e. 7/10 represents the same quantity on A's and B's scales.

While only minimal cardinality is needed for unit changes, maximal cardinality is needed for averages, e.g. to claim, "the English are happier than the Americans". It's unclear which sense of cardinality comparability is ordinarily meant, and additional assumptions are needed for the latter. As it happens, the hypothesis I propose in the next section leads to maximal cardinality; hence, from here onwards, when I refer to 'cardinality' without qualification I mean the maximal kind.

Even if the scales have cardinality, a further challenge is whether they are of ratio quality. I do not argue for or assume a ratio scale.

3. A Schelling-point story about subjective scale use

Much of the existing discussion of cardinality is quite mathematically complex (e.g. Bond and Lang (2018), Ferrer-i-Carbonell and Frijters (2004)). However, where the concern is how people report their inner states, what we need—and are currently lacking—is an explanation of how people interpret subjective scales. I propose an explanation after outlining the problems with scale interpretation.

Suppose I ask you "How happy are you?" and give you a 0 – 10 scale. To answer this question, you need to conscious or unconsciously fill in some details about the scale. The same will apply to other subjective scales but I will just use happiness for simplicity.

One is: what do the end points of the scale refer to? How happy(/unhappy) do you have to be to be a 10(/0) out of 10? There's no logical limit to how happy you could be—happiness is unbounded—but I have given you an apparently bounded scale—we return to this in a moment.

I might specify 10/10 to mean (say) 'very happy' and 0/10 as 'very unhappy', but you still need to decide what 'very happy' and 'very unhappy' mean.[7]

The other, as noted, is your reporting function: the relationship between your experienced and reported happiness. You might use a linear reporting function (every one-unit change on the self-reported scale represents the same change in happiness) but you might do something else.[8]

Once you've established how to use the scale, you then think of how happy you feel in terms of that scale. You've only been given a limited number of response categories, but happiness varies (nearly) continuously, so you pick the number closest to your internal score;[9] for example, you think you're 7.4/10, but you can only choose whole numbers, so you say you are 7/10.

How should we expect people will interpret subjective scales? An important observation made by Grice (1989), a philosopher of language, is that conversations are often cooperative endeavours, where speakers and listeners rely on each other to think and act in certain ways in order to be understood. 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.

We can expect individuals, when surveyed, will try to accurately communicate their inner states (Schwarz, 1995). There are, as noted, many ways to interpret the scale and individuals cannot communicate with each other on which one to use. Hence, to make their answers accurate, individuals will need to anticipate how *other* people will interpret the scale and then give their answers in terms of the scale they expect others to use. If I am confident my 6/10 represents a different amount of happiness from everyone else's 6/10, I am being uncooperative and can expect to be misunderstood.

Turning from philosophy to economics, in game theoretic terms individuals are seeking a *Schelling point*, or 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 Noble-prize winning economist after whom the term is named, asked a group of students this question and found that the most common answer was noon (at the information booth) at Grand Central Station. One could meet at any time and place, but certain answers are, for whatever reason, more salient and more likely to lead to successful coordination.

[7] Technically, one only needs to decide on the meanings of two points on the scale and extrapolate from there. As the number of points on the scales is fixed, deciding on the meaning of two points does generate meanings for the ends of the scale.

[8] In fact, you'll also need to decide on at least two non-numerical issues: what does 'happy' mean? What time period is being asked about (right now? In general? Recently? Etc.)? While such questions are relevant for comparing answers, they are beyond the scope of this essay. For some discussion, see Benjamin *et al.* (2020). Here, I concentrate only on the numerical issues of scale interpretation, i.e. I assume that individuals are answering the same qualitative question and then ask, further, whether we can conclude their answers are cardinally comparable.

[9] Note, however, Edgeworth's (1881) claim that there is a minimum 'just perceivable increment' in sensation.

What is the Schelling point for subjective scales? I hypothesise that, when respondents as asked to rate the value of a subjective state and given a limited number of options, e.g. a 0 – 10 scale, they will interpret the top and bottom of scales as, respectively, the lowest and highest values the state takes *in practice* (more on this in a moment), and then interpret the scale as linear, so that each point on the scale represents the same change in magnitude.

Why expect this?

Starting with the reporting function, linearity is the natural choice as it makes the scale cardinal; as Ferrer-i-Carbonell and Frijters (2004) note, cardinal scales are the type people are familiar with, e.g. what we use for measuring weight, height, income, etc. and hence the default option we would expect others to use. I discuss two alternative (but implausible) reporting functions later.

A linear reporting function is necessary but not sufficient for (minimal or maximal) cardinality comparability across time and between people. You and I might treat our own scales as equal-interval, but if yours is twice as 'long' as mine, then a one-point change for you is equivalent to a two-point change for me (and so minimal cardinality will not obtain).

Regarding endpoints, I might consider using 10/10 to mean the happiest it is logically possible to be. But there's no *logical* limit to happiness, just as there is no logical limit to how large numbers can be, so I cannot sensibly report my happiness on that scale. There is a *nomological* limit to happiness—a maximum within the laws of nature in this universe—but it is not obvious where it is, which makes it likewise impractical. Perhaps my current happiness is 5.00000001/10 on a scale where 10/10 is the nomological limit.

We do not want our scales to be too long *or* too short. If I use 10/10 to mean how happy I am on an average day, there's no way for me to convey how happy I am at my happiest. I could use 10/10 to mean the happiest I have ever been. But what if I expect to become happier tomorrow? If I say I'm 10/10 both days, I can't expect the person analysing the survey to know *I* mean two different things. The same concerns apply over time as they do across people. If you and I use our 10/10 to mean our personal maximums so far, but I know you have been happier than me, our scores will mean different things.

Hence, the 'goldilocks' solution here—not too short, not too long—seems to be to use the endpoints to represent the real limits, e.g. the happiest/unhappiest anyone is in the actual world. These are both the smallest *and* the largest magnitudes I and others would need if we all want to use the same length scale over time and so have cardinally comparable answers.

Of course, individuals will differ somewhat on what they think minimum and maximum in fact are. This may not matter, however, due to a 'wisdom of the crowds' effect. Galton (1907) famously observed that when a large group of people guessed the weight of a cow at a country fair, the median answer of the crowd was surprisingly accurate: within 1% of true answer. By extension, we can anticipate that, in aggregate, individuals' choices of subjective scale endpoints will approximate the true maximum and minimum.

Subjective scales are sometimes criticised for being a bounded measure of something unbounded ("why does happiness only go up to 10? *Surely* it's possible to be happier.") (Benjamin, 2020). We can now see this complaint rests on a misunderstanding. Subjective phenomena do not have

objectively measurable units. Hence, the numerical labels we attach to the ends of the scales are effectively arbitrary—the scales could run from 0 to 1,000, -50 to 50, etc.[10] If the endpoints cover the full range that is realistically possible, then the scales are not *realistically* bounded: while it is logically possible to be happier than 10/10, this is not practically possible, simply because we 'stretch' the scales to include whatever is actually possible.

We might wonder if people have a good idea of what the actual limits of happiness are. Further research should investigate this, but this assumption seems reasonable: we frequently communicate with each other about the highs, lows, and middles of our own lives and those of other people, in the latter case often through stories.

If the hypothesis about scale interpretation were true, the cardinality thesis will hold. A linear reporting function means each individual's scale at a given time is cardinal; if everyone uses the same length scale at all times and as each other, then a 1-unit reported change is the same for everyone everywhere—minimal cardinality; as the same levels are used for the end-points—the actual limits—we get to maximal cardinality.

As far as I know, the suggestion that subjective scale use be understood as a search for focal points has not been offered before. With it, we have an explanation of what individuals are doing when they answer subjective questions—they are not picking scales at random but are trying to make themselves understood.

While the focal point I have proposed—a linear reporting function and endpoints representing the real limits—seems the most intuitive if individuals want their answers to be understood, respondents might choose another.

An alternative hypothesis is for individuals to use their own personal maximum and minimum intensities to date for their endpoints; this could be motivated by respondents feeling they do not know what the collective limits are and instead relying on their own. If individuals do this and have different personal endpoints, the data will not be cardinally comparable.

This leaves us with two hypotheses. Before we discuss testing these empirically, let us break the cardinality thesis into its constituent parts.

4. Decomposing and assessing the cardinality thesis

From the discussion above, it may now be clear there are four conditions are individually necessary and jointly sufficient for *CT* holding *on average*. Let's spell these out.

> *C1 Phenomenal Cardinality*: the underlying subjective state is perceived in units of intensity.
>
> *C2 Linearity*: there is a linear relationship between the true and the reported subjective state.
>
> *C3 Intertemporality*: for each individual, the top and bottom of their self-reported scale represent the same magnitudes of the subjective phenomenon across time and this covers at least the full range of the phenomenon that is actually possible.

[10] At least, not practically. One approach, experimentally pioneered by Ng (1996) is to start by assuming Edgeworth's (1881) idea that all 'just perceived increments' in sensation are the same for everyone and then count up the number of those.

C4 Interpersonality: for different individuals, the top and bottom of their self-reported scales represent the same magnitudes of the subjective phenomenon at a time and this covers at least the full range of the phenomenon that is actually possible.

I now make few comments on the conditions and their evaluation.

What should we conclude about the cardinal comparability of subjective data if the conditions fail? This depends on which condition we are concerned with and how it fails. The first condition is fundamental: if the subjective phenomenon itself lacks cardinality, of course there could not be a cardinal measure of it; by analogy, C1 is about whether we can have a measuring stick at all. Continuing the analogy, C2 concerns whether the measuring sticks are bent, C3 is if the length of each of each stick changes over time, and C4 is whether different people have the same length sticks. Hence, if the first condition holds, but one or more of conditions two to four fail, that means the self-reported data are not cardinal to *some* extent and, therefore, the subsequent concern will be *how far* the data are from being cardinal (on average). It matters if our measuring sticks are slightly bent or very crooked. The cardinality thesis could fail to be exactly true, but nevertheless be approximately true. It's worth adding that, if we know how bent our sticks are, then we can straighten them.

CT requires stronger assumptions than are necessary for subjective scales to be cardinally comparable in certain circumstances. For instance, for comparisons between my happiness scores today and my happiness scores tomorrow, C4 isn't necessary. As such, *CT* might more fully be described the 'Raw, Universal Cardinality Thesis', the view the 'raw', unadjusted self-reports are 'universally' cardinally comparable, that is cardinally comparable across all times and individuals. Having noted this, I return to calling it just the 'Cardinality Thesis'.

Consistent with the Schelling point analysis, C3 and C4 do not specify exactly what range individuals must use, just so long as it covers at least the actual range and the same range is used across time and across people, respectively.

We might wonder if we can evaluate whether, or to what extent, the conditions hold. It is a necessary property of subjective states that they cannot be measured objectively. Therefore, we cannot empirically test for phenomenal cardinality, nor can we be certain if differences in *reported* happiness are due to (a) differences in experienced happiness, (b) differences in reporting behaviour or (c) or some combination of the two. Should we assume then, that no evidence or reasoning would have any bearing on this topic? Is assuming scale cardinality, in the end, a matter of faith?

This would be too fast. What we should do here is to evaluate hypotheses by using inference to the best explanation, the methodology philosophers of science have argued is the cornerstone of the scientific method (Boyd, 1980; Harman, 1965; Williamson, 2017). Although this may seem obvious, I stress it because I have been surprised to find, in discussion, that sceptics of *CT* seem to think that, because we cannot objectively measure subjective states, all hypotheses about subjective states are equiprobable. While it is always the case that more than one hypothesis will fit the facts, but that does not mean that all hypotheses are equally likely. Given our background

beliefs, including experience of our own mental states, certain pieces of evidence will raise or lower our credence in the various conditions.[11]

In the next section, I will survey the existing evidence that seems relevant to CT and explain what we should infer from. After that, I propose various predictions that follow from the Schelling point hypothesis; I explain how they could be empirically tested with further work and how the results of these tests will either confirm the hypothesis or, if they falsify it, indicate how subjective data could be adjusted to make it cardinally comparable.

5. Assessing the conditions

As there are several conditions that comprise CT, this assessment proceeds in several subsections.

5.1 Condition 1: Phenomenal Cardinality

As noted, if the subjective phenomenon at hand is not perceived in units of intensity, there could not a cardinal measure of it.

Regarding happiness, it is introspectively obvious we can assign magnitudes to the intensities of sensations. This is reflected in our language. As Ng (1997) points out, we do think it is coherent to make claims of the following type, "being thrown in a bath of sulphuric acid would feel at least twice as bad as stubbing my toe". Ng (2008) observes it is often hard to make these comparisons precisely, but I note that this poses no threat to there being, in reality, precise cardinal differences. If I try to guess how many times heavier Mount Everest is than Mont Blanc, the fact I do know precisely what the answer is does mean there is not, in reality, a precise difference in their relative weights.[12] If happiness were ordinal then all that could be said was that the sulphuric acid bath was *worse* than the toe-stubbing, not that it was worse by some *amount*.

One criticism of hedonism, the view well-being consists in happiness, where happiness refers to pleasurable experiences, is the *heterogeneity* objection: pleasurable experiences, such those from reading a book, falling in love, or eating ice-cream are so different that there is no single, common scale we can put them all on (Nussbaum, 2012; Heathwood, 2006). However, as Crisp (2006) straightforwardly counters, all these experiences do seem to share a common property, pleasurableness, or 'hedonic feeling tone', and this admits of degree, even if these experiences are nevertheless qualitatively different in some respect. As noted earlier, non-response rates for happiness and life satisfaction questionnaires are low, which indicates respondents have little difficult in putting their subjective experiences on a single scale (OECD, 2013). Perhaps all we can say to the sceptic who still disbelieves there are subjective units of intensity for happiness, is, as the saying goes, "if you have to ask, you'll never know".

It doesn't seem the feature of having units of felt intensity is peculiar to happiness or life satisfaction. Hence, to the critic who disbelieves condition 1 applies to happiness, the question is

[11] More technically, I am employing a Bayesian approach to epistemology. See Talbott (2016) and references therein. While a formal Bayesian analysis of this topic (where we specify our prior probabilities in certain beliefs and then say how evidence updates our posterior probabilities) is possible, I think it would add very little.

[12] In other words, the vagueness is epistemic, not metaphysical. I am grateful to Caspar Kaiser for suggesting this point.

whether it fails for other subjective states—such as tastiness, loudness, brightness, etc.—and, if so, what the relevant difference is.

5.2 Condition 2: Linearity

It is generally assumed the relationship between reported and underlying subjective states is linear, but only approximately (Kristoffersen, 2011). Why only approximately? As noted, individuals have only a limited number of response categories, which forces them to pick the response category closest to their subjective state. As such, for a given individual at a given time, reporting behaviour will follow a step-wise function as illustrated in figure 2. However, if we collect many reports, we can assume the relationship between reported and true SWB approximates a line on average: there will be about as many people who feel 6.8/10 and are pushed to report 7/10 as those who feel 7.2/10, and so on, hence the *average true* happiness score for those who *report* 7/10 is 7/10.[13]

There is at least some reason to think the reporting function might not be even approximately linear. I'll mention the two most plausible alternatives. After this, I argue linearity is supported by the evidence and clear up two confusions.

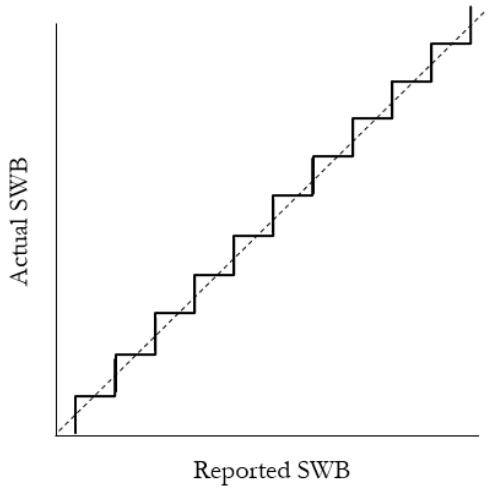

Fig 2. Linear stepwise relationship

One possibility is that the reporting relationship is logarithmic: that each one-unit reported increase of SWB represents a (say) doubling of the intensity of the true subjective state. This is shown in figure 3. This seems to be motivated by a *prima facie* similarity to the 'Weber-Fechner' law in psychophysics that the perceived intensity is proportional to the logarithm of the stimulus, thus physical forces, e.g. sound pressure, need to roughly triple for subjects to experience a one-unit change in subjective intensity, e.g. perceived loudness (Portugal and Svaiter, 2011).[14]

[13] In fact, and this seems not to have been noticed before, this will not quite be true. Supposing happiness is roughly normally distributed, then there will be slightly more people on the side of the cut-offs closer to the median score. For instance, while those with a true happiness of $7.5<x<8.5$ will report an 8/10, there will be more people in the range $7.5<x<8$ than $8<x<8.5$ so the average true happiness score of those who report 8 will be slightly lower than 8.

[14] I note the details of the law are unimportant for our purposes.

Ng (2008) suggests another possibility: the relationship between actual and measured SWB takes an arc-tangent form. This means the distance in actual SWB increases at the extremes of the scale. Thus, the actual difference between a self-reported 9 and 10 (and 1 and 2) is greater than the difference between a 3 and 4, a 5 and 6, etc. This is represented in figure 4. Ng's rationale is that, as there is no logical limit to happiness, the use of a linear representation which covered the full logical range would compress all changes in happiness from ordinary life event into a tiny range around the middle of the scale: e.g. becoming unemployed might take someone from 5/10 to 5.00002/10. The apparent advantage of the arc-tangent is that it makes the scale's middle comprehensive while still allowing very high happiness scores to be represented at the top of the range.

Three separate pieces of evidence indicate C2, the linearity condition is true, or nearly true. 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

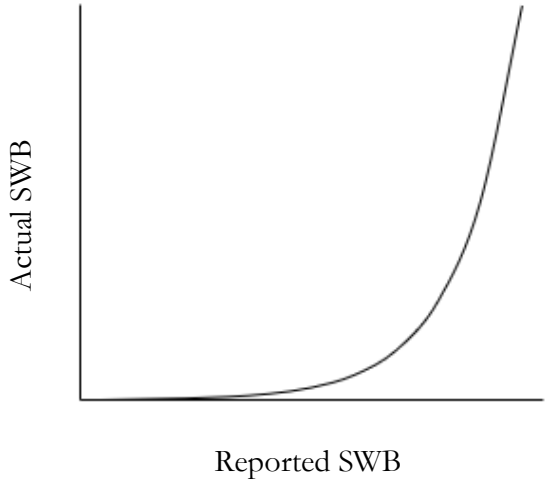

Figure 3. Logarithmic relationship

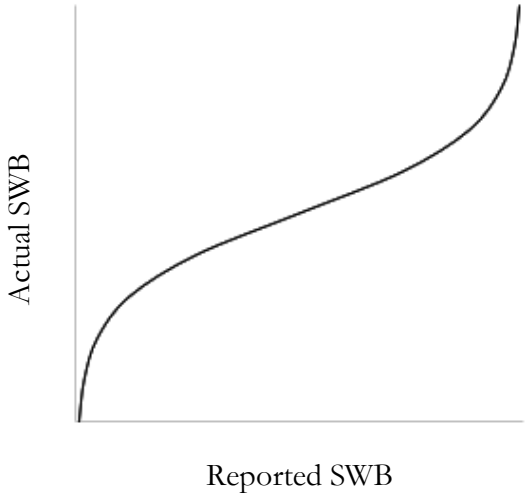

Figure 4. Arc-tangent relationship

height was effectively linear. This indicates individuals treat bounded scales of objectively measurable properties—in this case height—as linear.

We might wonder if people do the same for properties which are not objectively measurable, such as feelings. In a second study, van Praag (1991) gave subjects given ordered evaluative verbal labels ("very bad", "bad", "not bad", "not good", "good", "very good") and asked subjects to place this on a cardinal numerical scale labelled "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. Note that this finding leaves open whether each individual's scale had the same range and/or covered the same levels.

A third, compelling but somewhat indirect argument emerges from the *homoskedasticity* of errors in subjective reports. Krueger and Schkade (2008) conducted a test-retest of net affect—individuals are, in effect, asked how happy they are one day, asked again a week later, and the results are compared. Intuitively, what we would expect to find is that individuals' happiness varies by about the same week to week regardless of their level of happiness—we do not observe those who are very happy have wild swings in their moods whilst those who are unhappy have small changes, or *vice versa*. More technically, we would expect *homoskedasticity*, for the error in the regression model to be constant as the value of the predictor variable changes. Krueger and Schake find that the test-retest differences for *reported* levels of net affect are homoskedastic.[15] Given we *expect* homoskedasticity in actual SWB, if the relationship between reported and actual SWB were linear, then we would equally expect homoskedasticity in reported SWB, which was found. Given this is what is found, that suggests C2 is correct.

To be clear, homoskedasticity in the reported data does not prove C2 is the case—the finding is consistent with other reporting functions besides linearity. To illustrate, suppose a sceptic thinks the reporting relationship is exponential and, specifically, the magnitude changes doubles between each reported interval. To make this assumption consistent with the finding *reported* happiness is homoskedastic, what would need to be the case is that someone who *reported* their happiness as (say) 8/10 would need to have their *actual* happiness vary by about 32 times as much as someone who reports 3/10, which is not plausible.

Some evidence points against linearity. Lantz (2013) finds that the perceived distance between points on a 5-point Likert-style scale depends on how verbal anchors are used. For instance, using anchors only at the ends leads subjects to perceive a relatively larger distance between points near the ends of the scale than in the middle, i.e. going from 4 to 5 is larger than going from 3 to 4. However, subjects were not asked how big the perceived difference was, only if they perceived one at all; there being slight differences is consistent with the studies just mentioned and the reporting function being very close to linear. Further research could investigate the magnitudes of perceived differences.

I am unaware of any other non-anecdotal evidence pointing directly to logarithmic or arc-tangential reporting functions; for some anecdote-supported speculation on this topic, see Gómez-Emilsson (2019). Given the earlier Schelling point story, this is perhaps not a surprise. This issue for non-linear functions is that there is an infinity of specific non-linear functions to choose from, and one could not reasonably expect others to guess exactly which one you will use. For instance, for a logarithmic function, does each one-point increase represent a doubling of happiness, a tripling, a ten-fold increase, or something else? In terms of New York Schelling points, opting for non-linearity is analogous to expecting a stranger to meet you at your favourite coffee shop.

Two issues lurk. First, the Weber-Fechner law in psychophysics describes relationship between an *objectively* measure stimulus and a *reported subjective* intensity. Here, we are concerned about the relationship between *experienced* subjective intensity and *reported* subjective intensity. Hence, we have three relata—(1) changes in objective stimuli, (2) changes in reported experience; (3) changes in actual experience—and we are inquiring about their relationships.

[15] At p.18 they note "assumption of homoskedastic measurement error could be violated, but the deviation is probably slight".

In the psychophysics experiments, the normal interpretation is that when someone *reports* a 1-unit subjective increase in intensity—due to e.g. turning up the volume threefold on the speakers—they do, in fact, *experience* a 1-unit subjective increase in intensity. But we only observe the relationship between (1) and (2). To make the further inference the relationship between (1) and (3) is logarithmic, one must infer the relationship between (2) and (3) is linear. But C2 just is the claim there is a linear relationship between (2) and (3). Hence, it us puzzling to claim the Weber-Fechner law is evidence against linearity in the reporting function when the Weber-Fechner law assumes linearity in the reporting function to derive its conclusion, which is about the relationship between objective stimuli and subjective experience![16]

The second issue is where Bond and Lang's (2018) critique fits into this. Cutting through the mathematics, Bond and Lang make a hypothetical argument of the following general type: if one rejects the assumption that relationship between actual and reported SWB is linear, it is possible to reverse many of the key findings in the SWB literature. According to Kaiser and Vendrik (2020), the specific argument Bond and Lang use is as follows. As happiness is logically unbounded but individuals have only limited numbers of labels, reports in the top(/bottom) category could potentially be infinitely large(/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 categories below. Under these conditions, it is trivially easy to reverse any result which had assumed the scale is linear (i.e. had equal-interval cut-offs) by making suitably 'heroic' assumptions about the actual happiness levels of those in the top and bottom categories.

Bond and Lang do not, however, provide any evidence for their hypothetical argument. For the reasons already given, it is hard to believe individuals would use anything other than a linear reporting function if they wanted to be understood others.

5.3 Remarks on conditions 3 and 4

Conditions 3 and 4 would be met if individuals used the ends of the scales to refer to the actual maximum and minimum intensities of the subjective phenomenon over time. Why think individuals are willing and able to do that?

One piece of evidence supporting this is the worldwide distribution of subjective well-being scores. These follow an approximately normal distribution with responses in all the categories, as displayed in figure 5. Why does this suggest the reported maximum and minimum represent the real maximum and minimum? Recalling the earlier discussion, if people were using the logical or nomological limits, the self-reports would be compressed into a very small part of the scale, presumably near the middle. Conversely, if people's scales were sub-actual, for instance if they used 10/10 to mean their average happiness, then we would see a 'bunching' with lots of scores at ends of the scale. If people's scales used the actual range and their experiences are roughly normally distributed (which seems plausible), then we would expect to see *reported* SWB to be roughly

[16] The other possible explanation is that each doubling of objective stimulus in fact causes a doubling of subjective intensity but that individuals only report a 1-unit increase per doubling because the relationship between reported and experienced intensity is logarithmic. This is, in effect, the opposite pair of assumptions from those normally made. I do not think this possibility has been seriously entertained, precisely because we believe individuals can detect equal-interval changes in subjective intensities and honestly report them when they occur.

normally distributed and for these scores to use the full range. This is just what we see. Hence, the data indicate that individuals are, broadly speaking, accurately reporting their subjective well-being *and* that they are using the ends of the scale to represent the actual limits.

I only say, 'broadly speaking', however and raise three issues. First, there is, in fact, *some* bunching in the top category both looking at worldwide distribution of life satisfaction scores and in all sub-regions; more people say they have the top level of satisfaction than we would expect from a normal distribution. It is unclear if this represents accurate reporting, a positivity/self-delusion bias that should be corrected for, or something. This is a topic for further research.

Second, there is some evidence that individuals use their own previous happiness, rather than the happiness of others, as the reference points when answering happiness questionnaires (Steffel and Oppenheimer, 2009). This is in tension with the conclusion that individuals use the real (collective) limits for the ends of their scales. As noted, if individuals use themselves as the references points for their scale, this has troubling implication that scales may not be cardinal over time (if individuals change their scales) or across people (if individuals have different limits). As this is from a small, convenience sample of American college students, it should carry less weight than a global survey.

Perhaps relatedly, in the realm of health, there's evidence that people rate 'perfect health' relative to their age group, rather than the best health possible for anyone of any age. It is unclear if this is specific to health: speculatively, as heath varies so much between ages, respondents may expect the helpful way to answer the question is with reference to their age group.[17] These are two further topics for further research. Clearly, we do not always use ourselves as the reference points for all scales. If I ask, as Oswald (2008) did, how tall you are on a 0-10 scale, it would be perverse to say "10/10" on the grounds are currently the tallest you have ever been.

Third, even if the *general* pattern is for individuals to use the end points to represent the actual limit, this is consistent with various groups having different scale-lengths and levels, or people having different lengths and levels at different times. Hence, we need to say more about intertemporal and interpersonal comparisons.

[17] Note SWB also varies with age. See e.g. Blanchflower (2020)

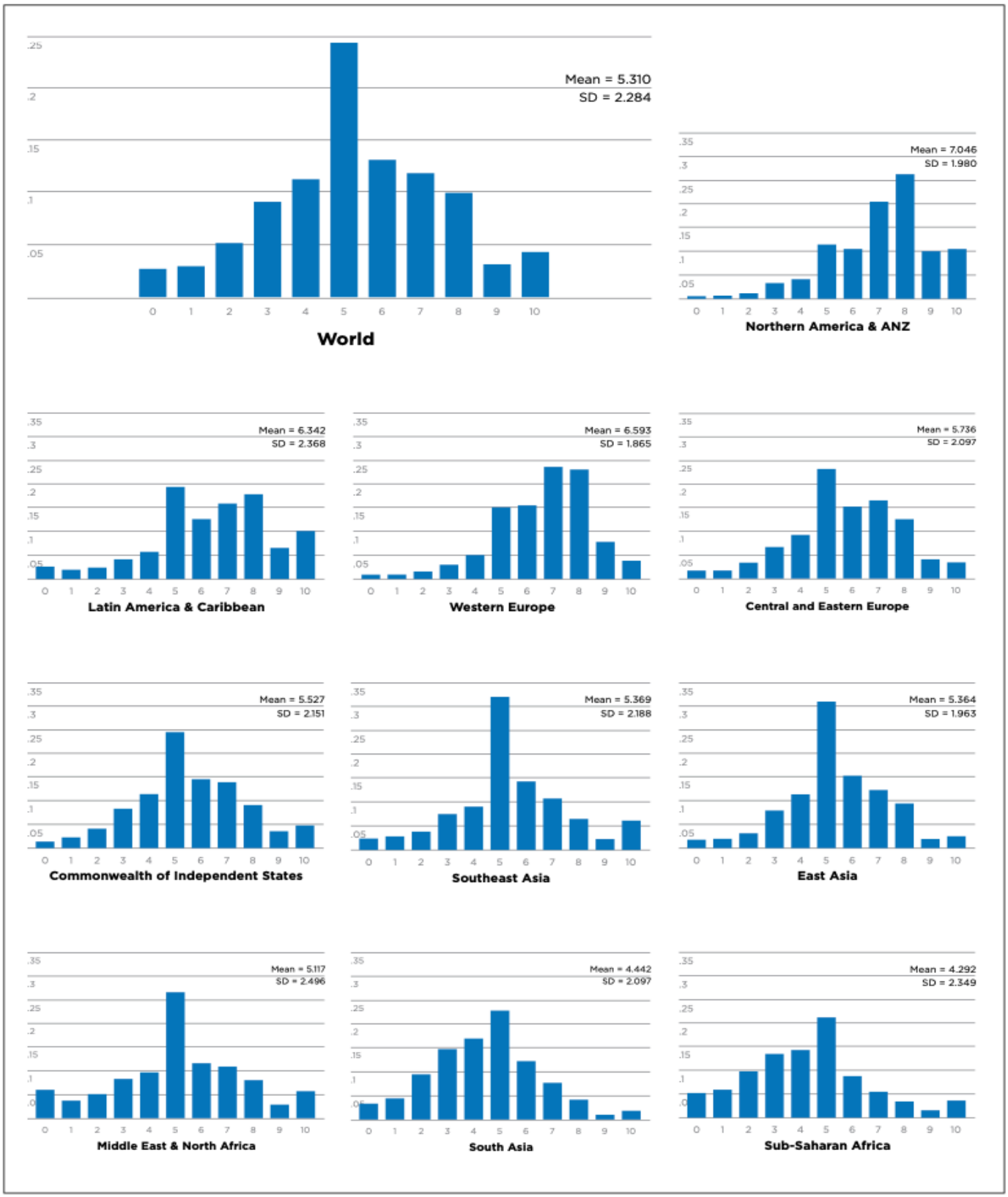

Figure 5. Worldwide distribution Cantril ladder scores. These are generated by the following prompt: "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" Graph reproduced from World Happiness Report 2017 (Helliwell, Layard and Sachs, 2017).

5.4 Condition 3: intertemporality

Do people use the same scale over time—might a '6/10' mean something different at different moments? Here, we need to distinguish between contextual factors—those related to the survey itself—and acontextual factors.

The psychological literature details several contextual factors that affect responses to subjective well-being questions, such as finding a coin, being asked about your love life, your satisfaction with politics, or whether you are surveyed in person or over the phone (Schwarz and Strack, 1999; Deaton, 2012; Dolan and Kavetsos, 2016.

Contextual factors pose no problem for intertemporal cardinality. Not only do mood and item-order effect have a relatively small impact, if questionnaires are asked in a standardised way, any contextual effects will 'wash-out' on average (Eid and Diener, 2004; Schimmack and Oishi, 2005).

A genuine threat is posed by acontextual factors. Ng (2008) observes that happiness researchers seem not to have noticed that individuals can *rescale*—alter what their scale's end-points represent—over their lives.

It's important to differentiate rescaling from *(hedonic) adaptation*, where the subjective impact of some event reduces over time. To illustrate, suppose Sam, a professional tennis player, reports he is 8/10 happy now. Sam then has an accident and is unable to play tennis. He is surveyed a year later and says he is 8/10 happy. One possibility is he had adapted to his new life and is genuinely as happy. Another is that he is less happy but has rescaled—specifically, he has shrunk his scale, lowering the level of happiness a 10/10 represents. A third is that some combination of rescaling and adaptation has occurred. To be clear, adaptation poses no threat to intertemporal cardinality as the same numbers still represent the same intensities of experiences.[18]

While there is a literature on reported adaptation to life shocks, e.g. divorce, unemployment, etc., authors seem to explicitly or implicitly assume intertemporal scale cardinality, rather than argue for it, and hence conclude adaptation genuinely occurs where it is reported. See Luhmann *et al.* (2012) for a review. In fact, I am unaware of any substantial discussion of whether reported adaptation is better explained by rescaling, genuine adaptation, or some combination of the two. As an example of typically brief discussion, Oswald and Powdthavee (2008, p16) state only that "There is probably no way to reject such concerns [about rescaling] definitively, but one objection to it is that in our data there is a continuing negative effect from longstanding disability; this seems inconsistent with the claim that disabled people fundamentally rescale their use of language".

To make progress here, we need to draw on and/or develop our theories of how adaptation and rescaling work.

Starting with rescaling, I not aware of any detailed suggestions of how and why rescaling might occur. Based on the earlier theory that individuals want to make themselves understood, I expect that individuals will be reluctant to rescale and will only do so if forced to by new evidence. Hence, we would only expect individuals to rescale if they experience extreme *and* unexpected events; if

[18] A separate worry, which originates from Sen (1987) pp45-6 is whether the fact people do (or could) be happy in deprived circumstances. If we think such people are happy but have low well-being, that means well-being cannot consist in happiness. This is not our concern here.

they had correctly anticipated the event, they would already have made their scale wide enough to accommodate it.

How many extreme, unexpected events do people run into? Intuitively, not many; we would expect people to have *some* idea of what the range of human experiences are—the joys of sex and love, the horrors of war and suffering—either first-hand or from others' reports. In this case, we would not expect to see much rescaling at all.

An alternative possibility is that individuals are frequently surprised by the subjective intensity of events. In this case, we would see rescaling occur to all sorts of major events as individuals recalibrate their endpoints.

One way to test these hypotheses about rescaling would simply be to ask a range of people if they had rescaled, see if anyone claims they have and, if so, what this is associated with. Another would be to find, in panel data sets, those who we expect would have had rescaling events—e.g. extreme pain. If we observed changes in other variables, such in income or relationship status, were associated with a smaller changed in subjective well-being, that would indicate a stretching of scales. These are topics for additional study.

Regarding hedonic adaptation, there are good evolutionary reasons to expect this occur, but only to occur for some events; for more detailed discussion see Perez-Truglia (2012) and Graham and Oswald (2010). The idea is that affective states are 'Mother Nature's' way to punish/reward animals for actions that increase/reduce our ability to survive and reproduce. Producing these sensations is costly in terms of energy, so hedonic adaptation is the solution that reduce costs whilst maintaining motivation. Hedonic adaptation can occur at the cognitive level too—people change their views on things (Wilson and Gilbert, 2005). We wouldn't expect hedonic adaptation to occur in response to a situation that continues to be good/bad for the creature's survival and reproduction; for instance, it would be disadvantageous to fully adapt to pain, as then pain would not be serving its warning function. As evidence of pain's usefulness, those with congenital immunity to pain, a rare medical condition, often end up severely damaging themselves (Udayashankar et al., 2012).

Turning to the data, what we see from life satisfaction scores is that people report adaptation to some things—e.g. becoming bereaved—and not to others—e.g. being unemployed or disabled (Clark et al., 2008; Clark et al., 2018; Luhmann et al., 2012). If people were often surprised at how good/bad extreme experiences were (the second hypothesis) then we would not see this; instead, we would observe reports of adaptation across the board. Further, when we look at which events people do and do not reportedly adapt to, these fit our evolutionary hypothesis of adaptation: we can understand why disability and unemployment would keep being be bad—the former makes life difficult, the latter feels shameful, and both increase isolation—and why (most) people eventually adjust to bereavement—it does not enhance survival and reproductive fitness to remain sad and unmotivated to find a new (reproductive) partner.

As it stands, appealing to the principle of Occam's Razor, we are pushed to abandon the hypothesis that rescaling occurs—it does not clearly help us explain the results.

A further test of rescaling comes from utilising data on memories. Prati and Senik (2020) compare *remembered* SWB—how satisfied individuals recall being—with *observed past* SWB—how satisfied

individuals they said they were at the time. Specifically, they analyse data from respondents of a German panel, who had been asked about their life satisfaction for years, were given the nine different pictures of changes in life satisfaction over time (see figure 5) and asked to pick one which best represented their own life.

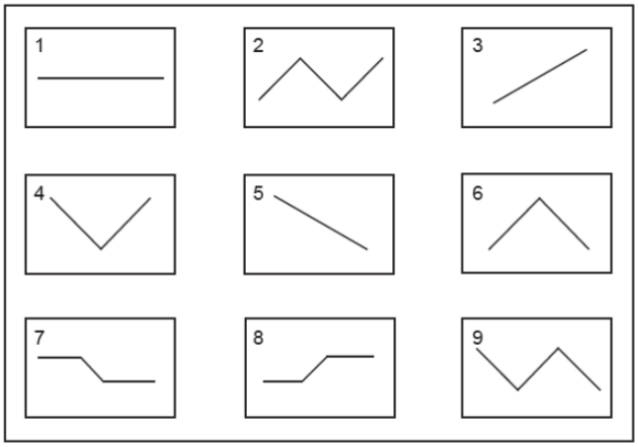

Figure 5. Potential patterns of recalled satisfaction

Figure 6 displays, for each group that picked a schematic pattern, what their average observed life satisfaction was.

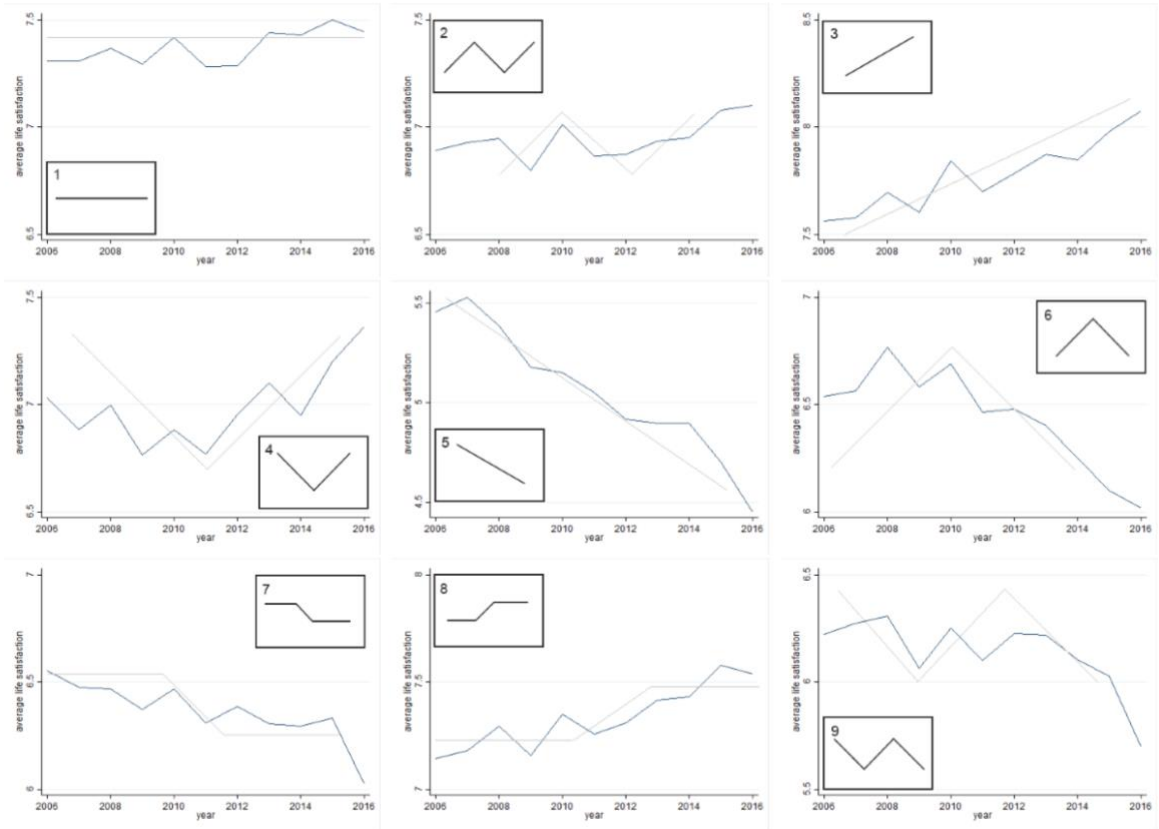

Figure 6. Observed past satisfaction, conditional on chosen pattern. Reproduced from Prati and Senik (2020)

It is worth stressing this is an extremely cognitively demanding tasks 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 extremely impressive.

Where does this tell us about whether rescaling occurs? If we accept there is consistency between remembered and observed past satisfaction, there seem to be only two ways to explain this.

Either (A) individuals both use the same scale over time *and* have good memories or (B) individuals change the scale use *and* have bad memories. If individuals used the same scales and had bad memories, or used different scales and had good memories, there would be an *in*consistency between the recalled and past observed patterns. Prati and Senik do not raise this second possibility.

Of the two options, (A) is the more elegant answer than (B): (B) requires not only that individuals change their scales over time—which, as noted, they may want to resist doing to ensure comparability of their answers—but that they have also poor memories. In fact, (B) requires quite specific and implausible patterns of memory failure.

To illustrate, suppose you're quite satisfied and your life satisfaction has been flat over time. This is shown by the horizontal arrow in Fig. 7a. Also suppose that, the maximum level of your life satisfaction scale has shrunk, illustrated by the change in length of the vertical arrows in Fig 7a. Because of this, your reported 0-10 level of satisfaction had therefore been rising over time (Fig 7b). To make your patterns of observed past satisfaction and recalled satisfaction consistent, given this upper-bound scale shrinkage, you would need to falsely recall that your satisfaction has increased (Fig 7b). If you instead erroneously recalled that your satisfaction had *de*creased, then there would be inconsistency between observation and recall.

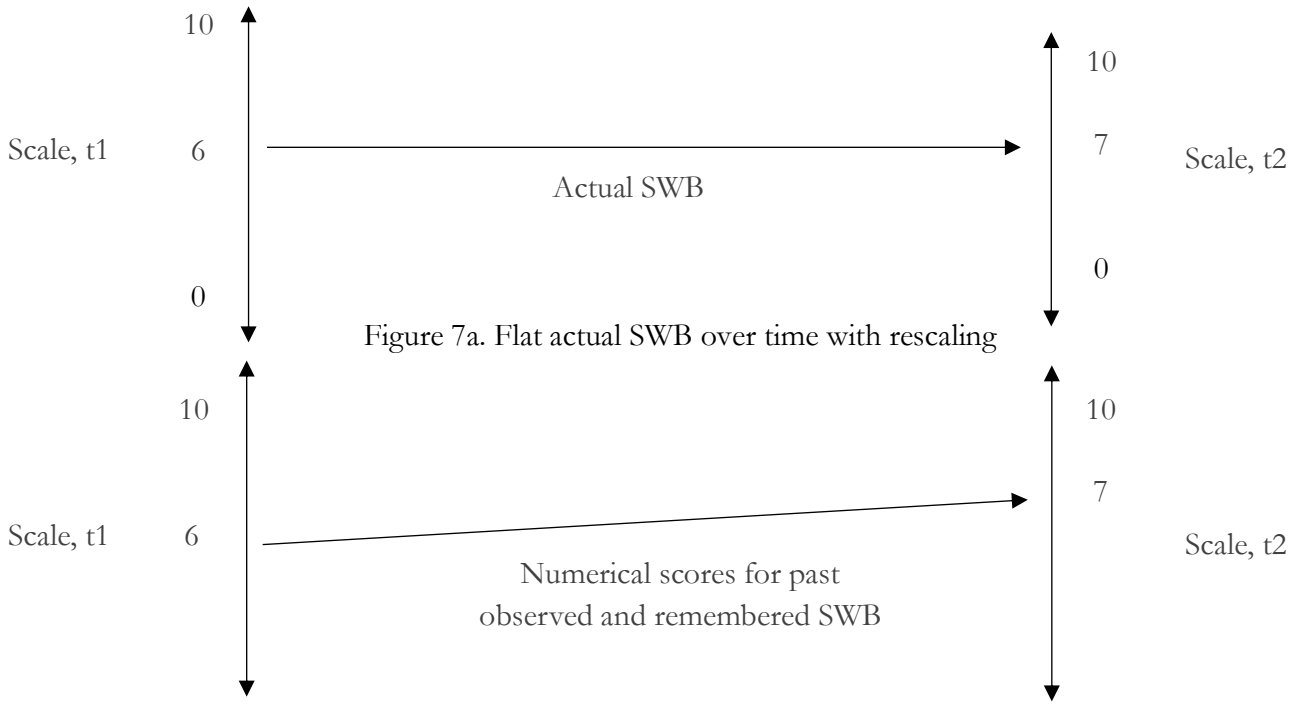

Figure 7a. Flat actual SWB over time with rescaling

Figure 7b. Changes in reported numerical scores resulting from constant actual SWB over time combined with rescaling

The same specificity applies the other way around. If your experienced satisfaction was flat but your scales had, in fact, stretched, then your reported satisfaction would go down; to make that consistent with recalled satisfaction going down, you'd need to falsely recall a decrease in satisfaction.

While it is possible to imagine we have *some* memory failures, it seems distinctly unlikely that rescaling would regularly be associated with a particular direction of memory failure. The most plausible hypothesis, then, is that individuals do use the same scales over time.

A tension here is that Prati and Senik's main conclusion is that individuals do exhibit a present bias in memory: those who are satisfied now are likely to recall being more satisfied than they said at the time. To derive this result, the authors assume intertemporality. The oddness here is how to make consistent the belief that (a) our memories are *good* (as evidenced by the match between the patterns of recalled and past observed satisfaction) with (b) that they are not *that* good (as evidenced by the present bias). I leave this topic for further investigation.

5.5 Condition 4: interpersonality

The final condition is whether different individuals use the same endpoints at a time. There are two types of concern here.

The first is whether there are what Nozick (1974, 41) called 'utility monsters', individuals who can and do experience much greater magnitudes of happiness (or any other sort of subjective state), than others.

I won't dwell on this as it seems unlikely there would be substantial differences in humans' capacities for subjective experiences. Presumably there are evolutionary pressures for each species to have range of sensitivity that is optimal for survival. To return to an example noted earlier, being immune to pain is an extremely problematic condition that would put someone at an evolutionary disadvantage. Further, even if there are differences, we would expect these to be randomly distributed, in which case they would wash out in large samples.[19]

The second concern is whether, continuing the terminology, there are 'language monsters', those who systematically using language very differently from each other. It seems unlikely there would be non-random differences between individuals within the same linguistic community, such as a nation-state. Why?

An influential idea from Wittgenstein (1953), another philosopher of language, is that the meaning of words is just their use in language ('meaning as use'). Hence, if we succeed in making ourselves understood via language, which we seem to do, then we do so in virtue of using words in the same way as each other. The application to this topic is as follows: given we seem to be able to communicate in general, it follows we are able to communicate about subjective states in particular. As a test of the reader's intuitions, I ask them to consider if they really believe that if someone said they were "2/10 happy" compared to "8/10 happy", they would really have no idea what the person was feeling. There does not seem to be anything special about using numbers, rather than

[19] It is worth noting that we might well expect there to be large differences between (a) current humans and both (b) non-human animals and (c) humans' whose genetics has substantially changed either naturally, due to the passage of much time, or deliberately through gene editing. It is an open question whether and how we could make cardinal comparisons between these groups

words, to convey meaning: we equally understand what someone means when they have a 'terrible' vs a 'wonderful' day.

That we use words the same way is less mysterious when we realise we often regulate each other's use of appropriate language, the process of *meta-linguistic negotiation*; for some discussion, see Plunkett (2015). For instance, if I were to say, in all earnestness, "I got a papercut, I am in agony", I might expect the response, "come on. That's not what *agony* feels like." Hence, we should only expect individuals to be language monsters with respect to subjective scale only if we cannot communicate with them in ordinary contexts.

However, even if the users of a language successful share meaning, we might still expect there to be systematic differences *between* linguistic groups, i.e. different cultures and/or nations.

There are mean-level differences in SWB across countries (Helliwell, Layard and Sachs, 2017). As Diener et al. (2017) observe, these could be explained either by culture-specific scale use, by differences in the conditions of those countries, or some combination of the two. How would we go investigating this?

In the current literature, three methods for testing interpersonality have been proposed to date. Each requires that we assume something is constant across individuals and then uses that constant to check if the scales change. I'll briefly explain these and what has been found using them. The results are inconsistent, so I then make further comments about which assumption to abandon.

One is *vignettes*, where survey participants are given a description of someone's life and then asked to rate how satisfied that person is. Here is an example vignette taken from Angelini *et al.* (2014)

> John is 63 years old. His wife died 2 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 for 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 difference in reports will be due to differential scale use (King et al., 2004).

Kahneman et al. (2004) observe surprisingly large differences in European data: for example 64% of Danes say they are "very satisfied" but only 16% of the French do. Angelini et al. (2014) provide vignettes to individuals in different Europeans countries and find relatively large response patterns to those vignettes. When these are accounted for and the scales correspondingly 'corrected', the differences between the Danes and the French (as well as many other nations) disappear. This indicates a lack of inter*country* cardinality.

The second is what we might call *variable equivalence*, where we assume the same variable—income, partnership, employment, etc.—have the same true impact in different places, at least if we control

for the other variables. If we accept this then, if the coefficients of the explanatory variables have the same size in different places, that indicates scales are the same range.[20] If we instead found, for instance, that the variables had half the coefficient size in France as Germany, we would assume the French had a scale that was twice as long.

In an analysis of this type, Helliwell et al. (2009) examined the predictors of life satisfaction across nations—e.g. income, levels of social capital, corruption—and found that the regression equations were effectively the same world over, which supports the sameness of scale length.

The third is to assume intertemporality, i.e. C3, and then observe reported changes from migrants. Assuming C3, any changes will be down to the real differences caused by living in one society compared to another.

Helliwell, Bonikowska and Shiplett (2016) studied immigrants moving from over 100 different countries to Canada, and found that, regardless of country of origin, the average levels and distributions of life satisfaction among the immigrants mimic those of Canadians. As their reports converge with those of each other, and the extant Canadians, that indicates that despite different cultural backgrounds, the individuals have the same scales as each other. To see the importance of the intertemporal cardinality assumption here, note that if we thought the new arrivals' scales changed on arrival, then that would be an explanation for their convergent scores.

These three sets of results are inconsistent: the vignette approach indicates there are international differences in scales, while the other two indicate identical scale use. In light of this, (at least) one approach must be faulty.

The approach that seems most likely faulty is the vignettes. Vignette equivalence assumes that respondents will agree how satisfied the people in the vignettes are and use that make inferences about differential scale use. However, respondents do not seem to agree. For instance, Angelini *et al.* (2014) find about 30% of Germans rate 'John' from the above vignette as satisfied or very satisfied, but 30% rate him dissatisfied or very dissatisfied. To assume that the respondents agree about John's life satisfaction requires us to conclude that respondents must mean the same thing by "satisfied" as "dissatisfied", which strains credulity seeing as one is positive and the other negative. Faced with a choice of vignette equivalence or *semantic equivalence* (that respondents attach the same meaning to words) the latter seems more plausible.

6. Does the existing evidence support the Cardinality Thesis? What further tests could be done? What would we do if it were false?

That concludes the survey of the existing evidence. While the evidence base is perhaps uncomfortably sparse in places, there does not seem be good reason to reject any of the conditions and accept alternatives to them. Hence, it seems reasonable to conclude the cardinality thesis is true, at least approximately, unless and until new evidence suggests others. The current crop of studies does, nevertheless, throw up some puzzles, notably: (1) how to make sense of the claim our memories are good but still biased; (2) apparent inconsistencies between different tests for interpersonal cardinality; (3) how to reconcile the conclusion CT holds with the apparent finding

[20] I am grateful to Caspar Kaiser for pointing out this would leave open what levels the scales covered.

individuals sometimes answer subjective scales relative to themselves or those like them, rather that the real limits for everyone.

The next question is how to further test the CT. I have already made some suggestions above. In addition, we can also generate testable predictions from the Schelling point hypothesis. I'll sketch these now.

What the Schelling point hypothesis requires is that individuals interpret scales in a very particular way. However, as far as I am aware, there is not research which directly asks individuals how they think they use subjective scales—specifically, what reporting function and endpoints they use—or tries to experimentally infer how people use them.

One avenue for investigation would be to give individuals a choice of different reporting functions and scale endpoints and then ask them to state which one is most similar to their own. For instance, whether they think 10/10 refers to the "happiest they've ever been", "the happiest they could be in reality", "the happiest anyone alive today could be in reality", "the happiest anyone could possibly be within the laws of nature", and so on. It would support the hypothesis if individuals think that they use linear reporting functions with the real limits for the end-points. Relatedly, individuals could be asked how they expect others to interpret subjective scales. Qualitative questions about why they picked their choices would also be illuminating. Even if we are worried that individuals think they behave in one way, but in fact behave in another, such research would have some value and build an evidential picture.

The other, related avenue is to see how individuals behave, rather than ask them how they think they behave. Differently labelled scales could be presented, both to the same or to different groups of individuals, and those individuals asked to give ratings for themselves. For instance, respondents could be asked to rate their happiness on a series of 0-10 scales, where the 10s are variously: left unlabelled; marked "the happiest you've been so far"; "the happiest any human has ever been in history"; "the happiest it is possible for any life form to be within the laws of nature" and so on (with equivalent markings for the 0). The reporting function could be tested in a similar way: the same and different individuals would be presented with an array of explicitly specified or entirely unspecified options and asked to give subjective ratings. We would then conclude that the labelled scale and reporting functions which has answers most similar to the unlabelled ones would be how individuals, by default, interpret subjective scales. It would support the hypothesis if, for instance, scores were most similar between the unlabelled reporting function and the reporting function labelled as linear.

Of course, further work might indicate the cardinality thesis is false. Can we, as Stone and Krueger wonder, adjust for this? As stated in section 4, if the problem is C1, the answer is No. However, if it is C2, C3, or C4 the answer is Yes. As noted in section 4, there are only a few ways our 'measuring sticks' can go wrong; hence, if we understand where and what type of error there is, we can correct for it.

I will elaborate with an example. Suppose individuals state and/or we infer from their responses that they use a particular non-linear reporting function, say a specific logarithmic one. In this case, as Kristofferson notes, we could simply apply the relevant mathematical function to the raw self-

reported data so that the transformed scores are then cardinally comparable. A similar approach could be applied for the intertemporality and interpersonality conditions.

7. Conclusion

Much of our lives, including much of science, relies on the use of subjective data. These data are often thought to be cardinally comparable. If they were not, that would substantially limit what we would be able to infer from them and necessitate a stark reconsideration of what we think we know. This topic seems to have attracted much attention; as a result, there were gaps in theoretical and practical understanding. This paper has tried to fill some of these gaps. I have proposed a theory of how individuals might intuitively interpret subjective scales in order to be understood and how this would lead to their answers being cardinal comparable. I then identified several distinct conditions for cardinality, explained how we could, in theory, assess them, and then went on to do so using the existing evidence. This highlights various areas for further research but left us with the tentative conclusion that subjective scales are best understood as cardinally comparable. I closed by explaining how a lack of cardinality in the 'raw' subjective data is fixable and should not cause us to abandon its use. The conclusion of this essay is therefore optimistic. Not only have we not found a problem where we feared there might be one, but we may well be able to fix the problem if we later discover it does exist.

Bibliography

Adler, M. D. (2013) 'Happiness Surveys and Public Policy: What's the Use?', *Duke Law Journal*, pp. 1509–1601.

Alexandrova, A. (2012) 'Well-Being as an Object of Science', *Philosophy of Science*, 79(5), pp. 678–689. doi: 10.1086/667870.

Alexandrova, A. (2016) 'Is well-being measurable after all?', *Public Health Ethics*. Narnia, 10(June), pp. 1–15. doi: 10.1111/1467-9973.00225.

Angelini, V. *et al.* (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*. Blackwell Publishing Ltd, 76(5), pp. 643–666. doi: 10.1111/obes.12039.

Angner, E. (2013) 'Is it possible to measure happiness?: The argument from measurability', *European Journal for Philosophy of Science*, 3(2), pp. 221–240. doi: 10.1007/s13194-013-0065-2.

Benjamin, Dan *et al.* (2020) 'Self-reported wellbeing indicators are a valuable complement to traditional economic indicators but are not yet ready to compete with them', *Behavioural Public Policy*. Cambridge University Press (CUP), 4(2), pp. 198–209. doi: 10.1017/bpp.2019.43.

Benjamin, D *et al.* (2020) *What Do Happiness Data Mean? Evidence from a Survey of Happiness Respondents**.

Blanchflower, D. (2020) *Is Happiness U-shaped Everywhere? Age and Subjective Well-being in 132 Countries*. Cambridge, MA. doi: 10.3386/w26641.

Bond, T. and Lang, K. (2018) *The Sad Truth About Happiness Scales: Empirical Results*. Cambridge, MA. doi: 10.3386/w24853.

Boyd, R. (1980) 'Scientific Realism and Naturalistic Epistemology', *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. The University of Chicago PressPhilosophy of Science Association, pp. 613–662. doi: 10.2307/192615.

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

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

Clark, A. E. *et al.* (2008) 'Lags and leads in life satisfaction: a test of the baseline hypothesis', *The Economic Journal*, 118(529), p. F243.

Clark, A. E. *et al.* (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), pp. 619–645.

Davies, W. (2015) *The happiness industry: How the government and big business sold us well-being*. Verso Books.

Deaton, A. (2012) 'The financial crisis and the well-being of Americans', *Oxford Economic Papers*. Cambridge, MA, 64(1), pp. 1–26. doi: 10.1093/oep/gpr051.

Van De Deijl, W. (2017) 'Which Problem of Adaptation?', *Utilitas*. Cambridge University Press, 29(4), pp. 474–492. doi: 10.1017/S0953820816000431.

Diener, E. *et al.* (2017) 'Findings all psychologists should know from the new science on subjective well-being.', *Canadian Psychology/Psychologie canadienne*, 58(2), pp. 87–104. doi: 10.1037/cap0000063.

Dolan, P. and Kavetsos, G. (2016) 'Happy talk: Mode of administration effects on subjective well-being', *Journal of Happiness Studies*, pp. 1–19.

Dolan, P., Peasgood, T. and 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), pp. 94–122. doi: 10.1016/j.joep.2007.09.001.

Dolan, P. and White, M. P. (2007) 'How Can Measures of Subjective Well-Being Be Used to Inform Public Policy?', *Perspectives on Psychological Science*. SAGE PublicationsSage CA: Los Angeles, CA, 2(1), pp. 71–85. doi: 10.1111/j.1745-6916.2007.00030.x.

Edgeworth, F. Y. (1881) *Mathematical Psychics*. London: Kegan Paul.

Edwards, E. (1964) 'On the theory of scales of measurement', *Ergonomics*. American Association for the Advancement of Science, pp. 504–505. doi: 10.1080/00140136408956259.

Eid, M. and Diener, E. (2004) 'Global judgments of subjective well-being: Situational variability and long-term stability', *Social Indicators Research*. Springer, 65(3), pp. 245–277. doi: 10.1023/B:SOCI.0000003801.89195.bc.

Ferrer-i-Carbonell, A. and Frijters, P. (2004) 'How Important is Methodology for the estimates of the determinants of Happiness?*', *The Economic Journal*. Blackwell Publishing Ltd, 114(497), pp. 641–659. doi: 10.1111/j.1468-0297.2004.00235.x.

Ferrer-i-Carbonell, A. and Frijters, P. (2004) 'How important is methodology for the estimates of the determinants of happiness?', *The Economic Journal*.

Galton, F. (1907) 'Vox populi', *Nature*, 75(1949), pp. 450–451. doi: 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*. Available at: 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/ (Accessed: 21 September 2020).

Graham, L. and Oswald, A. J. (2010) 'Hedonic capital, adaptation and resilience', *Journal of Economic Behavior and Organization*. North-Holland, 76(2), pp. 372–384. doi: 10.1016/j.jebo.2010.07.003.

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

Harman, G. H. (1965) 'The Inference to the Best Explanation', *The Philosophical Review*. JSTOR, 74(1), p. 88. doi: 10.2307/2183532.

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

Haybron, D. (2008) *The pursuit of unhappiness: The elusive psychology of well-being*. OUP.

Haybron, D. M. (2016) 'Mental State Approaches to Well-Being', in Adler, M. D. and Fleurbaey, M. (eds) *The Oxford Handbook of Well-Being and Public Policy*. Oxford University Press. doi: 10.1093/oxfordhb/9780199325818.013.11.

Heathwood, C. (2006) 'Desire satisfactionism and hedonism', *Philosophical Studies*, 128(3), pp. 539–563. doi: 10.1007/s11098-004-7817-y.

Helliwell, J. *et al.* (2009) *International evidence on the social context of well-being*. w14720.

Helliwell, J., Bonikowska, A. and Shiplett, H. (2016) *Migration as a Test of the Happiness Set Point Hypothesis: Evidence from Immigration to Canada*. Cambridge, MA. doi: 10.3386/w22601.

Helliwell, J. F., Layard, R. and Sachs, J. (2017) *World happiness report 2017*. Sustainable Development Solutions Network.

Helliwell, J., Layard, R. and Sachs, J. (2017) *World Happiness Report 2017*.

Kahneman, D. *et al.* (2004) 'Toward national well-being accounts', in *American Economic Review*, pp. 429–434. doi: 10.1257/0002828041301713.

Kaiser, C. and Vendrik, M. (2020) *How threatening are transformations of happiness scales*…. 2020–19.

King, G. *et al.* (2004) 'Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research', *American Political Science Review*, 98.

Kristoffersen, I. (2011) *The Subjective Wellbeing Scale: How Reasonable is the Cardinality Assumption?*, *Economics Discussion / Working Papers*. 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*. Springer Netherlands, 130(2), pp. 845–865. doi: 10.1007/s11205-015-1200-6.

Krueger, A. B. and Schkade, D. A. (2008) 'The reliability of subjective well-being measures', *Journal of Public Economics*. Elsevier, 92(8–9), pp. 1833–1845. doi: 10.1016/j.jpubeco.2007.12.015.

Lantz, B. (2013) *Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations*, *Electronic Journal on Business Research Methods* .

Layard, R. (2003) 'Happiness: has social science a clue? Lecture 1: what is happiness? Are we getting happier?', in *Lionel Robbins memorial lecture series*.

Luhmann, M. *et al.* (2012) 'Subjective well-being and adaptation to life events: a meta-analysis.', *Journal of personality and social psychology*, 102(3), p. 592.

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

Ng, Y. K. (1996) 'Happiness surveys: Some comparability issues and an exploratory survey based on just perceivable increments', *Social Indicators Research*. Springer Netherlands, pp. 1–27. doi: 10.1007/BF00293784.

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

Nozick, R. (1974) *Anarchy, state, and utopia*. New York: Basic Books.

Nussbaum, M. C. (2012) 'Who is the happy warrior? Philosophy, happiness research, and public policy', *International Review of Economics*. Springer-Verlag, 59(4), pp. 335–361. doi: 10.1007/s12232-012-0168-7.

OECD (2013) *Guidelines on Measuring Subjective Well-being*. OECD Publishing. doi: 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), pp. 369–372. doi: 10.1016/j.econlet.2008.02.032.

Oswald, A. J. and Powdthavee, N. (2008) 'Death, happiness, and the calculation of compensatory damages', *The Journal of Legal Studies*, 37(S2), p. S251.

Perez-Truglia, R. (2012) 'On the causes and consequences of hedonic adaptation', *Journal of Economic Psychology*, 33(6), pp. 1182–1192. doi: 10.1016/j.joep.2012.08.004.

Plunkett, D. (2015) 'Which Concepts Should We Use?: Metalinguistic Negotiations and The Methodology of Philosophy', *Inquiry (United Kingdom)*. Routledge, 58(7–8), pp. 828–874. doi: 10.1080/0020174X.2015.1080184.

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

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

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

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

Schimmack, U. and Oishi, S. (2005) 'The influence of chronically and temporarily accessible information on life satisfaction judgments', *Journal of Personality and Social Psychology*, 89(3), pp. 395–406. doi: 10.1037/0022-3514.89.3.395.

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), p. 153. doi: 10.2307/1403610.

Schwarz, N. and Strack, F. (1999) 'Reports of subjective well-being: Judgmental processes and their methodological implications', *Well-being: The foundations of hedonic psychology*, 7, pp. 61–84.

Sen, A. (1987) *On Ethics and Economics*. Oxford.

Steffel, M. and Oppenheimer, D. M. (2009) 'Happy by what standard? The role of interpersonal and intrapersonal comparisons in ratings of happiness', *Social Indicators Research*. Springer, 92(1), pp.

69–79. doi: 10.1007/s11205-008-9289-5.

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

Talbott, W. (2016) 'Bayesian Epistemology', *Stanford Encyclopedia of Philosophy*. Winter. Edited by E. Zalta.

Udayashankar, C., Oudeacoumar, P. and Nath, A. (2012) 'Congenital insensitivity to pain and anhidrosis: A case report from South India', *Indian Journal of Dermatology*. Wolters Kluwer -- Medknow Publications, 57(6), p. 503. doi: 10.4103/0019-5154.103080.

Williamson, T. (2017) 'Semantic paradoxes and abductive methodology', in Armour-Garb, B. (ed.) *Reflections on the Liar*. Oxford: OUP, pp. 325–346. doi: 10.1093/oso/9780199896042.003.0013.

Wilson, T. D. and Gilbert, D. T. (2005) 'Affective Forecasting', *Current Directions in Psychological Science*. SAGE PublicationsSage CA: Los Angeles, CA, 14(3), pp. 131–134. doi: 10.1111/j.0963-7214.2005.00355.x.

Wittgenstein, L. (1953) *Philosophical investigations*. Edited by G. Anscombe and R. Rhees. Oxford: Blackwell.