



What If Well-Being Measurements Are Non-Linear?

Daniel Wodak

To cite this article: Daniel Wodak (2018): What If Well-Being Measurements Are Non-Linear?, Australasian Journal of Philosophy, DOI: [10.1080/00048402.2018.1454483](https://doi.org/10.1080/00048402.2018.1454483)

To link to this article: <https://doi.org/10.1080/00048402.2018.1454483>



Published online: 08 Apr 2018.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



What If Well-Being Measurements Are Non-Linear?

Daniel Wodak 

Virginia Polytechnic Institute and State University

ABSTRACT

Well-being measurements are frequently used to support conclusions about a range of philosophically important issues. This is a problem, because we know too little about the intervals of the relevant scales. I argue that it is plausible that well-being measurements are non-linear, and that common beliefs that they are linear are not truth-tracking, so we are not justified in believing that well-being scales are linear. I then argue that this undermines common appeals to both hypothetical and actual well-being measurements; I first focus on the philosophical literature on prioritarianism and then discuss Kahneman's Peak-End Rule as a systematic bias. Finally, I discuss general implications for research on well-being, and suggest a better way of representing scales.

ARTICLE HISTORY Received 10 June 2017; Revised 17 February 2018

KEYWORDS wellbeing; welfare; well-being measurement; Kahneman; prioritarianism; psychological measurement; scales

1. Introduction

Philosophers, psychologists, economists, and policy-makers wish to infer a great deal from measurements of well-being about which policies are better for people, about ways in which decision-making is systematically irrational, and even about the plausibility of moral theories. My central contention in this paper is not that we can learn nothing from well-being measurements. It is that we can learn less than is often thought, for a neglected reason: we do not know whether well-being scales are *linear*.

In [section 2](#), I explain the difference between linear and non-linear scales. In [section 3](#), I argue that it is plausible that well-being scales are non-linear, and that common beliefs that they are linear are not truth-tracking, so we are not justified in believing that well-being scales are linear. In [section 4](#), I show how this undermines appeals to hypothetical well-being measurements in debates about utilitarianism and prioritarianism. In [section 5](#), I turn to actual well-being measurements in social science: I argue that widely accepted inferences from such measurements unjustifiably assume linearity, focusing on work by Daniel Kahneman. In [section 6](#), I argue that this problem has important implications for research on well-being, and in [section 7](#) I suggest a solution that turns on how well-being scales are represented.

2. What Are Non-Linear Scales?

With mental states, as with crowd sizes, it is often easy to know that *A* is greater than *B*, but hard to know the magnitude of the difference between them. We might know that

Obama's 2009 Inauguration drew a larger crowd than Trump's 2017 Inauguration did, or that Michelle is happier than Melania, while being ignorant of *how much* larger Obama's crowd was, or *how much* happier Michelle is. In such cases, an *ordinal scale* is appropriate. This is a rank order: Michelle's happiness > Melania's happiness.

When the magnitudes of differences are meaningful, an *interval scale* is appropriate. *Equal* magnitudes of differences give us a *linear scale*, such as the Celsius scale for temperature: the magnitudes of differences (in mean kinetic energy) between 2 and 3 and between 6 and 7 are the same. As Stevens [1959: 31–4] noted, 'logarithmic interval scales' have meaningful but unequal magnitudes of differences between intervals: on the Richter scale, for instance, the magnitude of the difference (in terms of the energy released by earthquakes) between 6 and 7 is not equal to the magnitude of the difference between 2 and 3; it is a million times greater.

To assume that well-being scales are linear is to assume that they have equal magnitudes of differences between intervals. (This is how 'linearity' is often used in the literature: for example, Myles et al. [1999].) For my purposes, any scale without this feature is 'non-linear': this includes ordinal scales (for instance, Mohs's hardness scale) and logarithmic scales (for instance, the Richter scale).

A note about terminological variance is warranted here. Imagine a series of scores—3, 5, 7—on a well-being scale. If one infers that the total and average *well-being* for the series are 15 and 5, respectively, one is assuming linearity. Sometimes the assumption that would license such inferences is described by using other terms ('interval', 'ratio', 'cardinal'). This is not the place to analyse or quibble with classifications of scales; all that matters for our purposes is whether one makes the assumption that licenses inferences like the one above, regardless of how it is described.

3. Are Well-Being Measurements Linear?

Many scales are used to measure well-being and putative subjective components thereof, like pleasure and pain.¹ Whether these scales are linear is an empirical question. But it is rarely tested (section 6). Instead, 'most measures of subjective well-being are assumed to be ordinal, rather than cardinal', yet 'treating [them] as if they were cardinal' is commonplace [OECD 2013: 189–90]. Well-being measurements are summed and averaged as if they are linear. This raises the question: in the absence of empirical evidence, are we justified in believing that well-being scales are linear?

No. I offer three arguments to support this. The first two may support a stronger conclusion—that we are justified in believing that they are *non-linear*—but I don't need that claim. That we are not justified in believing that well-being scales are linear will suffice for my drawing significant implications for philosophical and psychological practices (sections 4–5).

First, that we use ordinal scales for subjective components of well-being relies on weaker assumptions—namely, that the relevant psychological attributes are merely ordinal [Michell 2012], or that our introspection and self-reporting about such attributes does not encode the complex relational information required by any interval scale.

¹ I am striving to be neutral between different conceptions of well-being. I leave open whether my arguments extend to measuring *non-subjective* components of well-being.

In the absence of empirical evidence to the contrary, we should rely on conservative assumptions.

Second, logarithmic scales are commonly used in reporting subjective phenomena. This has been a dominant view since psychophysics emerged [Fechner 1966]. We default to using non-linear scales when reporting brightness [Bartleson and Breneman 1967; Pinoli 1997]. ‘Untrained observers’, asked to report subjective loudness, ‘use a scale that is nearer to logarithmic than to linear’, which ‘biases ... arithmetic means’ [Poulton et al. 1980: 96]. Why? Because, as Gelfand [2009: 4] explains,

the sound pressure of the loudest sound that we can tolerate is on the order of 10 million times greater than that of the softest audible sound. One can immediately imagine the cumbersome task that would be involved if we were to deal with such an immense range of numbers on a linear scale. The problems involved with and related to such a wide range of values make it desirable to transform the absolute physical magnitudes into another form ... which make the values both palatable and rationally meaningful.

Similar reasoning can apply to well-being. The spectrum from indifference to noticeable pain to agony is vast. To represent this by using numbers linearly would be a ‘cumbersome task’; but we could represent those values logarithmically in a way that is ‘palatable’ and ‘meaningful’.

If we are not justified in believing that well-being measurements are linear, why are they so often treated as linear? One plausible explanation turns on how well-being measurements are represented by using devices with linear properties, like numerals. Consider ‘the representational fallacy’—the fallacious inference from the premise that some representative device has salient feature F to the conclusion that the represented feature of the world also has salient feature F .² To illustrate, say that someone believed that 32 degrees Fahrenheit is twice as hot as 16 degrees Fahrenheit, because these temperatures are represented with the integers 32 and 16, and 32 is 16 times 2. The inference is fallacious. (The conclusion translates to ‘0 degrees Celsius is twice as hot as –9 degrees Celsius.’) The use of integers in the Fahrenheit and Celsius scales is potentially misleading because these scales don’t have all of the salient features of this representative device (they lack non-arbitrary zero points, so are not ‘ratio’ scales), yet we could easily *infer* that they do.

Whether someone is subject to the representational fallacy is an empirical issue. I do not know how frequent such mistakes are with temperature.³ But Joel Michell’s work suggests that psychologists often make similar fallacious inferences. He describes ‘the psychometrician’s fallacy’ as the fallacious inference of an interval scale from an ordinal scale (see, *inter alia*, Michell [2009a, 2009b, 2012]). Michell provides historical and contemporary examples of psychologists measuring attributes on an ordinal scale, then treating the measurements *as if* they are on an interval scale, such that we can aggregate data. There is disagreement about this fallacy. For instance, Borsboom and Mellenbergh [2004] and Michell [2004] disagree about whether it is pervasive enough to make psychometrics a ‘pathological science’.⁴ But both agree, as Borsboom and Mellenbergh

² This differs from Dyke’s [2014: 14] use of the same term to refer ‘to a general philosophical tendency to place too much emphasis on language when doing ontology’.

³ There are plenty of threads and columns about such mistakes online. For instance, Larry Scheckel’s 22 January 2014 ‘Ask Your Science Teacher’ column, on *The Tomah Journal*, described ‘What temperature is twice as hot as zero degrees?’ as a ‘tricky question’.

⁴ I would like to thank an anonymous referee for pushing me on this point.

[2004: 118] note, that ‘in much psychological research, item scores are simply summed and declared to be measurements of an attribute, without any attempt being made to justify this conclusion.’

Many have expressed concerns about similar practices in relation to the psychological measurement of subjective components of well-being. In their overview of pain measurement, Chapman et al. [1985: 9–10] wrote:

The quantification of subjects’ responses on rating scales can be problematic. Although pain experiences are classified into categories in scales, the categorization implies a rank ordering. The category boundaries are not known and the approximation of the ranked categories to equal intervals is often assumed but not demonstrated. Some investigators simply assign numbers to the categories that rank in magnitude with the category descriptor and score subject judgments by statistically manipulating the numbers.

In other words, investigators assign numbers to represent categories in scales, then treat those categories as if they have a salient feature of the representative device: they are treated ‘like equally spaced numbers’, such that ‘statistically manipulating the numbers’ is legitimate.

One interpretation of this practice is that investigators are attempting to *disguise* the limitations of their research (as Michell suggests: [2004: 122]). But this does not fit common justifications for using numerals (see section 6). Nor does it explain why measurements are assumed to be *linear* (‘like equally spaced numbers’), rather than *not ordinal*. ‘The statistics applicable to measurements made on a logarithmic interval scale’, Stevens [1959: 33] noted, ‘include those appropriate to a linear interval scale, except that we would need to work with the logarithms of the scale values rather than the scale values themselves’. Moreover, it does not generalize: investigators’ incentives do not explain why others, like philosophers, treat well-being measurements as if they are linear.

A more charitable and generalizable interpretation of this practice is that investigators and philosophers succumb to the representational fallacy. They believe that scales are linear (rather than not ordinal) *because* scales are represented with a device that has linear properties.

An upshot of this explanation is that if well-being measurements were non-linear, we would be disposed to believe that they are linear. This upshot is imprecise.⁵ Who is this ‘we’? Presumably, it is those who engage with well-being measurements. What is it to believe that measurements are linear? This need not involve conscious thoughts with that content; it suffices that we are disposed to draw inferences as if well-being measurements are linear. These dispositions could be ‘masked’, or corrected, in any particular agent: just as we may learn that the zero point on the Fahrenheit scale is arbitrary, we may learn to suspend judgment about whether the intervals on well-being scales are equal.

The point is that, in so far as one would believe that well-being measurements are linear even if they are not, one’s beliefs are not sensitive to the truth; at best, those beliefs are accidentally true. So, even if one’s belief that well-being measurements are linear is true, it is unjustified.⁶ This is the final argument for why we

⁵ I would like to thank an anonymous referee and editor for pushing me to clarify this.

⁶ This step in the argument could be defended on the basis of *modal* conditions, such as versions of the Sensitivity principle (see Nozick [1981: 179]), or non-modal *explanatory* conditions (see, e.g., Shafer [2014]). It is often thought that such conditions on knowledge can extend to conditions for justified belief (see, e.g., Setiya [2012: 139]). I am grateful to an anonymous referee for pushing me to clarify the relevant epistemic standard.

should not believe that well-being measurements are linear: unless they are based on empirical evidence, such beliefs are not truth-tracking.

4. Hypothetical Measurements of Well-Being

Does it matter whether we are justified in believing that well-being scales are linear? Yes. If we aren't, this undermines common appeals to well-being measurements to support philosophically significant conclusions.

Consider appeals to hypothetical well-being measurements in philosophy, such as Roger Crisp's [2003: 745–6] appeal to this 'pair of distributions' of well-being:

	Group 1	Group 2
Equality	50	50
Inequality	10	90

Assume that each group contains the same number of people (say, 1,000) and that there are no questions of desert at issue. The numbers represent the welfare of each individual in each group: the individuals in Equality have equally good lives, while those in Inequality have lives that are either much better or much worse than the lives of those in Equality.

In a footnote, Crisp stipulates that the *numerical units on this scale* should be treated in the same way as we treat *numbers*. This stipulation is crucial. Without it, we could not infer that the total utilities in Equality (50+50) and Inequality (10+90) are equal, and hence that '[a]ccording to traditional utilitarianism ... there is no reason to choose one over the other' [ibid.: 746].

This exemplifies a widespread practice. As Michael Otsuka notes, in the extensive debate about utilitarianism et al., 'it is often left unspecified what constitutes a greater, lesser, or equal improvement in a person's utility': it is 'stipulate[d]' that there are 'numerical benefits of different magnitudes that comprise intervals along a whole number cardinal scale that is meant to represent the absolute levels of people's utility in linear fashion' [2015: 1–2].

I argue that, because we are not justified in believing that this stipulation is true, such appeals to well-being measurements are either potentially misleading or redundant. Either way, they should play no role in evaluating moral theories.

Some may resist this. Philosophers are entitled to stipulate details about thought experiments, including distributions of well-being! True. But thought experiments work by eliciting intuitions, and the object of those intuitions is some proposition *about the scenario we imagine* [Gendler 2010; Brown and Fehige 2017], which may not comport with every stipulation. There can be a gap between what's imagined and what's stipulated.

To see how, take an example from Sen [1979: 473]:

considering states of affairs *a* and *b*, let *r* be a romantic dreamer and *p* a miserable policeman. In *b* the policeman tortures the dreamer; in *a* he does not. The dreamer has a happy disposition ... and also happens to be rich, in good health, and resilient, while the policeman is morose, poor, ill, and frustrated, getting his simple pleasures out of torturing. The utility values for *p* and *r* happen to be:

	<i>a</i> (no torture)	<i>b</i> (torture of <i>r</i> by <i>p</i>)
<i>r</i> 's utility	10	8
<i>p</i> 's utility	4	7

Intuitively, *a* is better than *b*. Does this show that a smaller change in well-being (10 to 8) is less important than a larger change (4 to 7), or that a state of affairs with 14 ‘utils’ is morally better than one with 15 ‘utils’?

No. That Sen stipulates that torture is less bad for the victim than it is good for the perpetrator (as a ‘simple pleasure’) does not guarantee that this is true in the scenario that we imagine; and our intuitions concern the imagined scenarios, not the stipulations themselves.

In Sen’s example, it may be obvious that the scenario imagined is unlikely to comport with the author’s stipulations. But that is not always so. We can have mistaken beliefs about what we imagine and intuit. Opacity of mind is commonplace [Carruthers 2011]. And many philosophers have persuasively argued that we falsely believe that some imagined scenarios fit the stipulations of thought experiments. Consider Kripke [1980: 150] on ‘the illusion that water might not have been hydrogen hydroxide’, or Woodward and Allman [2007: 185] on putative counterexamples to consequentialism that ‘stipulat[e] away ... considerations that would be present in real life’, or Weijers [2013: 22] on ‘thought experiments that stipulate features that are so unrealistic that we have not experienced anything like them’, resulting in information influencing ‘intuitions’ in a manner that is ‘contrary to the point of the experiment itself’. Where we falsely believe that what we imagine comports with what was stipulated, intuitions about thought experiments are potentially misleading.

Is this true of Crisp’s example? Might what’s imagined fail to comport with his stipulated linear well-being measurements? Yes. The first two arguments from [section 3](#) suggest that default scales for well-being are plausibly non-linear. This could undermine the point of the thought experiment.

To see how, consider the possibility that reported well-being on the relevant scale is a concave function of actual well-being, as in [Figure 1](#):

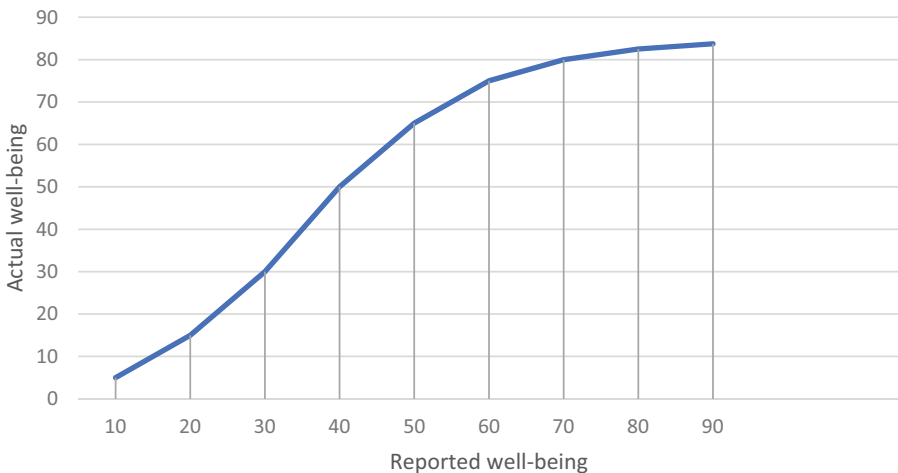


Figure 1: Diminishing marginal returns on reported well-being

In the scale above, ‘10’+‘90’ units of *reported* well-being translates to 5+83.75 (88.75) units of *actual* well-being, which is less than ‘50’+‘50’ (130). On any scale like this, utilitarians have decisive reason to prefer *Equality*. So, ‘a strong case for *Equality* over *Inequality*’ needn’t indicate that ‘equality is itself to be preferred’, *contra* Crisp [2003: 746].

Do people use a scale like the one in *Figure 1* in Crisp’s example? Perhaps. Scales for actual well-being measurements are explicitly bounded on both ends (for example, from ‘0’ to ‘10’). Crisp’s scale is not. Implicitly, it may have a lower bound, but one that lacks a verbal ‘anchor’ (for example, ‘not happy at all’, or ‘completely unhappy’). This makes it hard to determine what a unit on the scale represents [Kahneman and Sugden 2005]. And it allows for interpretations wherein the scale is taken to have a *negative* lower bound (‘completely unhappy’) and no upper bound, resulting in the compression of the lower section of the scale, but not the upper portion of the scale. This would produce a scale roughly like the one above; many scales like this would skew results, undermining the thought experiment.

Worse yet, if what we imagine fails to comport with Crisp’s stipulated linear well-being measurements, we might not believe that this is so. Plausibly, we would falsely believe that the scale is linear because it is represented with numerals that have linear properties. That 50+50 = 10+90 makes it easy to assume that total well-being in *Equality* and *Inequality* are equivalent, even when that assumption is false.

This is critical. Crisp’s seemingly innocuous thought experiment is potentially misleading. We could *believe* that it elicits intuitions about a scenario that are inconsistent with utilitarianism; but that belief may well be false, because what’s imagined may not comport with what’s stipulated, without this being introspectively obvious because of his use of numerals.

Not all appeals to hypothetical well-being measurements are potentially misleading. Crisp’s scenario, like many thought experiments, may be problematic because it ‘is inadequately described’ [Wilkes 1988: 8]. Perhaps other thought experiments in the literature are adequately described. Consider an example from Nagel [1979: 123–4]. A parent has two children—one healthy, one unhealthy. She could move to a city where the second child will receive treatment, or move to the suburbs where the first child will flourish. The former gives the healthy child a greater benefit, but the unhealthy child is worse off. When Derek Parfit discusses this example, he ‘use[s] figures’:

	<i>healthy child</i>	<i>unhealthy child</i>
<i>move to the city</i>	20	10
<i>move to the suburb</i>	25	9

Parfit noted that ‘such figures misleadingly suggest precision.’ For the well-being measurements to be probative, we must assume that [2002: 83]

[e]ach extra unit is a roughly equal benefit, however well off the person who receives it. If someone rises from 99 to 100, this person benefits as much as someone who rises from 9 to 10. Without this assumption we cannot make sense of some of our questions. We cannot ask, for example, whether some benefit would matter more if it came to someone who was worse off.

He's right. Without assuming linearity, why think that the change from 20 to 25 represents a greater benefit than the change from 9 to 10 does? This detail is crucial to the point of the example.

If this example is not misleading, we must be justified in believing that this detail is true of the scenario that we imagine. Any justification for doing so would come from the literal description from Nagel, not from the numeral figures that Parfit added. In other words, if the description is adequate to make the use of hypothetical well-being measurements *not* misleading, those measurements are redundant. So, why use them?

5. Actual Well-Being Measurements

What about appeals to actual well-being measurements? There is a 'growing conviction among psychologists and economists that people's happiness can be measured in sufficient detail for the results to be used in, for example, guiding governmental decisions' [de Boer 2014: 703]. Many argue at length for the use of well-being measurements in policy-making [Layard 2006; Diener et al. 2009], and even for the establishment of National Well-Being Accounts [Kahneman et al. 2004b]. While confidence in well-being measurements may be growing, it is not new. As Angner [2011a: 4] argues, subjective measures of well-being 'are part of an uninterrupted research stream going back at least to the 1920s and 1930s', which was always intended to guide public policy.⁷

It matters if we are not justified in believing that actual well-being measurements are linear. This affects whether and how such measurements can guide public policy and support conclusions about other philosophically significant issues, such as irrational biases. For brevity's sake, let's focus on the claim that our judgments and decisions are *systematically biased* in two ways [Kahneman 2011: 409]:

PEAK BIAS. We are subject to a 'bias that favors a short period of intense joy over a long period of moderate happiness' and a corresponding 'bias [that] makes us fear a short period of intense but tolerable suffering more than we fear a much longer period of moderate pain'.

END BIAS. We are 'prone to accept a long period of mild unpleasantness because the end will be better', and to 'giving up the opportunity for a long happy period if it is likely to have a poor ending.'

Kahneman combines these putative biases under one label (the 'Peak-End Rule'). I treat the two separately, as I will argue that (a) Kahneman's evidence might support END BIAS, but (b) his actual well-being measurements do not support PEAK BIAS, because we should not believe that these measurements are linear. To use well-being measurements to infer that PEAK BIAS is true, we need to know the magnitudes of the differences between 'peaks' on the scale.

Why focus on these two claims? First, because they are used to justify paternalistic interventions (Kahneman [1999: 15, 2011: 381]; cf. Broome [1996]). Second, because they are widely endorsed: Langer, Sarin, and Weber [2005: 157] identify 'the tendency to weigh the peak and the end of a sequence too heavily' as a 'systematic bias'. Many philosophers accept that PEAK BIAS and END BIAS are supported by '[r]obust

⁷ I am grateful to an anonymous referee for drawing my attention to this.

experimental evidence’ [Hales and Johnson 2014: 512–16]. This evidence supposedly reveals a systematic ‘distorting effect of memory’ [Tiberius 2006: 498; emphasis mine]:

It turns out that in assessing past painful experiences ... we tend to follow the Peak End Rule. That is, in retrospective assessments of pain we put more weight on *the worst part* and *the very end* of the experience.

Third, because even the philosophical criticism that Kahneman’s work has received almost uniformly targets his theory of well-being.⁸ The only objection to his aggregative methodology comes from de Boer [2014: 715]: ‘affective space is not bipolar’, so it is ‘unwarranted’ to calculate ‘wellbeing by subtracting negative emotion scores from positive ones.’ Kahneman needn’t assume that affective space is bipolar, or add units of pleasure to units of pain, to infer PEAK BIAS (or END BIAS). But he must add units of pleasure (pain) to units of pleasure (pain) *as if they are linear* to infer PEAK BIAS; that this is problematic has gone unnoticed.

Finally, PEAK BIAS and END BIAS provide a helpful comparison in exploring the limits of what we can learn from actual measurements of well-being in the absence of empirical information about whether scales are linear. If we can infer END BIAS but not PEAK BIAS, those limits are neither non-trivial nor so extensive that well-being measurement is pointless.

5.1 The Cold Pressor Experiment

Let’s start with the ‘cold pressor’ experiment [Kahneman et al. 1993], which is standardly used to illustrate END BIAS. Subjects were exposed to two painful experiences: first, one hand was immersed in water at 14 degrees Celsius for 60 seconds; second, the other hand was immersed in water at 14 degrees for 60 seconds, then kept in the water for 30 seconds longer as its temperature was raised to 15 degrees. Subjects rated the second condition as better than the first, and chose to repeat it rather than the first. As Kahneman and Frederick [2002: 79] note, these judgments and choices ‘violate dominance’: the second condition shares all of the bad features of the first, plus an additional period of pain. This suggests an irrational disposition to choose more pain because the end is better.

Other experimental evidence involving measurements of pleasure or pain can also support END BIAS,⁹ without assuming linearity. To see how, consider the following hypothetical measurements on the Richter scale:

	Region A	Region B
<i>Earthquakes in 2016</i>	5, 5, 5, 4, 4, 3	5, 5, 5, 4, 4

The total energy released by earthquakes in 2016 was greater in A than B. We can know this via dominance reasoning. If people systematically judged that sequences like

⁸ See, *inter alia*, Beardman [2000], Kelman [2005], Alexandrova [2008, 2012], Barrotta [2008], Feldman [2010: ch. 3], Hausman [2010], and Angner [2011b].

⁹ See Kahneman [1997: 386]. This is in line with by Kahneman’s method of ‘confirming judgmental biases’ via ‘comparisons of subjective happiness to independent assessments of objective happiness’ [1999: 22, 19 and references therein].

B's (which 'ended' on a worse earthquake: 4, rather than 3) were worse than sequences like *A*'s, this would suggest an 'end bias'. That the Richter scale is logarithmic would be irrelevant.

5.2. The Colonoscopy Experiment

The 'colonoscopy' experiment [Redelmeier and Kahneman 1996] can illustrate PEAK BIAS. While patients underwent colonoscopies,

[they] were prompted every sixty seconds to report the intensity of their current pain. They were to use a scale [from 0 to 10] where 10 was 'intolerable pain' and 0 was 'no pain at all'.¹⁰

In discussing this experiment, Kahneman often presents 'raw data' from a representative pair of patients, A and B, with graphs like the following (Figure 2):¹¹

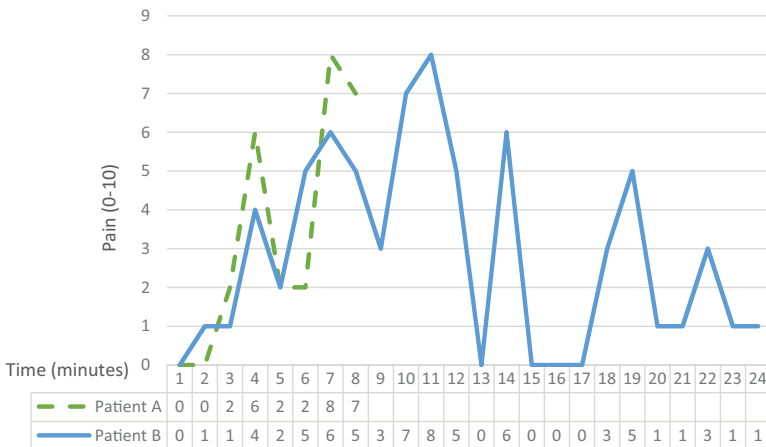


Figure 2: Kahneman's representative pair of patients

The x-axis represents the duration of the procedure; the y-axis represents patients' real-time reports of pain intensity on the aforementioned scale. Based on these data, Kahneman [2011: 379] asks 'an easy question':

which patient suffered more? No contest. There is general agreement that patient B had the worse time. Patient B spent at least as much time as Patient A at any level of pain, and the 'area under the curve' is clearly larger for B than for A.¹²

The final sentence offers two ways of answering the 'easy question'. The first appeals to dominance reasoning: 'B spent at least as much as A at any level of pain', as well as

¹⁰ This explanation of the methodology, from Kahneman [1999: 4], omits and simplifies some details from Redelmeier and Kahneman [1996: 4]. Nothing hangs on this.

¹¹ See Redelmeier and Kahneman [1996: 4] and Kahneman [1999: 4, 2011: 379]. I have combined his two graphs into one that is easier to read.

¹² Interestingly, Kahneman's [1999: 6] view suggests a different answer: the profiles differ in 'length', and 'the average height' for A is higher than for B.

16 minutes longer in pain. This is analogous to the Cold Pressor experiment. And, for that reason, it cannot support PEAK BIAS: A and B had the same peak pain intensity (8).

To support PEAK BIAS, the data must involve different peaks. To illustrate this, imagine a fictitious third patient, C, whose reports are similar to A's (Figure 3):

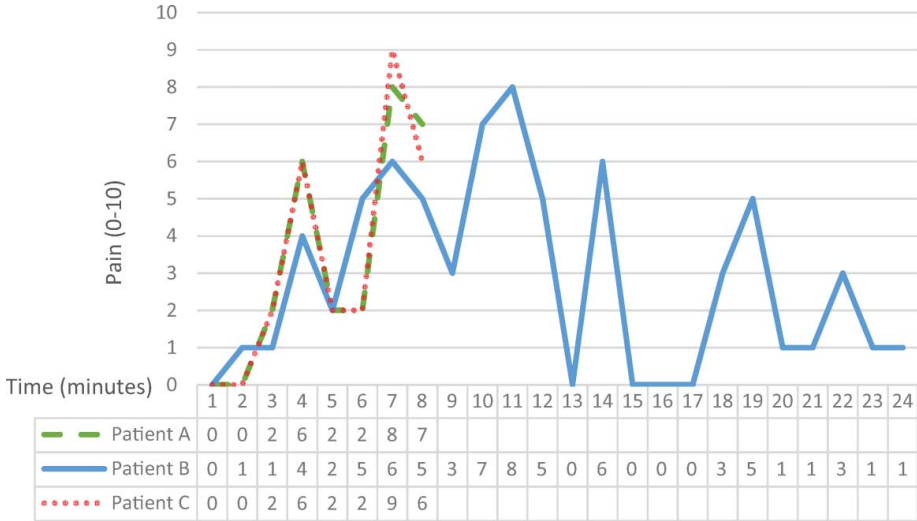


Figure 3: Adding 'Patient C'

Who suffered most? We cannot appeal to dominance reasoning, as C's profile has a higher peak (9). But we could appeal to Kahneman's second way of answering the question—namely, considering the 'area under the curve'. This is the method that Kahneman [ibid.: 83] uses to measure 'objective happiness': 'For an objective observer evaluating the episode from reports of the experiencing self, what counts is the 'area under the curve' that integrates pain over time; it has the nature of a sum.'

If we follow Kahneman's method, treating the measurements as linear, B experienced more pain (68.1) than A or C did (27.2). However, if instead we treat the measurements as ordinal, we cannot add up the units. And if we treat them as logarithmic, C experienced more pain: C's peak (one minute at 9) would represent *ten times more pain* than B's (one minute at 8). To cancel out this difference, B would need to have endured the equivalent of 9 more minutes at 8, or 90 more minutes at 7, or 900 more minutes at 6.

Because C's profile has a higher peak, we cannot aggregate the data without *some* assumption about the magnitude of the differences between peaks. And the case for PEAK BIAS goes through only if we assume linearity in particular—that is, that the difference between 9 and 8 is the same as between 7 and 8, and so on.¹³

¹³ There is further evidence that Kahneman assumes linearity. Redelmeier and Kahneman [1996: 5] report *mean values* for pain—e.g. 'Average Pain' (3.1)—in the same way as they report mean values for duration in minutes [ibid.: 23]. Kahneman [2011: 380] appeals to the 'average of the level of pain reported at the worst moment of the experience and at its end' ('Peak-End Rule'), which would be identical for A and C (7.5).

This is a problem, because we are not justified in assuming linearity. Indeed, there is reason to believe that this scale is nearer to logarithmic. The scale is from ‘0’ (‘no pain at all’) to ‘10’ (‘intolerable pain’). A patient should report ‘1’ as soon they experience *some* pain, however slight: a scratch on one’s finger is more than no pain at all. But, since the scale has an upper bound, ‘10’ should be reserved for agony, which is worse than the equivalent of ten scratches on one’s finger. This suggests that a greater magnitude of pain is needed to move from 9 and 10 than from 0 and 1. And the reasoning iterates: it requires increasingly larger magnitudes of differences between intervals. Our default scale for pain, like for loudness, may be nearer to logarithmic than to linear; in which case, our systematic emphasis on peaks is not irrational, *contra* Kahneman.

This result partially undermines Kahneman’s view that human decision-making is *systematically* irrational. Experimental evidence may support END BIAS. But it is uncommon. Reported *peak* levels of pain have more predictive power than *end* levels of pain in the colonoscopy experiment [Redelmeier and Kahneman 1996: 6]. This is not unusual. In many studies, ‘peak affect emerged as the best predictor of global evaluations’, such that ‘end affect was no longer a significant predictor’ [Fredrickson 2000: 581]. So, the data’s implications for understanding decision-making or justifying paternalistic policies are underwhelming.

6. Implications for Social Science

What implications does this discussion have for social scientific research on well-being? My objection to Kahneman’s case for PEAK BIAS is that (a) it depends upon the assumption of linearity, which (b) is empirically unsupported. How far does this generalize? Do (a) and (b) hold for alternatives to Kahneman’s method for measuring well-being? These issues warrant more attention than I can give them here, as the many methodologies for measuring well-being are complex and heterogeneous.

Regarding (a), many but not all measurements of well-being assume linearity. In interpreting answers on well-being surveys, ‘psychologists have by and large interpreted the answers as cardinal, i.e. that the difference in happiness between a 4 and a 5 for any individual is the same as between an 8 and a 9 for any other individual’; by contrast, ‘cardinality is still considered very suspect’ in similar research in economics [Ferrer-i-Carbonell and Frijters 2004: 641]. With that said, many prominent studies in economics ‘compare aggregates of satisfaction over countries and hence also implicitly rely on cardinality’ [ibid.: 646].

It is sometimes argued that such studies’ conclusions—for example, about the relationship between well-being and income [Layard et al. 2008]—can be supported without assuming linearity. Ferrer-i-Carbonell and Frijters [2004: 642] argue that assuming linearity generally ‘makes little difference to the results’. Why think this? Because, as Diener and Tov [2012: 145] explain,

the use of nonparametric ordinal statistics to treat well-being data has typically not led to different conclusions from those based on parametric statistics that assume equal scale intervals. More research in this area is needed.

This use of nonparametric statistics is no panacea (for general discussion, see [Michell 2009: 45] and references therein). And, as Diener and Tov note, in occasional cases the assumption of linearity leads ‘to altered conclusions’ [2012: 145]; this holds for

Kahneman's work, as I argued above.¹⁴ But more research is needed to determine how widespread this problem is.

What about (b)? Various methods for adducing empirical evidence about the intervals on well-being scales have been proposed. Here's one [Eich et al. 1999: 161]:

a subject is instructed to squeeze a hand-grip dynamometer, then the loudness of a tone is adjusted to correspond to the verbal pain descriptor, such as 'very intense', 'weak', and so forth. The scores derived from the hand-grip dynamometer and tone loudness are then plotted on a log-log scale to produce a numerical scale or magnitude estimation for each descriptor. In this way, meaningful statements can be made about the relative magnitudes of different pain descriptors.

For our purposes, all that we need to know about such approaches is that 'they are time-consuming and tedious to develop.' Since '[e]ase of scale construction and use are critically important for clinical applications', such methods are rarely employed [ibid.]. So, in principle, empirical evidence regarding whether well-being measurements are linear may be available; but in practise it is not gathered when well-being is measured.

The same point applies to a proposal from Kahneman [1999]. He is by no means oblivious to the issue that well-being scales could be non-linear, or that whether they are linear is an empirical matter.¹⁵ Indeed, he acknowledges that, in measuring objective happiness, 'the intervals may be arbitrary: a pain rating of 7 is reliably worse than a rating of 6, but the interval between 7 and 6 need not be psychologically equivalent to the interval between 3 and 2' [ibid.: 5]. He argues that 'a consistent rescaling is possible, yielding a ratio scale for instant utility that is calibrated by its relation to duration' via what he calls 'temporal integration'; however, this is 'a theoretical possibility, not a practical procedure' [ibid.: 6].

This is odd. According to Kahneman, we can treat the 'original profiles' of Patients A and B as linear 'only after a rescaling that incorporates a judgment about the equivalence of intensity and duration' [ibid.]. But he treats such measurements as linear without rescaling them. Other psychologists have followed his lead. Kemp et al. [2008: 132] use Kahneman's methodology to 'measure of the total happiness experienced' on vacations by considering 'the sum of the happiness of all the different moments', without mentioning 'rescaling'. Fredrickson [2000: 585, 589] is a curious case: citing a range of studies, she says that there is 'empirical support' for 'biases and mistakes', including PEAK BIAS. Yet she grants that *for policy-makers* to 'aggregate [well-being measurements] they must convert them into a ratio scale' [ibid.: 598]. If *policy-makers* need to do this, why don't *psychologists* need to do it? There is a disconnection between the modest conclusion that temporal integration shows that 'the measurement of experienced utility should be viewed as a difficult technical problem, not a hopeless quest' [Kahneman et al. 1997: 394], and the confidence in ambitious conclusions of research that purport to measure 'experienced utility' linearly without any rescaling.

Other proposals are likely to encounter similar problems. A procedure that allows us to test whether well-being measurements are linear is unlikely to be suitable for clinical

¹⁴ Redelmeier and Kahneman [1996: 4] use parametric Pearson correlation statistics.

¹⁵ Notably, Kahneman et al. [1997: 393] provide a representation theorem that shows that *if* certain axioms obtain, there is 'a suitable monotonic transformation of instant utility (and disutility) to ratio scales ... with the same zero point.' But the representation theorem does not tell us *whether* the axioms hold for (say) Patient C, or *what* the appropriate monotonic transformation of C's utility scores would be. We need empirical evidence to rescale C's original profile: hence the procedure in Kahneman [1999].

applications. This fits the complaint from Michell, Borsboom, and Mellenbergh: in psychological research, scores are often summed and declared to be measurements of an attribute, without any attempt to justify this conclusion empirically.

Some might object that empirical evidence is not required here. In so far as it is problematic to assume that well-being measurements are linear, this problem solves itself. What makes researchers prone to *interpreting* well-being scales as if these are linear also makes subjects *use* numerical well-being scales as if these are linear. In both cases, well-being scales are represented with devices like numerals, which makes it easy to treat intervals on the scale as if they have the linear properties of numbers. If the scale is used and interpreted linearly, the problem solves itself.¹⁶

Both steps of this objection are problematic. Regarding the first, well-being scales are represented differently when used and interpreted. When such scales are used, intervals on the scale are often accompanied by verbal labels. On a life satisfaction scale from 1 to 7, ‘1’ represents the proposition ‘In general, I consider myself not a very happy person’ and ‘7’ represents the proposition that ‘In general, I consider myself a very happy person’ [Lyubomirsky and Lepper 1999: 151]. On some well-being scales, *every* interval has a verbal label like ‘very happy’, ‘quite happy’, ‘not very happy’ and ‘not at all happy’ (see OECD [2013: 82–5]). But when the scale is interpreted, these intervals are ‘coded as numbers’, ignoring verbal labels [ibid.: 173]. This difference in how scales are represented when used and interpreted is important.¹⁷ Chapman et al. [1985: 10] find it ‘questionable’ to treat well-being scales as if they are linear ‘unless the investigator has evidence that subjects treat the categories like equally spaced numbers or ... non-parametric ranking statistics are employed’. In part, this is because empirical evidence suggests that, for subjects, ‘scale items are not equally spaced when labeled with words commonly used to describe pain’ [ibid.]. As Michell [2009a: 44–5] argues, this militates against interpreting scales linearly: ‘each datum is not an isolated number, it is a proposition’, and ‘[t]abulated numbers are shorthand for a set of propositions.’

What about the second step? If the scale is represented in the same way when used and when interpreted, will it be treated in both cases as if it is linear? Not necessarily. When *interpreting* a scale, the tasks include making inferences about aggregates, which requires judgments about magnitudes of differences between intervals. This makes the linear properties of numbers salient. The tasks involved in *using* scales are different. On Kahneman’s approach, subjects report levels of pain at particular points in time. Assigning a value to a discrete experience is a different task from assigning a value to a set thereof, as the latter is not directly experienced, but is instead constructed [1999: 15].

A final problem with the objection is that it proves too much. Empirical evidence shows that, when we *use* numerical scales to report phenomena like loudness and brightness, we treat them as if they are non-linear. (This is true even when we are asked to report magnitudes of equal distance.)¹⁸ But evidence suggests that if we are later asked to *interpret* such measurements, we treat them as if they are linear. Kahneman discusses cases in which judgments of the total brightness experienced over time are

¹⁶ I am grateful to Ralf Bader for raising this in personal communication.

¹⁷ A similar concern applies to the use of verbal labels for intervals on Likert scales. See Wu and Leung [2017] for discussion of whether such scales can be treated as linear.

¹⁸ See O’Shaughnessy [1987: 150] on the logarithmic mel scale for pitch.

determined by the *numerical average* of the perceived brightness of the peak and the end [ibid.: 15].

In sum, more work is needed to determine how often well-being measurements require the assumption of linearity; but, where they do, we should require empirical evidence supporting this assumption, and no practical procedure for providing that evidence has been established.


7. A Solution

Well-being measurements are frequently used to support philosophically important conclusions. But these conclusions often rest on the unwarranted assumption of linearity. This problem is easy to miss because of how well-being scales are represented.

Here's a tentative solution to this problem. When we do not have evidence to support linearity, change how the scale is represented: replace numerals with letters. Letters encode a rank ordering from lower to higher (a, b, c) without surplus structure—they encode no information about the magnitudes of differences between intervals.

It is hard to see why this solution should be opposed when well-being measurements are used in philosophical thought experiments. There may be more resistance from practitioners of social science. But the proposal does not place impossible constraints on such research. The main justification for using numerals when measuring well-being is that numerals are easier for subjects to remember than sequences of verbal descriptions are [OECD 2013: 82–4]. That applies equally to letters. And researchers can still (a) draw inferences that do not assume linearity, and/or (b) empirically test for linearity. What they cannot responsibly do is draw inferences that depend on an untested assumption of linearity, then represent such measurements with devices that suggest linearity.¹⁹

ORCID

Daniel Wodak  <http://orcid.org/0000-0001-8797-1106>

References

- Alexandrova, Anna. 2008. First-Person Reports and the Measurement of Happiness, *Philosophical Psychology* 21/5: 571–83.
- Alexandrova, Anna. 2012. Well-Being as an Object of Science, *Philosophy of Science* 79/5: 678–89.
- Angner, Erik. 2011a. The Evolution of Eupathics: The Historical Roots of Subjective Measures of Well-being, *International Journal of Wellbeing* 1/1: 4–41.
- Angner, Erik. 2011b. Are Subjective Measures of Well-Being ‘Direct’? *Australasian Journal of Philosophy* 89/1: 115–30.
- Barrotta, Pierluigi. 2008. Why Economists Should Be Unhappy with the Economics of Happiness, *Economics & Philosophy* 24/2: 145–65.
- Bartleson, C.J. and E.J. Breneman. 1967. Brightness Perception in Complex Fields, *Journal of the Optical Society of America* 57/7: 953–57.
- Beardman, Stephanie. 2000. The Choice between Current and Retrospective Evaluations of Pain, *Philosophical Psychology* 13/1: 97–110.

¹⁹ Thanks to Ralf Bader, Guy Fletcher, Gil Hersch, Ben Jantzen, Lydia Patton, Govind Persad, Kelly Trogdon, participants at Georgetown University’s Workshop on Methodology in Applied Ethics, and two anonymous referees.

- Borsboom, Denny, and Gideon Mellenbergh. 2004. Why Psychometrics Is Not Pathological: A Comment on Michell, *Theory & Psychology* 14/1: 105–20.
- Broome, John. 1996. More Pain or Less? *Analysis* 56/2: 116–18.
- Brown, James, and Yiftach Fehige. 2017. Thought Experiments, *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. URL = <https://plato.stanford.edu/archives/sum2017/entries/thought-experiment>
- Carruthers, Peter. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford: Oxford University Press.
- Chapman, C.R., K.L. Casey, R. Dubner, K.M. Foley, R.H. Gracely, and A.E. Reading. 1985. Pain Measurement: An Overview, *Pain* 22/1: 1–31.
- Crisp, Roger. 2003. Equality, Priority, and Compassion, *Ethics* 113/4: 745–63.
- de Boer, Jelle. 2014. Scaling Happiness, *Philosophical Psychology* 27/5: 703–18.
- Diener, Ed, R. Lucas, U. Schimmack, and J. Helliwell. 2009. *Well-Being for Public Policy*, Oxford: Oxford University Press.
- Diener, Ed, and William Tov. 2012. National Accounts of Well-Being, in *Handbook of Social Indicators and Quality-of-Life Research*, ed. K.C. Land, A.C. Michalos, and M.J. Sirgy, Heidelberg: Springer Dordrecht: 137–57.
- Eich, Eric, I.A. Brodtkin, J.L. Reeves, and A.F. Chawla. 1999. Questions Concerning Pain, in *Well-Being: Foundations of Hedonic Psychology*, ed. Daniel Kahneman, Ed Diener, and Norbert Schwarz, New York: Russell Sage Foundation: 155–68.
- Fechner, Gustav Theodor. 1966. *Elements of Psychophysics, Volume I*, ed. Davis H. Howes and Edwin G. Boring, trans. Helmut E. Adler, New York: Holt, Rinehart and Winston.
- Feldman, Fred. 2010. *What Is This Thing Called Happiness?* Oxford: Oxford University Press.
- Ferrer-i-Carbonell, Ada, and Paul Frijters. 2004. How Important Is Methodology for the Estimates of the Determinants of Happiness? *The Economic Journal* 114/497: 641–59.
- Fredrickson, Barbara. 2000. Extracting Meaning from Past Affective Experiences: The Importance of Peaks, Ends, and Specific Emotions, *Cognition and Emotion* 14/4: 577–606.
- Gelfand, Stanley. 2009. *Hearing: An Introduction to Psychological and Physiological Acoustics, 5th edn.*, London: Informa Healthcare.
- Gendler, Tamar Szabó. 2010. *Intuition, Imagination, and Philosophical Methodology*, Oxford: Oxford University Press.
- Hales, Steven D., and Jennifer Adrienne Johnson. 2014. Luck Attributions and Cognitive Bias, *Meta-philosophy* 45/4–5: 509–28.
- Hausman, Daniel M.. 2010. Hedonism and Welfare Economics, *Economics & Philosophy* 26/3: 321–44.
- Kahneman, Daniel. 1999. Objective Happiness, in *Well-Being: Foundations of Hedonic Psychology*, ed. Daniel Kahneman, Ed Diener, and Norbert Schwarz, New York: Russell Sage Foundation: 3–25.
- Kahneman, Daniel. 2011. *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux.
- Kahneman, Daniel, and Shane Frederick. 2002. Representativeness Revisited: Attribute Substitution in Intuitive Judgement, in *Heuristics and Biases: The Psychology of Intuitive Judgment*, ed. Thomas Gilovich, Dale Griffin, and Daniel Kahneman, Cambridge: Cambridge University Press: 49–81.
- Kahneman, Daniel, B.L. Fredrickson, C.A. Schreiber, and D.A. Redelmeier. 1993. When More Pain Is Preferred to Less: Adding a Better End, *Psychological Science* 4/6: 401–5.
- Kahneman, Daniel, A.B. Krueger, D.A. Schkade, N. Schwarz, and A.A. Stone. 2004a. A Survey Method for Characterizing Daily Life Experience, *Science* 306/5702: 1776–80.
- Kahneman, Daniel, A.B. Krueger, D.A. Schkade, N. Schwarz, and A.A. Stone. 2004b. Toward National Well-Being Accounts, *American Economic Review* 94/2: 429–34.
- Kahneman, Daniel and Robert Sugden. 2005. Experienced Utility as a Standard of Policy Evaluation, *Environmental and Resource Economics* 32/1: 161–81.
- Kahneman, Daniel, P.P. Wakker and R. Sarin. 1997. Back to Bentham? Explorations of Experienced Utility, *The Quarterly Journal of Economics* 112/2: 375–406.
- Kelman, Mark. 2005. Hedonic Psychology and the Ambiguities of ‘Welfare’, *Philosophy & Public Affairs* 33/4: 391–412.
- Kemp, Simon, C.D.B. Burt, and L. Furneaux. 2008. A Test of the Peak-End Rule with Extended Autobiographical Events, *Memory & Cognition* 36/1: 132–8.
- Kripke, Saul A. 1980. *Naming and Necessity*, Cambridge, MA: Harvard University Press.

- Langer, Thomas, R. Sarin, and M. Weber. 2005. The Retrospective Evaluation of Payment Sequences: Duration Neglect and Peak-and-End Effects, *Journal of Economic Behavior & Organization* 58/1: 157–75.
- Layard, Richard. 2006. *Happiness: Lessons from a New Science*, New York: Penguin.
- Layard, Richard, G. Mayraz, and S. Nickell. 2008. The Marginal Utility of Income, *Journal of Public Economics* 92/8–9: 1846–57.
- Lyubomirsky, Sonja, and Heidi Lepper. 1999. A Measure of Subjective Happiness: Preliminary Reliability and Construct Validation, *Social Indicators Research* 46/2: 137–55.
- Michell, Joel. 2004. Item Response Models, Pathological Science and the Shape of Error: Response to Borsboom and Mellenbergh, *Theory & Psychology* 14/1: 121–9.
- Michell, Joel. 2009. The Psychometricians' Fallacy; Too Clever by Half, *British Journal of Mathematical and Statistical Psychology* 62/1: 41–55.
- Michell, Joel. 2012. 'The Constantly Recurring Argument': Inferring Quantity from Order, *Theory & Psychology* 22/3: 255–71.
- Myles, Paul S., S. Troedel, M. Boquest, and M. Reeves. 1999. The Pain Visual Analog Scale: Is It Linear or Nonlinear? *Anesthesia & Analgesia* 87/6: 1517–20.
- Nagel, Thomas. 1979. *Mortal Questions*, Cambridge: Cambridge University Press.
- Nozick, Robert. 1981. *Philosophical Explanations*, Cambridge, MA: Harvard University Press.
- Organisation for Economic Co-operation and Development (OECD). 2013. *OECD Guidelines on Measuring Subjective Well-Being*, Paris: OECD Publishing.
- O'Shaughnessy, Douglas. 1987. *Speech Communication: Human and Machine*, Reading, MA: Addison-Wesley.
- Otsuka, Michael. 2015. Prioritarianism and the Measure of Utility, *Journal of Political Philosophy* 23/1: 1–22.
- Parfit, Derek. 2002. Equality or Priority?, in *The Ideal of Equality*, ed. Matthew Clayton and Andrew Williams, Basingstoke: Palgrave: 81–125.
- Pinoli, Jean-Charles. 1997. The Logarithmic Image Processing Model: Connections with Human Brightness Perception and Contrast Estimators, *Journal of Mathematical Imaging and Vision* 7/4: 341–58.
- Poulton, E.C., R.S. Edwards, and T.J. Fowler. 1980. Eliminating Subjective Biases in Judging the Loudness of a 1-kHz Tone, *Perception and Psychophysics* 27/2: 93–103.
- Redelmeier, D.A. and D. Kahneman. 1996. Patients' Memories of Painful Medical Treatments: Real-Time and Retrospective Evaluations of Two Minimally Invasive Procedures, *Pain* 66/1: 3–8.
- Sen, Amartya. 1979. Utilitarianism and Welfarism, *The Journal of Philosophy* 76/9: 463–89.
- Setiya, Kieran. 2012. *Knowing Right from Wrong*, Oxford: Oxford University Press.
- Shafer, Karl. 2014. Knowledge and Two Forms of Non-Accidental Truth, *Philosophy and Phenomenological Research* 89/2: 373–93.
- Stevens, S.S. 1959. Measurement, Psychophysics and Utility, in *Measurement: Definitions and Theories*, ed. C.W. Churchman and P. Ratoosh, New York: John Wiley & Sons: 18–63.
- Tiberius, Valerie. 2006. Well-Being: Psychological Research for Philosophers, *Philosophy Compass* 1/5: 493–505.
- Weijers, Dan. 2013. Intuitive Biases in Judgments about Thought Experiments: The Experience Machine Revisited, *Philosophical Writings* 41/1: 17–31.
- Wilkes, Kathleen V. 1988. *Real People: Personal Identity without Thought Experiments*, Oxford: Clarendon Press.
- Woodward, James and John Allman. 2007. Moral Intuition: Its Neural Substrates and Normative Significance, *Journal of Physiology-Paris* 100/4–6: 179–202.
- Wu, Huiping and Shing-On Leung. 2017. Can Likert Scales Be Treated as Interval Scales? *Journal of Social Service Research* 43/4: 527–32.