Scales over our eyes

Daniel Wodak
Virginia Tech—dwodak@vt.edu

I—A Tale of Two Fallacies

Some things are hard to measure with precision. With mental states, as with crowd sizes, it is often easy enough to know that *A* is greater than *B*, even though it very hard to know the magnitude of the difference between *A* and *B*. We can know that Obama's 2009 Inauguration drew a larger crowd than Trump's 2016 Inauguration, and perhaps we can also know that Michelle is happier than Melania, even when we are ignorant of *how much* larger the crowd was at Obama's Inauguration, and also ignorant of *how much* happier Michelle is than Melania.

At this point it will be helpful to introduce some terminology about measurements and scales. When our best measurements tell us that *A* is greater than *B* (and *B* greater than *C*, and so on), but do not tell us the magnitude of the differences, an *ordinal scale* is appropriate. An ordinal scale is only a measure of the rank order: Michelle's happiness > Melania's happiness. If we knew the magnitude of the difference, an *interval scale* would be appropriate. With temperature, for instance, we know that the difference between 2 and 3 degrees Celsius is the same as the difference between 6 and 7, and is half the difference between 8 and 10.

The difference between ordinal and interval scales is significant for a range of reasons. For instance, when we only know rank order, we do not know how to aggregate data points. Only knowing that Michelle's happiness > Melania's happiness does not put us in a position to know the total or average of Michelle and Melania's happiness. If we are to aggregate psychological data about happiness, we need to be in a position to measure happiness on an interval scale.

While this is all widely known, it is also frequently ignored. Joel Michell coined the phrase "the psychometrician's fallacy" to describe the fallacious inference by psychometricians of an interval scale from an ordinal scale.[1] Michell has provided plenty of historical and contemporary examples of psychologists who measure some attribute on an ordinal scale, then treat the measurements as if they are on an interval scale, such that we can aggregate data points.

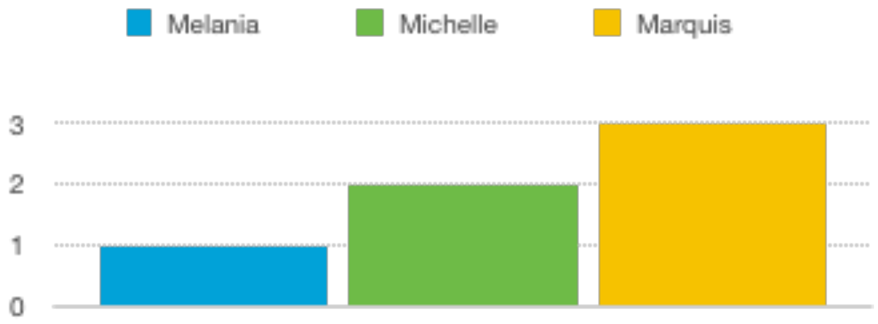
[1] See Joel Michell, 'The psychometricians' fallacy: Too clever by half?', *British Journal of Mathematical and Statistical Psychology*, 62 (2009) pp. 41–55, and references to earlier work on the same topic therein. I use this phrase with some hesitation for the following reason. I am concerned with an epistemic claim (about knowledge or justified beliefs regarding the magnitude of the difference between *A* and *B*). At times, Michell suggests that his true concern is a metaphysical claim (about whether there is a magnitude of the difference between *A* and *B*). He writes, for instances, that "the terms of the psychometricians' fallacy are these: the premise is the proposition that some attribute, *A*, is ordered (i.e. the kind of attribute admitting ordinal scaling); and the conclusion is the proposition that *A* is quantitative (i.e. the kind of attribute that *would be measurable* at least on an interval scale *were more known about it*)" (p. 42, emphasis mine)." Other remarks also suggest this interpretation: "Merely ordinal attributes do not suffer the ontologically crippling debility of not being able to stand alone in the light as real structures without the crutch of additive structure to lift them from the shadowy underworld of non-being" (p. 52). I have no commitments regarding whether *happiness itself* is quantitative or a "merely ordinal attribute".

The psychometrician's fallacy is a methodological problem in psychology. My interest here concerns how it relates to a methodological problem in practical and theoretical ethics.

One helpful way of framing the problem will involve seeing how the psychometrician's fallacy can be plausibly understood as a special instance of a more general fallacy which we might call "the representational fallacy".[2] For my purposes, the representational fallacy is 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*.

For illustrative purposes, imagine that we constructed an ordinal scale where we coded Michelle's level of happiness as "2" and Melania's level of happiness is "1". Now imagine someone who inferred from this that Michelle was twice as happy as Melania. That would be an understandable mistake. It is a mistake because data points on an ordinal scale do not put us in a position to know ratios. For that we would need a *ratio scale*: i.e., an interval scale on which there is a unique and non-arbitrary zero point, such as the scales for length and mass. But the mistake is understandable because numbers exist on a ratio scale on which two is twice as large as one. Hence the tempting albeit fallacious inference from the known premise about the salient feature of the representative device (the number two is twice as large as the number one) to the unknown and possibly false conclusion about the feature of the world represented (Michelle's happiness is twice as great as Melania's happiness).

We can also illustrate the representational fallacy with graphic representations. To make this point, let's add to our ordinal scale. Imagine someone who is extremely pretty happy with the state of the world. Call them "Marquis de Sade", or "Marquis" for short. Say we know that Marquis' happiness > Michelle's happiness, and we encode this on our scale by adding "3" for Marquis' level of happiness. Now we could graphically represent our ordinal scale as follows:



A salient feature of this graphic representation is the physical equivalence of distances between points ("1", "2", "3") on the *y* axis. From this, one might fallaciously infer that the magnitude of the difference between "1" and "2" is the same as the magnitude of the difference between "2" and "3". In other words, one might fallaciously infer an interval scale from a rank order.

[2] See Heather Dyke, *Metaphysics and the Representational Fallacy* (Routledge 2007). I use this phrase with some hesitation too, as Dyke offers more and less committal formulations of the fallacy, some of which are too tightly bound to *linguistic* representations—as opposed to, say, graphic representations—for my purposes.

We have now seen two ways in which common practices might put scales over our eyes. By representing ordinal measurements using numbers and graphs, we are easily led to infer that—like the representative devices—the features of the world represented have known intervals.[3]

So far I have explained the psychometrician's fallacy, and how it can be understood as a special instance of the representational fallacy. The former is clearly a methodological problem in psychology. But it is not yet obvious why either constitutes a methodological problem in ethics. In fact, the above forms the basis for two related methodological problems in ethics. Or at least, so I will contend. I will spend the bulk of the paper discussing the first, which concerns how actual measurements of wellbeing are used to support conclusions about issues in practical ethics (understood broadly to include conclusions about what outcomes are better for people, what we ought to do as individuals or policy-makers, and what decisions are rational). I close by considering the second problem, which concerns how hypothetical measurements of wellbeing are used to support conclusions in normative ethics (i.e., conclusions about the truth or plausibility of first-order normative theories, such as utilitarianism or prioritarianism).

II—Actual Measurements of Happiness

Francis Edgeworth was famously optimistic that in principle one could invent a "hedonimeter" which would "grant to the science of pleasure what is granted to the science of energy":

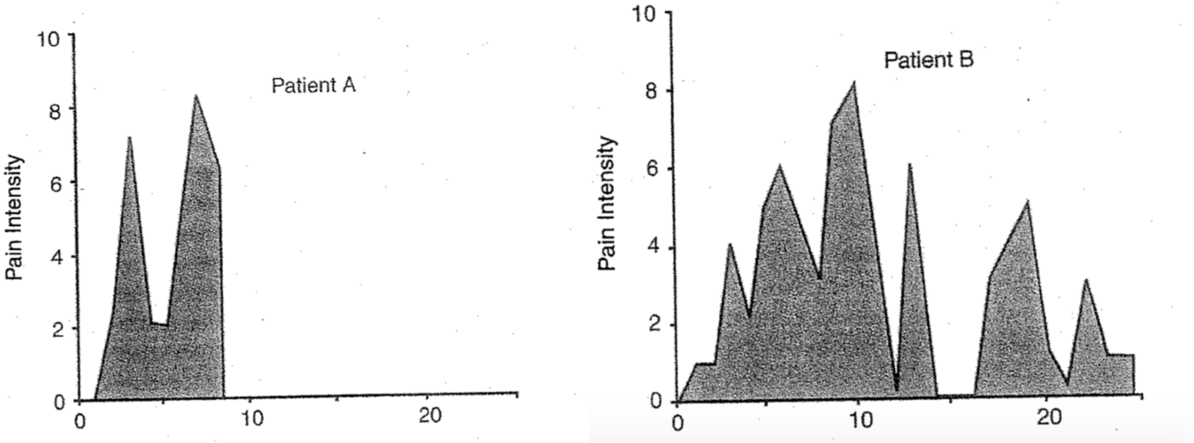
> an ideally perfect instrument, a psychophysical machine, continually registering the height of pleasure experienced by an individual, exactly according to the verdict of consciousness… From moment to moment the hedonimeter varies; the delicate index now flickering with the flutter of the passions, now steadied by intellectual activity, low sunk while hours in the neighbourhood of zero, or momentarily springing up towards infinity. The continually indicated height is registered by photographic or other frictionless apparatus upon a uniformly moving vertical plane. Then the quantity of happiness between two epochs is represented by the area contained between the zero-line, perpendiculars thereto at the points corresponding to the epochs, and the curve traced by the index.[4]

For a century, most were quite pessimistic that anything like this was possible even in principle.

[3] Michell's writing suggests the above connection to the representational fallacy, though to my knowledge he does not discuss it, and only focusses on numerical representations: "The salient point about ordinal scales is that if all that is known about an attribute is its ordinal structure, and if this is coded numerically, giving an ordinal scale, then conclusions inferred about the attribute from the numbers used that are not also validly entailed by the relevant, ordinal relations numerically unclothed are an artefact of the scale." He later continues: "each datum is not an isolated number, it is a proposition", and "[t]abulated numbers are shorthand for a set of propositions" ('The psychometricians' fallacy', pp. 44-45).

[4] Francis Edgeworth, *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences* (Kegan Paul and Co, 1881), p. 101.

The last few decades have seen a return to optimism, led by (among others) Daniel Kahneman.[5] Kahneman has even compared the methodology he champions to Edgeworth's hedonimeter.[6] This methodology has gone by various names as it has been developed and tweaked. At one point, Kahneman famously (albeit misleadingly) called it a measure of "objective happiness".[7] More accurately, it has been called Ecological Momentary Assessment. To illustrate the idea, let's briefly outline what became a fairly famous experiment. Donald Redelmeier and Kahneman recorded the pain reported by patients undergoing colonoscopies. The patients were prompted every sixty seconds to report the intensity of their current pain, on a scale of "0" ("no pain at all") to "10" ("intolerable pain"). Here are the graphed results for what they took to be a representative pair of patients (where the *x* axis shows duration in minutes):[8]



This is not a paper about colonoscopies. Nor is it about methodological problems in psychology. My concern here lies with how measurements of wellbeing are used to support conclusions about what outcomes are better for people, what we ought to do as individuals or policy-makers, and what decisions are rational. And the graphs above have been reproduced repeatedly in support of such conclusions in scholarly and popular work by Kahneman.

Kahneman's claims about what's good or better for people have been criticized, typically on the assumption that he takes happiness, wellbeing, and "experienced utility" to be equivalent, and identical to pleasure (and *mutatis mutandis* for their negative counterparts and pain).[9] For our

[5] Others have made similar assessments: "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" (Jelle de Boer, 'Scaling happiness', *Philosophical Psychology*, 27(5) 2014 p. 703).

[6] Daniel Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011), p. 392.

[7] Daniel Kahneman, "Objective Happiness" in *Well-Being: The Foundations of Hedonic Psycholog*y, Daniel Kahneman, Ed Diener, and Norbert Schwarz (eds.), (Russell Sage Foundation, 1999), pp. 3–25.

[8] Donald A. Redelmeier and Daniel Kahneman, 'Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures, *Pain,* 66 (1996), p. 4.

[9] It is important to recognize here that Kahneman seems to understand instant utility in terms of some conjunction of qualia and behavioral dispositions, at discrete points in time; philosophers object to different parts of this package. See especially Fred Feldman's 'Kahneman's "Objective Happiness"' (in *What is this Thing Called*

purposes we can sidestep that concern and treat pleasure as a plausible component of wellbeing (and *mutatis mutandis* and for pain and "ill-being"). That might even be Kahneman's considered view. In light of that, I take the following to be a question about wellbeing: "Assuming that the two patients used the scale of pain similarly, which patient suffered more?" Kahneman explicitly thinks claims that is "an easy question": "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."[10]

On this basis, Kahneman also endorses conclusions about what individuals and policymakers ought to do. Individuals should choose colonoscopies like Patient A's over ones like Patient B's. And policy-makers should force or nudge them to do so: "If the objective is to reduce the amount of pain actually experienced, conducting the procedure swiftly may be appropriate even if doing so increases the peak pain intensity and leaves patients with an awful memory."[11] Kahneman clearly believes that is an important policy objective: he takes an "implication of [his] analysis" to be "that the goal of policy should be to increase measures of objective wellbeing, not measures of satisfaction or subjective happiness".[12]

Kahneman's claims about what's better for people also form the basis for his conclusions about rationality. We are subject to a "bias" that "makes us fear a short period of intense but tolerable suffering more than we fear a much longer period of moderate pain."[13] This is because if we endured experiences like Patient A's and Patient B's, we would be likely to remember the former as worse than the latter even though this is false. Similar effects apply to other putative components of happiness or wellbeing: for instance, our decisions are subject to "a bias that favors a short period of intense joy over a long period of moderate happiness."

Happiness?, Oxford University Press, 2010), as well as A. Alexandrova, 'First-Person reports and the measurement of happiness', *Philosophical Psychology*, 21 (2008, pp. 571–583, P. Barotta, 'Why economists should be unhappy with the economics of happiness', *Economics and Philosophy*, 24 (2008), pp. 145–164, D. Hausman, 'Hedonism and welfare economics', *Economics and Philosophy*, 26 (2010), pp. 321–344, D. Haybron, *The pursuit of unhappiness: The elusive psychology of well-being* (Oxford University Press, 2008), and M. Kelman, 'Hedonic psychology and the ambiguities of welfare', *Philosophy and Public Affairs*, 33 (2005), pp. 391–412. This is by no means the only objection philosophers have offered to Kahneman. For instance, others have objected that, *contra* Kahneman, patients' reports only offer us *indirect* access to their mental states: Erik Angner, 'Are subjective measures of well-being 'direct'?', *Australasian Journal of Philosophy*, 89(1) (2011), pp. 115-130. The closest to the objection raised here comes from Jelle de Boer, who argues that we cannot aggregate units of pleasure and units of pain: "Positive and negative emotions are independent—affective space is not bipolar. Calculating affective wellbeing by subtracting negative emotion scores from positive ones … is therefore unwarranted" ('Scaling Happiness', p. 715). My objection concerns whether we can aggregate units of pain alone, or units of pleasure alone.

[10] Kahneman, *Thinking, Fast and Slow*, p. 379. It is surprising that Kahneman thinks that this answer is obviously right, because it is in tension with his more general view that when comparing periods of different durations the comparison point should be the average not the total ("Objective Happiness", p. 6); if we can aggregate the data in the way he suggests, the average level of pain for Patient A would be greater than the average for Patient B.

[11] Kahneman, *Thinking, Fast and Slow*, p. 381

[12] Kahneman, "Objective Happiness", p. 15. Kahneman also makes the slightly hedged claim that it "could be justifiable" to prolong colonoscopies unnecessarily at a low-level of pain to increase "patients' willingness to undergo further colonoscopies when their treatment required it" (p. 20). This is early *Nudge*-style reasoning.

[13] Kahneman, *Thinking, Fast and Slow*, p. 409

These memorial biases also explain why Kahneman objects to what he calls measures of satisfaction or subjective happiness, which are *ex post* global assessments of episodes.

An important observation about the representative pair of patients above is that they cannot be used to illustrate all of these claims. As Kahneman said, "Patient B spent at least as much time as patient A at any level of pain". There is no difference in the relevant *peak* pain intensity. The difference between them was "the bad luck of patient A that the procedure ended at a bad moment, leaving him with an unpleasant memory", which patient B was lucky enough to avoid.

A brief aside. It is worth explaining why this important observation has not been made. Kahneman has for over twenty years coupled together two separate effects which should be treated separately for several reasons: according to his "Peak-end rule" as it was "observed" in the colonoscopy experiment, "[t]he global retrospective rating was well predicted by the average of the level of pain reported at the worst moment of the experience and at its end."[14] As you can see, Kahneman is coupling together the predictive power of the worst moment (the peak) and the final moment (the end). Let's call these the Peak Rule and End Rule, respectively.

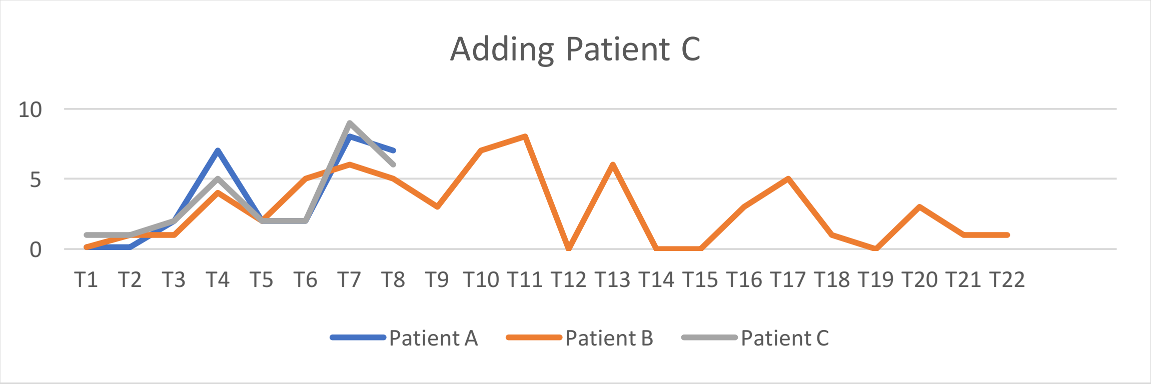
There are at least three reasons to de-couple the Peak Rule and End Rule which I will mention here, but not fully defend right now. First, in many contexts, the Peak Rule has far more predictive power than the End Rule.[15] This is obscured if we couple the two effects together. Second, Kahneman claims that both rules constitute biases that distort retrospective evaluations and subsequent decisions. Indeed, Kahneman believes that the psychological data itself provides evidence that both Rules are fact biases—that our decision-making is irrational in "its exaggerated emphasis of peaks and ends".[16] I think this is a mistake in relation to the Peak Rule, for reasons that do not apply to the End Rule. Third, since the pain measurements above are not on an interval scale, it is inappropriate to average "the level of pain reported at the worst moment of the experience and at its end", which the Peak-End Rule requires us to do.

Now back to our main thread. We're concerned with a methodological problem with using psychological data to support claims about what's good for us, what we ought to do, and what decisions are irrational. We've seen that for representative cases to be able to support Kahneman's claims about these issues would require a difference in peak pain. So let's introduce a fictitious third patient to the representative pair from the colonoscopy experiment:

[14] Kahneman, *Thinking Fast and Slow*, p. 380.

[15] For instance, Peak Pain has far more predictive power than End Pain in Redelmeier and Kahneman's colonoscopy experiment ('Patients' memories', p. 6). And in some studies, "peak affect emerged as the best predictor of global evaluations", such that "end affect was no longer a significant predictor" (Barbara L. Fredrickson, 'Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions', *Cognition and Emotion*, 14(4) 2000, p. 581); indeed, "sometimes a simpler "peak only" rule appears to apply", such as "when people are bracing themselves for the anticipated re-experience of an aversive episode" (p. 588).

[16] Kahneman, *Thinking, Fast and Slow*, p. 409. This judgment seems to be widely shared. For instance, "the tendency to weigh the peak and the end of a sequence too heavily" is listed as a "systematic bias" by Thomas Langer et al. in 'The retrospective evaluation of payment sequences: duration neglect and peak-and-end effects', *Journal of Economic Behavior & Organization*, 58 2005, p. 157.



Now let's turn back to Kahneman's main claims from before. Assuming that the three patients used the scale of pain similarly, which patients suffered most? Is this still an easy question? It's certainly not as easy as before. Patient B did not spend at least as much time as patient C at any level of pain.[17] But you might still think that it is easy enough. Patient B had the worst time; the total pain that B experienced over the longer duration of her colonoscopy is greater than what Patient A or Patient C experienced over their comparatively short colonoscopies. As Kahneman would say, the area under the curve is clearly larger for Patient B than for Patient A or Patient C.

However, this answer makes a significant assumption. It assumes that the patients' reported levels of pain are on an interval scale, and that we know the magnitudes of the differences. We've already seen why this assumption would be tempting given the representative devices used by Kahneman. The use of numerals for levels of pain makes it easy to assume that the difference between "2" and "3" will be the same as the difference between "6" and "7" *units of pain*. Notably, Kahneman repeatedly encourages this way of thinking about the relevant scale: "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."[18] Such sums are easy. We simply add up the integers to reach the total. The graphic representation of the scale—and the linguistic references to the "area under the curve"—also suggest that we can and should simply add up the relevant blocks of pain. On the *y* axis, the physical distance between 2 and 3 is the same as the physical distance between 6 and 7, just as it is on the *y* axis, which actually does use an interval scale (to represent time in minutes).

The problem, to be clear, is not that the difference between "2" and "3" is not equivalent to the difference between "6" and "7". It's that (a) we do not know what the magnitudes of such differences are, and so (b) we cannot know conclusions that depend upon such knowledge.

[17] In the graph above, the *x* axis represents time in intervals of minutes, while the *y* axis represents pain reports on the same scale as before. The Profiles for Patient A and B are intended to be the same as before. I have added a Profile for Patient C which is identical to the profile for Patient A except that: (a) Patient C reports pain levels of 1 for the first two minutes, when Patient A reported levels of 0; (b) Patient C reports a pain level of 5 for minute four, when Patient A reported a level of 7; (c) Patient C reports a pain level of 9 for minute seven, when A reported a level of 8; and (d) Patient C reports a pain level of 6 for minute eight, when A reported a level of 7.

[18] Kahneman, *Thinking, Fast and Slow*, p. 383

(Similar claims could be made in terms of justified beliefs about these matters.)[19]

To see why this matters, consider a fairly familiar example of a scale in which the difference between "2" and "3" is not equivalent to the difference between "6" and "7". Imagine that scientists gave you a single graph of earthquake activity in Regions A, B, and C. The *y* axis was intervals on the Richter Scale, and the *x* axis were calendar days. Imagine further that the numbers on the graphs were identical to the numbers on the graphs above. (Regions A, B and C are very earthquake prone, it turns out.) Now ask: Which region had worst earthquakes?

If we care about the total amount of energy released, the answer is actually C. This is because the Richter scale is logarithmic. In terms of the underlying feature of reality represented (energy released), the difference between 6 and 7 units on the Richter scale is far greater than the difference between 2 and 3; a million times greater, in fact. So the fact that Region C has the highest peak (namely, 9) should have much more influence on our evaluations, and on any subsequent decisions, than the fact that Region B suffered many more smaller quakes between 1 and 5. This is simply because you would need far more of those smaller quakes to make up the difference. This is easy to overlook. We're used to thinking that two 5s is greater than one 9. But on the Richter scale, one million 5s only equals the total amount of energy released by one 9. So when we aggregate to determine the total amount of energy released, or the average of any particular values on the scale, we would make significant mistakes if we assumed that we were dealing with a linear scale. We would put far too little emphasis on the peaks.

Now let's turn back to how we should evaluate the experiences of Patients A, B and C. We don't know whether the relevant scale is like the Celsius scale or like the Richter scale. And these obviously don't exhaust the possibilities. So we don't know how to aggregate the psychological data to compare the relevant total or average amounts of pain that the patients experienced.

From reading Kahneman's popular book *Thinking, Fast and Slow*, you would not think that there was any such methodological problem with aggregating the relevant psychological data. Here's how the Peak-end rule supposedly applies to the profiles of patients A and B: "The worst rating (8 on a 10-point scale) was the same for both patients, but the last rating before the end of the procedure was 7 for patient A and only 1 for patient B. The peak-end average was therefore 7.5 for patient A and only 4.5 for patient B."[20] Notice that C's peak-end average would be the exact same as A's on this view: the average of 9 and 6 is also 7.5. But it would be incorrect to treat the average of 9 and 6 units on the Richter scale as the same as the average of 8 and 7 units on the Richter scale. And if we do not know if this scale is like the Richter scale, then it is simply irresponsible to write as if we can aggregate the data so easily.

So far I have illustrated how we need to aggregate psychological data to support a particular

[19] In case this is unclear, the basic idea here is that in order to gain knowledge by inference, we must know the premises as well as we know the conclusion; if inferences to a conclusion (e.g., patient B suffered more) implicitly rely on the unknown premise that the scale has linear intervals, then we do not gain knowledge of the conclusion.

[20] Kahneman, *Thinking, Fast and Slow*, p. 380

prudential evaluation—the claim that Patient B's experience is worse than Patient C's—but have no responsible way of doing so. It should be obvious that this point applies to any other patients' pain profiles that do not have the same peak pain intensity. But those are precisely the profiles that are relevant to Kahneman's claims about what we should do and when we make decisions irrationally. In forming medical policies where "the objective is to reduce the amount of pain actually experienced", we simply do not know whether "conducting the procedure swiftly may be appropriate even if doing so increases the peak pain intensity"—because we do not know how much worse that peak pain intensity is unless we know the intervals on the scale. Likewise, we do not know whether one is *biased* when one puts a great deal of emphasis on the peak levels of pleasure or pain, because we do not know whether a short period of intense but tolerable suffering is actually far worse than a longer period of moderate pain. *Mutatis mutandis* for pleasure and other components of wellbeing.[21]

III—Is the Problem Soluble?

If the point above is right, we have a methodological problem in practical ethics. Conclusions about what's good for us, what we should do, and what decisions are irrational are defended on the basis of psychological data coupled with a fallacious inference about the relevant scales. My main target has been a psychologist, but Kahneman's views have been very influential in philosophy. And many philosophers have just accepted that the data he presents does in fact support the conclusions he has reached. For instance, here's Valerie Tiberius:

> It turns out that in assessing past painful experiences, for example, 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. To counteract the distorting effects of memory, some social scientists favor a type of measurement known as Ecological Momentary Assessment (EMA) to get at people's actual experiences, as they happen.[22]

Note that Tiberius accepts that putting more weight on "the worst part" is a "distorting effect".

And of course, Kahneman is just one example of a psychologist whose measurements of putative components of wellbeing have been used to support conclusions in practical ethics. The methodological problem described above generalizes to any such conclusions that rely on aggregations of psychological data in the absence of knowledge of the intervals of the scale.

If we have a methodological problem, is there a viable solution? I think not. I will address one potential solution from a psychologist here, and another potential solution to the related problem about hypothetical wellbeing measurements later. The reasons they fail are similar.

[21] In case this is not clear, Kahneman's claims about The Peak Rule are in fact defended on the basis of such simple aggregations of psychological data: one method of "confirming judgmental biases" is to "compare judgments to true value", which here requires "comparisons of subjective happiness to independent assessments of objective happiness" ("Objective Happiness", p. 22). See p. 19 for references to studies using this method that purport to show that the peak level of pain or pleasure has a distorting influence on our evaluations and decisions.

[22] Valerie Tiberius, 'Well-Being: Psychological Research for Philosophers', *Philosophical Compass* 1(5) 2006, p. 498 (emphasis added). Few philosophical discussions of Kahneman's work mention the intervals of the scale.

Addressing this potential solution will also help counteract a problematic implication via omission. The above might have been read to suggest that Kahneman himself succumbed to the psychometrician's fallacy. That would be misleading. Kahneman has explicitly claimed that he does not need to make assumptions about magnitudes of the intervals on the scale. Here are all of the "stringent assumptions" that Kahneman recognizes that he requires for his method of using patients' ratings to measure pain (or more generally, "instant utility"):

> these ratings must contain all of the relevant information required for its temporal integration to be a plausible measure of the total utility of an extended period. It is also assumed that the scale has a stable and distinctive zero point ("neither good nor bad", "neither approach nor avoid"), and that the measurement of positive and negative deviations from zero is ordinal. The subjects' ratings correctly order experiences in terms of Good or Bad, but 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*.[23]

He goes on to defend these stringent assumptions. I won't contest anything other than his discussion of the intervals on the scales, because this discussion is quite strange. Kahneman is not assuming that we have an interval scale; he recognizes that the intervals may be arbitrary. But he proposes that "a consistent rescaling is possible, yielding a ratio scale for instant utility that is calibrated by its relation to duration".[24] The key idea here is temporal integration.

> The core idea that leads to temporal integration is straightforward, and the colonoscopy example can be used to illustrate it. Obviously there are two ways of making a painful medical procedure worse: by increasing the level of pain, or by making the procedure last longer. Thus, an equivalence can be established between changes of pain intensity and of duration. Furthermore, because duration is measured on a ratio scale in physical units, it is possible in principle to rescale pain intensity in terms of duration.[25]

This is possible in principle because we can "rescale the reports made by the subject" by asking her to make trade-offs between higher levels of pain for shorter periods and lower levels of pain for higher periods. If "one minute of pain at level 7 is as bad as two minutes of pain at level 6", this implies that "the original reports of pain should be rescaled, assigning level 7 a value that is twice as high as the value assigned to level 6".[26] Doing so for every level of the scale would give us knowledge of the intervals on the scale. Is the methodological problem solved?

Not at all. Showing that the problem is soluble falls shy of providing the solution. Kahneman recognizes this: "The formal analysis describes a theoretical possibility, not a practical procedure", and since no such procedure has been performed, "the rule of temporal integration may not apply to the original profiles" of Patients A and B. "It applies," he says, "only after a rescaling that incorporates a judgment about the equivalence of intensity and

[23] Kahneman, 'Objective Happiness', p. 5, emphasis mine.
[24] Kahneman, 'Objective Happiness', p. 6.
[25] Kahneman, 'Objective Happiness', p. 5.
[26] Kahneman, 'Objective Happiness', p. 6.

duration".[27] What's strange about this discussion is that Kahneman has repeatedly relied on the original profiles to support prudential evaluations, deontic claims, and claims about biases, even though he acknowledges that this is unwarranted until those profiles are rescaled. And they have never been rescaled. So at best, temporal integration offers an urgent research project: rescale subjects' profiles and see if his conclusions stack up. Temporal integration should not, without engaging in that project, make us confident in such conclusions. And yet Kahneman still treats the conclusions reached through aggregating *instant utilities* as "objective" and "true" measures of total utility, without having rescaled the profiles.[28]

Notably, many other studies using Kahneman's own methodology simply aggregate the relevant psychological data without even mentioning why this might be problematic. For instance, conclusions about a subject's evaluations of extended experiences have been reached on the basis of "moment-by-moment account[s] of how happy he or she is feeling", which is used to reach a "measure of the total happiness experienced" that is simply "*the sum* of the happiness of all the different moments".[29] "Rescaling" the profiles was not discussed at all.

Some psychologists do recognize that rescaling is required, but do not recognize how this undermines the conclusions that they draw from the data. For instance, Barbara Fredrickson writes that "empirical support for the peak-and-end rule is robust", and shows that our decisions are plagued by "biases and mistakes": "A variety of studies have demonstrated that these mistakes may cause people to … choose more pain rather than less". Yet she also recognizes that *for policy-makers* to "aggregate those good-bad ratings they must convert them into a ratio scale that calibrates affect intensity relative to duration."[30] If *they* need to do this, it is baffling why psychologists apparently do not. Why doesn't she need to rescale subjects' ratings in order to know whether subjects were actually choosing "more pain rather than less"?

A more important issue, however, is that temporal integration does not even show that it is possible in principle to rescale reports of instant utility on an interval scale. The reason why comes from Kahneman's own work on "duration neglect." The variable of duration "had little or no effect on retrospective global evaluations" in all of the studies Kahneman mentions".[31] Kahneman claims that "people apparently construct and evaluate a representative moment and use the evaluation of this moment as a proxy for the evaluation of the entire episode. Duration is effectively deleted from this representation".[32] Now note that Kahneman's method of temporal integration does not rely on using duration (measured on a ratio scale) itself to calibrate reported instant utility. It relies on using *remembered duration* (which is, he says, "effectively deleted" from our construction of the relevant "representative moments") to calibrate reports of instant utility. The latter does not offer us a route to a ratio scale.

[27] Kahneman, 'Objective Happiness', p. 6.
[28] Kahneman, 'Objective Happiness', p. 4.
[29] Simon Kemp et a, 'A test of the peak–end rule with extended autobiographical events', *Memory and Cognition*, 26(1) 2008, p. 132 (emphasis mine); the study uses this method to measure experiences on holidays.
[30] Barbara Fredrickson, 'Extracting meaning from past affective experiences', pp. 585, 589, 598.
[31] Kahneman, 'Objective Happiness', p. 19.
[32] Kahneman, 'Objective Happiness', p. 20.

We don't need to know that Kahneman is right about duration neglect for this to be a problem. (Though I think he is right.) We just need to know that it is a live possibility. The claim I am making here is not that temporal integration would lead to *false* judgments about the relevant scales. It is that temporal integration could so easily lead to false judgments about the relevant scales that it cannot provide the knowledge of the intervals of the scales that we require.

In fact, even if Kahneman were wrong about duration neglect, this rescaling procedure still would not succeed even in principle. To see why, we must recognize that this rescaling procedure elicits judgments about a finite series of tuples: <pain level, duration>. From these judgments, Kahneman is making an inductive leap to a ratio scale that predicts the ordering of an *infinite* series of tuples, and guarantees that this ordering has certain formal features; for instance, it would have to be transitive.[33] It is hard to see what justifies this inductive leap. I find it extremely doubtful that people's actual preferences over a sufficiently large finite series of judgments about *x* level of pain for *y* duration would yield a transitive ordering.[34] One minute of pain at level 10 might be worse than two minutes of pain at level 9, while two minutes at level 9 would be worse than four minutes at level 8, and so on and so forth until we reach eight and a half hours of pain at level one; but *that* might be worse than one minute of pain at level 10. This problem might not emerge if we stick to a small set of pair-wise judgments. But unless we have some prior justification for starting with that particular set this is simply *ad hoc.* And any ratio scale that emerges from this rescaling procedure will be just as arbitrary as its starting points.

To be clear, this problem is fully general. It applies to many other conclusions reached about issues in practical ethics that implicitly depend on premises about the scale, which we do not know. But not all conclusions about issues in practical ethics that appeal to psychological data have this feature. It is possible to support philosophically interesting and surprising claims about what's good for us, what we ought to do, and what decisions are irrational, on the basis of psychological data that does not need to be aggregated into a total or an average.

Kahneman's work on the End Rule is a case in point. You can demonstrate the End Rule, and provide good evidence that it is a bias, without knowing whether patients' pain reports are on an interval scale. Recall the original graphs of Patient A and Patient B. Whatever pain Patient A experiences, Patient B experiences too; but Patient B also experiences additional periods of pain. Dominance reasoning suggests that Patient B's experience is worse than A's. This is even more obvious in Kahneman's earlier "cold pressor" experiments, where patients endured a painful stimulus (having a hand submerged for 60 seconds in cold water), and some patients endured that same stimulus immediately followed by an additional albeit less painful stimulus (having a hand submerged for 30 seconds in less cold water). Patients rated the second condition as better than the first even though it shares all the bad features of the first, plus an

[33] I am very grateful to Ben Jantzen for helpful discussions of this point. Any mistakes here are mine, however.

[34] For relevant discussion, see Larry S. Temkin, *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*, (Oxford University Press, 2012), especially the one-person spectrum case on pp. 134-139.

additional period of pain.[35] On its face, this is an irrational violation of a dominance principle.[36] And *prima facie*, it suggests that patients irrationally neglect the duration of pain episodes.

One implication of my discussion so far is that if practical ethicists are to appeal to psychological data about putative components of wellbeing, they should seek out experiments that resemble the cold pressor study: studies where the aggregation of data is not required. Such experiments will be hard to come by. Subjects like Patient C are difficult to avoid in large-scale surveys or experiments that occur outside of carefully controlled laboratory conditions. And dominance reasoning will not apply as soon as we introduce subjects like Patient C. So this suggests that the methodological problem in practical ethics is not inevitable for philosophers who wish to appeal to psychological data. But if it cannot be solved, it is difficult to avoid.

IV—Actual and Hypothetical Aggregations

So far I have argued that we have a methodological problem with appeals to actual data about wellbeing to support conclusions in practical ethics. Now I want to argue that we have a similar methodological problem in the purely theoretical domain of normative ethics when philosophers rely on hypothetical wellbeing measurements to defend normative theories.

It will be helpful to make two preliminary points about thought experiments in general before turning to thought experiments in normative ethics that involve hypothetical wellbeing measurements. First, thought experiments work by eliciting intuitions, and the object of those intuitions is some proposition about the scenario we imagine.[37] It is important to recognize this because it is possible that the scenario imagined fails to comport with stipulated details of the thought experiment. For instance, consider the following hypothetical scenario:

> *David has never handled a gun before. Because his friends dare him to do so, David blindfolds himself and fires a machine gun at a crowd of people. David does not hit anyone. Let's stipulate that everyone knew that David would not hit anyone. Intuitively, what David does is wrong, so actions can be wrong when we know no one will be harmed.*

I take it that this thought experiment obviously fails to support the conclusion in the final sentence. And I take it that the reason why it fails to support this conclusion is that (a) the final sentence elicits an intuition about the case imagined ("what David does"), and (b) it is obvious that the case imagined is unlikely to comport with the stipulation in the penultimate sentence. At least in the absence of further description, it is obvious introspectively that the scenario we

[35] Daniel Kahneman et al, 'When More Pain Is Preferred to Less: Adding a Better End', *Psychological Science* 4(6), (1993), pp. 401-405.

[36] Why only "on its face"? Because as Stephanie Beardman has argued ('The choice between current and retrospective evaluations of pain', *Philosophical Psychology*, 13(1) 2000), there is "an alternative interpretation of the data": "why can't it be that subjects remember the relative duration of the pain, and don't care about it? What evidence is there to say that caring more about the *shape* of an experience than its duration represents a cognitive error?" (pp. 97, 105). A similar point is made by Kelman in 'Hedonic Psychology and the Ambiguities of "Welfare"'.

[37] For discussion of the relationship between thought experiments and imagination, see James Robert Brown and Yiftach Fehige's 'Thought Experiments', *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), and Tamar Szabó Gendler, *Intuition, Imagination, and Philosophical Methodology* (Oxford University Press, 2010).

imagine is one in which it is not true that everyone knows that no one will be harmed by David.

Second, we can have mistaken second-order beliefs about the contents of our intuitions. Opacity of mind is at least possible, if not quite common: there are plenty of plausible philosophical views on which we can have mistaken beliefs about our own mental states,[38] and plenty of plausible cases in which we do not know crucial details of what we imagine.[39] This matters because when we respond to thought experiments, it is possible that we falsely believe that the scenario imagined does comport with stipulated details of the thought experiment. Saul Kripke, for instance, argued quite persuasively that such mistakes were widespread in relation to thought experiments about natural kinds: philosophers mistakenly believed that they had imagined that water was not H_2O, when in fact they had imagined that the local watery stuff—the stuff with the same surface properties as water—was not H_2O.[40]

These two points illustrate a common concern about thought experiments in philosophy: inferences drawn from responses to thought experimenting are potentially misleading if the hypothetical scenario "is inadequately described."[41] This is a general methodological problem. If you like, you can consider the following to be specific instance of this general problem—I have no strong views about how we individuate problems. My contention is that there is a methodological problem with uses of hypothetical wellbeing measurements in normative ethics, and that it is hard to recognize because of the representational fallacy. Such uses of hypothetical wellbeing measurements are at worst misleading, and at best redundant.

To focus matters, let's stick to the use of hypothetical wellbeing measurements to support conclusions about a particular theory in normative ethics, prioritarianism, which is the view that improvements in wellbeing of worse off individuals matter more, morally speaking. An illustrative example here is the following thought experiment from Roger Crisp.

> Consider the following pair of distributions, called *Equality* and *Inequality*:
>
> | | *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*

[38] Peter Carruthers, *The Opacity of Mind: An Integrative Theory of Self-Knowledge* (Oxford University Press, 2011). See also Timothy Williamson on anti-luminosity in *Knowledge and its Limits* (Oxford University Press, 2000).

[39] See, for instance, Peter Mendelsund's *What We See When We Read* (Vintage, 2014) for a series of examples of literary fictions where most readers are surprised to learn that they did not imagine certain details.

[40] Saul Kripke, *Naming and Necessity* (Reidel, 1972).

[41] Kathleen Wilkes, *Real People: Personal Identity without Thought Experiments* (Oxford University Press, 1988), p. 8). For our purposes, to be adequate a description must at least make it very plausible that the imagined scenario comports with the author's stipulations about that scenario—e.g., hypothetical wellbeing measurements.

have lives that are either much better or much worse than the lives of those in *Equality*.[42] Attached to that final sentence is the following footnote: "No commitment to precise measures or to any particular view of welfare itself is intended by the use of numbers; throughout the article, numbers may always be understood in terms similar to those used in the sentence to which this note is attached in the text." Crisp is, in effect, stipulating that the *numerical units on this scale* should be understood in the same way we treat *numbers generally*: in other words, he stipulates that we have a ratio scale for wellbeing.

You might think that we simply form probative intuitions about these stipulated distributions of wellbeing, and can appeal to them directly. I think that is a mistake. Thought experiments are devices of the imagination, not stipulation. It is at least possible that whatever scenario we imagine about this thinly described case does not comport with the stipulated wellbeing measurements on a ratio scale. (That's the first general point from before.) Indeed, it is possible that the scenario we imagine does not comport with such stipulations, without our knowing that this is so. (That's the second general point from above.) And in fact, given our earlier discussion of the representational fallacy, such mistakes would be far too easy to make. After all, the hypothetical wellbeing measurements are an additional layer of representation, and it is easy to assume that a salient feature of the representational device (i.e., that the magnitude of the differences between 10 and 50 and between 50 and 90 are the same) is shared by the attribute that they are used to represent (i.e., that the magnitude differences between Equality and Inequality for Group 1 is the same as differences in these scenarios for Group 2).

My contention is that the above shows that Crisp's thought experiments are potentially misleading, for a reason that is rarely recognized. We are likely to have the second-order belief that the scenario we imagined comports with the stipulated wellbeing measurements on a ratio scale. My contention, to be clear, is that not that this belief is false. It is that *we do not know* whether it is true. We could easily be in the same position as those who mistook intuitions about watery stuff for intuitions about water. And if we were in that position, we would have the same second-order belief, because it is far too easy to infer, fallaciously, that the scenario represented shares a salient feature of the numbers that have been used to represent it.

To be clear, there are plenty of cases in which we might mistakenly believe that the scenario we imagine comports with how the thought experiment was described. Philosophers routinely idealize in various ways to simplify things and focus on one particular variable. But such mistakes should not concern us unless they may undermine the probative force of the intuitions in question. When the question is whether actions can be wrong when we know that no one will be harmed, it matters if the action we actually imagined was reckless. When the question is whether water is H_2O, it matters whether our intuitions concern the possible molecular composition of *water* or *watery stuff*. And when the question is whether improvements in the wellbeing of worse off individuals matter more, it matters whether in the imagined scenarios the differences in wellbeing are of lesser, equal, or greater magnitudes.

[42] Roger Crisp, 'Equality, Priority and Compassion', *Ethics*, 113(4), (2003), pp. 745-746.

That's the main reason why I think uses of hypothetical wellbeing measurements in normative ethics like Crisp's are potentially misleading. And to be clear, this is an illustrative instance of a fairly widespread practice, at least in the discussion of theories like prioritarianism. As Michael Otsuka recently noted in a related context, "[i]n discussions of prioritarianism, it is often left unspecified what constitutes a greater, lesser, or equal improvement in a person's utility": it is simply "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."[43] If the main point I made above is right, these uses of hypothetical wellbeing measurements in normative ethics are methodologically problematic in a similar respect to the problem with uses of actual wellbeing measurements in practical ethics.

In fact, I think there is one respect in which hypothetical wellbeing measurements in normative ethics are worse. The scale of actual wellbeing measurements is bounded on both ends (e.g., from "0" for "no pain at all" to "10" for "intolerable pain"). Hypothetical wellbeing measurements typically include no upper bound (it is a scale *from* "0", but not *to* anything), and no external reference point, and these features are widely recognized to make it harder for individuals to determine what a unit on the scale is meant to represent.[44] This increases the likelihood that such uses of hypothetical wellbeing measurements will be misleading.

That said, I do not want to contend that *all* uses of hypothetical wellbeing measurements are misleading. Consider, for instance, a central example in this literature, which originally comes from Thomas Nagel. He imagines a parent who has two children, one of whom is happy and healthy, while the other suffers from a painful handicap. The parent could either move to a city where the second child could receive special treatment, or move to the suburb where the first child will flourish. Crucially, moving to the suburb is meant to give the healthy child a greater benefit, but the unhealthy child is meant to be worse off.[45] When Derek Parfit discusses this example, he "use[s] figures" to make this vivid by adding these hypothetical measurements:

	move to the city	*move to the suburb*
healthy child	20	25
unhealthy child	10	9

Interestingly, Parfit noted right upfront that "such figures misleadingly suggest precision." In particular, he recognized that for the figures to be probative, we must assume the following:

> Each 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

[43] Otsuka, 'Prioritarianism and the Measure of Utility', *The Journal of Political Philosophy*, 23(1) 2005, pp. 1-2. Otsuka's challenge to prioritarianism in that paper is more complex than, and orthogonal to, the challenge here.

[44] For discussion of why these features of scales are problematic (using examples from psychology and economics), see Daniel Kahneman and Roger Sudgen, 'Experienced Utility as a Standard of Policy Evaluation', *Environmental & Resource Economics* (2005) 32: 161–181.

[45] See Nagel, *Equality and Partiality* (Oxford University Press, 1991).

> who was worse off.[46]

He's right. Without the assumption that we have ("at least roughly") an interval scale, we cannot know whether the change from 20 to 25 represents a greater benefit than the change from 9 to 10. If in the scenario we imagine the improvement in wellbeing for the less well-off child were greater, the intuition that it matters more to move to the city would be irrelevant.

Let's assume that this is not a misleading thought experiment. That is, let's assume that we know that in the scenario we imagine the improvement in wellbeing for the less well-off child is smaller. Why would we be justified in believing that this is the case? I want to contend that it would not be because of the stipulated wellbeing measurements on an interval scale. These represent the imagined scenario, and make it easy to assume that the imagined scenario comports with the relevant stipulation; but there's no reason to think that the stipulation itself helps us actually imagine a scenario in which there is a lesser benefit to the less well-off child. What would make it plausible that in the scenario we imagined there was a lesser benefit to the less-well of child would be the description of the case from Nagel, not Parfit's use of figures. If this is right, then the addition of such figures adds nothing; if the description is adequate to make the use of figures *not* misleading, then those figures are just going to be redundant.

So far I have outlined a methodological problem with the use of hypothetical wellbeing measurements in normative ethics. This practice is potentially misleading (because it is easy to fallaciously infer that the magnitude of differences between numbers maps on to the magnitude of the differences between the levels of wellbeing of the imagined people). And where such measurements are not misleading, they are redundant; the hypothetical scenario must already be adequately described to ensure that the measurements are not misleading, in which case it is hard to see how the measurements add anything to thought experiment.

If this is a methodological problem, is it soluble? Can we ensure that such measurements are not misleading without making them redundant? Interestingly, the best suggestion for a solution that I have seen came in the discussion of a related objection to prioritarianism that was made in one of the earliest discussions of the view. In *Weighing Goods* (1991), John Broome argued that prioritarians need a measure of a person's wellbeing that is distinct from the value of her wellbeing. Here's Broome's more recent summary of his objection:

> Compare these two distributions of wellbeing:
>
> *I* = (3, 3)
> *J* = (2, 4)
>
> A prioritarian will think *I* better than *J*. Imagine a change from *J* to *I*. In this change, the first person gains one unit of wellbeing and the second person loses one. But the first person has priority, because she is worse off than the second. A unit change in her wellbeing is more valuable than a unit change in the second person's wellbeing. So in a change from *J* to *I*, the two people's wellbeings change by the same quantity – positive for one and negative for the other – but the changes differ in value.

[46] Derek Parfit, 'Equality or Priority', *The Lindley Lecture* (Lawrence, Kansas: University of Kansas, 1991), reprinted in *The Ideal of Equality*, ed. Matthew Clayton and Andrew Williams (Basingstoke: Palgrave, 2002), p. 83.

> Prioritarianism presupposes a quantitative scale for the quantity of wellbeing: a *cardinal* scale, to be exact. Philosophers often take it for granted that we have a rough quantitative notion of wellbeing. No doubt that is so, but we need to ask what its source is. To have a cardinal scale, we have to be able to make sense of statements that compare the quantity of changes in people's wellbeing. The example contains one change from 2 to 3 and another from 4 to 3, and these changes are supposed to be the same in quantity: one unit. What exactly does it mean to say they are the same in quantity?[47]

Broome then considers various candidate answers on behalf of the prioritarian—such as that they are the same in quantity because they have the same moral value—and rejects them. Most interesting, for our purposes, is his discussion of "the valuation of uncertain prospects". We could consider the perspective of a single agent valuing a choice between *J* (where an outcome of 3 is certain) and *I* (where the outcomes 2 and 4 are equally probable). If the agent is indifferent between those options, that could support the contention that the difference between 2 and 3 is the same as the difference between 3 and 4. And we can then "assign meaning to quantities of wellbeing by generalizing this idea."[48]

Broome responds that "the prioritarian cannot get her quantities of wellbeing this way" due to a complicated theorem from Harsanyi. For our purposes, two simpler points can be made, each resembling the two responses to Kahneman's appeal to temporal integration from before.

The first is that if we need to appeal to valuations of uncertain prospects to "assign meaning to quantities of wellbeing", then we need to actually do so *in each thought experiment*. It is one thing to say that in principle, we can support the claim that a change from 2 to 3 and a change from 4 to 3 units of the relevant scale are in fact changes of the same in quantity of wellbeing. It is another to say that when we respond to the thought experiment, the scenario we imagine is one in which such changes in units represent changes of the same quantity of wellbeing.

The second is more important. Even in principle, I doubt that we can appeal to valuations of uncertain prospects. This is largely due to the work of Lara Buchak, who argues that actual agents do, and rational agents can, pay special attention to worst- and best-case scenarios.[49] A simple way to put this point is as follows. Two agents for whom the difference between 2 and 3 is the same as the difference between 3 and 4 might (rationally) be more aversive to *risk itself*, and so differ in the gambles they would be willing to make between *J* and *I*. Likewise, two agents for whom 2 and 3 is *not* the same as the difference between 3 and 4 might (rationally) be more aversive to *risk itself*, and so be willing to make the same gambles between *J* and *I*.

So, just as I doubt that valuations of different levels of wellbeing for different durations allows us to calibrate scales of wellbeing in actual subjects, I doubt that valuations of different probabilities of different levels of wellbeing allows us to calibrate scales of wellbeing in

[47] John Broome, 'Equality versus Priority: A useful distinction', *Economics and Philosophy*, 31(2) (2015), p. 226.
[48] Broome, 'Equality versus Priority', pp. 226-227.
[49] Lara Buchak, *Risk and Rationality,* (Oxford University Press, 2013)

hypothetical subjects. In both cases, we cannot appropriately isolate the variable of wellbeing from the relevant confounding feature that is meant to introduce precision: duration or risk.

VI—Conclusion

I have sought to identify two methodological problems in ethics. The first concerns inferences to conclusions about practical ethics from premises about actual wellbeing measurements, where an implicit premise of the argument is that the measurements are on an interval scale. I have offered an explanation for why that implicit premise is easily assumed to be true—because the scale is represented with devices that have known intervals—even when we have no actual or possible evidence that would show that it is true. Since this implicit premise is not known, and we are not justified in believing it, the same holds for any conclusions it supports.

The second problem concerns inferences to conclusions about normative ethics from premises about hypothetical wellbeing measurements, where a stipulated feature of the measurements is that they are on an interval scale. I have offered an explanation for why it is easy to assume that our intuitions about the relevant scenarios comport with that stipulation—because the scale is represented with devices that have known intervals—even when that assumption is not supported by adequate evidence. Further, I have argued that the evidence that would make that assumption adequate would also make hypothetical measurements, as an additional representational device, redundant. So such uses of hypothetical wellbeing measurements in normative ethics are at worst potentially misleading, and at best redundant.

If I am right that these are methodological problems, what should we do about them? I will end by suggesting that we should use a helpful heuristic, "The Letters for Numerals Heuristic": when we only have enough information to use an ordinal scale, but that scale is represented with numerals, simply replace the numerals with letters. A salient feature of letters is that they provide a rank ordering from lower to higher (a, b, c) without intervals or ratios. That is, they encode precisely the amount of information that an ordinal scale is meant to encode.

This heuristic can be used for preventative and diagnostic purposes. If we replace the numbers in wellbeing measurements with letters, we can quickly tell whether inferences drawn from those measurements are unwarranted. Recall our original example of an ordinal scale. We knew that Marquis' happiness > Michelle's happiness > Melania's happiness, and represented their levels of happiness with numerals ("3", "2", and "1"). Replace these with letters ("C", "B", and "A") and it is transparent that inferences that involve ratios (e.g., Michelle is twice as happy as Melania), intervals (e.g., the differences between their levels of happiness are equal), and addition (Michelle's happiness + Melania's happiness = Marquis' happiness) are unwarranted. By contrast, when we consider Kahneman's cold pressor experiment, the inferences drawn do not seem problematic when we replace numbers with letters: to judge that A + B is worse than A while knowing that A and B both represent episodes of pain is, at least *prima facie*, irrational. So if we wish to avoid common practices that put scales over our eyes, we need not eschew scales entirely; we simply need to avoid representing scales in an irresponsible manner.