European Journal for Philosophy of Science (2022) 12: 7
https://doi.org/10.1007/s13194-021-00437-7

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

Democratising Measurement: or Why Thick Concepts Call for Coproduction

Anna Alexandrova[1] · Mark Fabian[1]

Received: 20 July 2021 / Accepted: 8 December 2021 / Published online: 20 January 2022

Abstract

Thick concepts, namely those concepts that describe and evaluate simultaneously, present a challenge to science. Since science does not have a monopoly on value judgments, what is responsible research involving such concepts? Using measurement of wellbeing as an example, we first present the options open to researchers wishing to study phenomena denoted by such concepts. We argue that while it is possible to treat these concepts as technical terms, or to make the relevant value judgment in-house, the responsible thing to do, especially in the context of public policy, is to make this value judgment through a legitimate political process that includes all the stakeholders of this research. We then develop a participatory model of measurement based on the ideal of co-production. To show that this model is feasible and realistic, we illustrate it with a case study of co-production of a concept of thriving conducted by the authors in collaboration with a UK anti-poverty charity Turn2us.

Keywords Thick concepts · Wellbeing · Measurement · Coproduction · Values in science

1 Introduction

Value-laden phenomena – wellbeing, resilience, biodiversity, sustainability, vulnerability, quality of care, and so on – are ubiquitous in contemporary social and life sciences. In these cases, the very definition of a scientific term requires an evaluative standard, often a controversial one. Judgments about moral, political or aesthetic value thus enter into the most technical aspects of research, namely measurement. The ongoing efforts to develop an evidence base around, for instance, well-being,

✉ Anna Alexandrova
aa686@cam.ac.uk

Mark Fabian
mf723@cam.ac.uk

[1] University of Cambridge, Cambridge, UK

requires defining it as either a subjective judgment measured by self-reports, or an objective state measured by behavioural or social indicators, or some combination of the two. Which self-reports and which indicators to select is a matter of controversy, often pitting against each other competing visions of the good life.

Philosophers engage with these issues through the notion of ‘thick concepts’, those that describe and evaluate simultaneously. Their presence in science is widely recognised and well documented.[1] Commentators typically use thick concepts to challenge traditional views of the objectivity and value-freedom of science. But beyond this it is less clear what practical recommendations to follow in the presence of thick concepts. What is responsible practice when it comes to the measurement of phenomena they pick out? Should these concepts be eliminated, subjected to special methods, celebrated and multiplied? In this paper we lay out the options and make concrete the idea that measurement of thick concepts should be democratised.

In Part I we articulate three strategies open to researchers working with thick concepts. They are to redefine thick concepts as technical terms that lose their evaluative content, to assume full responsibility for making the relevant value judgment, and finally to make the value judgment through a legitimate political process. We submit that these options, once articulated with care, are likely exhaustive, with all viable approaches falling in one or the other of the three. They also helpfully systematise various proposals other commentators articulated less explicitly.

In Part II we argue that the third strategy is the responsible choice, other things being equal. Our proposal takes its cue from the growing theory and practice of participatory science, but focuses specifically on measurement. Our core claim is that measures of variables picked out by thick concepts can and should be co-produced in collaboration with stakeholders who bring distinctive types of expertise, each relevant to measurement. The resulting instruments should blend values of all stakeholders with technical and practical constraints on the instruments themselves.

By way of a proof of concept we end with a case study. We report on a process of co-producing a conception of ’thriving’ that the authors implemented with Turn2us, a national anti-poverty charity in the UK. While the theoretical argument in Parts I and II shows that the co-production of measurement scales is desirable, the case study shows that there is nothing inherently impossible in our proposal to democratise measurement.

2 Part I: Three available strategies

Consider the definition of thick concepts given by Elizabeth Anderson:

> A concept is thickly evaluative if (a) its application is guided by empirical facts; (b) it licenses normative inferences; and (c) interests and values guide

[1] See Kirchin (2013) and Tiberius (2013) on thick concept in ethics; Reiss (2010), Root (2007), Abend (2019) on social sciences; Dupré (2007), Hawthorne (2013), Kingma (2014), Stegenga (2015a), Djordjevic and Herfeld (2021) on life and medical sciences. See Väyrynen (2013) for a dissenting view that thick concepts are not inherently but only pragmatically evaluative.

> the extension of the concept (that is, what unifies items falling under the concept is the relation they bear to some common or analogous interests or values) (2002, 504-505).

This is a general definition which Anderson goes on to apply to the concept of ’intelligence’, but it fits ’well-being’, our example throughout this paper. Users of this concept certainly aim to ascribe it on empirical grounds, as Anderson requires in condition (a). Informally, parents eyeball their child’s behaviour to check if they are well, whereas more formal indices that have proliferated in the recent decades guide the application of this concept in contexts of development, healthcare, management, and policy making. Once these assessments are made, they certainly feed into practical decisions about what needs to be done to improve wellbeing. That is condition (b) and the point behind any evidence-based endeavour whether in parenting, self-help, or wellbeing public policy.[2] Finally, as condition (c) stipulates, definitions of wellbeing require judgments about what is good for the people whose wellbeing is in question. This is why there are deep and longstanding disagreements among researchers of wellbeing about, among other issues, whether wellbeing is a mental state and if so which one. Value commitments such as hedonism, utilitarianism, liberalism, and eudaimonism are regularly invoked as inspirations for adopting one or another approach to wellbeing.[3] So wellbeing is certainly a thick concept in Anderson’s sense.

Much of the discussion of thick concepts in philosophy of science, including Anderson’s own writings, has been dedicated either to showing their presence or to arguing for their ineliminability and legitimacy in the face of traditional demands on science.[4] We find it helpful to systematise these discussions into three strategies.

2.1 Strategy One: Turn thick concepts into technical terms.

Often the most natural way for scientists to proceed is to get rid of the evaluative element of thick concepts, thereby turning them into technical terms. Examples from economics include the “discount rate” and “cost of living” (Stapleford, 2009; Deringer, 2018). Essentially this amounts to denying that thick concepts exist as such, since, if they do, such a separation is not supposed to be possible (Putnam, 2004). This approach comes naturally because high profile success of scientific theories often consists in postulating new concepts and showing their fruitfulness through application.[5] If so, it makes no sense to demand that a concept properly captures some pre-theoretical notion because conceptual change is the whole point.

[2] See Dolan and Peasgood (2008), Dolan and White (2007), Clark et al. (2018) and Frijters et al. (2020) among many calls for evidence-based wellbeing public policy.

[3] See Haybron and Tiberius (2015) and Adler and Fleurbaey (2016) on the political theory behind wellbeing policy.

[4] Anderson (2004), Douglas (2011), Brown (2020) and Alexandrova (2017) (chapter 4) are some examples.

[5] The idea that the main scientific achievement is in conceptualising nature in fruitful new ways is a staple in history and philosophy of science, especially the Kantian strands. It is central to Carnap (1950), a more recent restatement is Friedman (2001).

Is Newtonian mass really what people mean by ‘mass’? Maybe not, but that does not matter if Newtonian mass enables as many epistemic achievements as it does.

In the sciences with thick evaluative terms, we rarely see such reasoning explicitly. No wellbeing researcher literally says: "It does not matter if life satisfaction is really wellbeing. We use ‘wellbeing’ to mean ‘life satisfaction’ because life satisfaction is a more fruitful concept". Nevertheless it is possible to pick up traces of such reasoning in the way that scientists justify their operationalisations in the methodology sections of research articles (Cohen Kaminitz, 2018). It is common to encounter researchers adopting a particular definition of wellbeing and justifying it because it fits best their measurement tools, or enables the use of new dataset, or because it is theoretically interesting, or because it fits previous definitions, or models.[6] Absent in such reasoning is any explicit recognition of the evaluative element in the meaning of the concept and absent is an attempt to supply an argument that justifies this element in a way that evaluative concepts should be justified.

While it is rare to encounter this strategy in its pure form in published wellbeing research, there exists an attempt to defend such a stance explicitly. Ernst Nagel did so with his distinction between *appraising* and *characterising* value judgments (Nagel, 1961). Scientists appraise when they approve or disapprove of something on the basis of a commitment to an ideal – for example, when they use the thick term ‘anemic’ to highlight how poorly an organism is faring. In contrast, scientists characterise when, to use Nagel’s own words, their “value judgement expresses an estimate of the degree to which some commonly recognised … type of action object or institution is embodied in a given instance”(Nagel, 1961, 492). In the first case the scientist endorses the value, while in the second they merely report that an animal is anemic according to an agreed definition. Nagel puts forward this distinction to vindicate the possibility of value-freedom of science even when its central concepts are thick. He says that even if the two kinds of value judgments will in practice bleed into each other, it is still logically possible to stick to characterising rather than appraising. Nagel also thinks this is desirable because scientific knowledge should be objective in the sense of being "value-free and unbiased" (ibid, 502).

Although he does not use our language, Nagel’s proposal is effectively to place appraising value judgments outside science and treat all thick terms in a manner that is agnostic about their evaluative element. These terms thus become technical in the sense that their everyday evaluative connotation is erased and they are judged only by the more familiar epistemic virtues of scope, simplicity, empirical adequacy, etc. They become scientific terms first and foremost.

2.2 Strategy Two: Keep the value judgment in-house

We said that the first strategy comes naturally to scientists with traditional views of science, but it is also common to encounter the second strategy. This is when researchers mount, first, an explicitly normative argument in favour of adopting one

[6] Arguably Daniel Kahneman’s (1999) defence of the concept of ‘objective happiness’ followed this strategy. Instead of making an ethical case that wellbeing is happiness, he positioned it as an interesting measurable quality and a theoretical contrast to remembered or expected utility.

or another operationalisation of a thick concept and, secondly, do so by appeal to their own personal normative intuitions or the consensus of their discipline. Both parts are important because, as we shall see shortly, it is possible to have the first without the second. As an example of this strategy, consider the following from Oishi et al. (2018, p. 164-165):

> What is a good society? From the perspective of the science of happiness, a good society is a society that makes its citizens happy. Various policy ideas can be evaluated in terms of happiness.

In the well-being space, Strategy Two leads traditional economists to intentionally adopt a preference-satisfaction account of welfare, psychologists to adopt mental-state accounts, and so on. Crucially, this adoption is not agnostic, as in the first strategy, but rather it is mindful and deliberate. It comes with an attempt to defend a given operationalisation by marshalling arguments about its ethical appropriateness. In the wellbeing sciences the second strategy has been prominent ever since the field matured in the 1990s. The proponents of life satisfaction often justify it by saying that it empowers respondents to decide what matters (Diener et al., 2009). For example, Clark et al. (2018, p. 4) give following three reasons for using life satisfaction metrics over measures of affect or meaning in life:

> First, it is comprehensive—it refers to the whole of a person's life these days. Second, it is clear to the reader—it involves no process of aggregation by researchers. Third, and most important, it is democratic—it allows individuals to assess their lives on the basis of whatever they consider important to themselves.

Finally there are also famous deployments of Aristotelian considerations when defending accounts of wellbeing in terms of character and virtues (Seligman, 2012) or in terms of capabilities (Alkire et al., 2015).

These attempts by social scientists to build a normative case for thick concepts do not always satisfy professional ethicists and there is thus a whole literature of philosophers challenging the justifications of measures of wellbeing given by scientists.[7] More recently, there have been calls for philosophers and psychologists to collaborate more closely in a process of conceptual engineering to develop an account of wellbeing that is descriptively, empirically, and normatively adequate for psychological science.[8] Our point is only to note that sometimes scientists do take it upon themselves to mount normative arguments based on their own visions of the good life and to the extent that they do, they see this strategy as open to them qua scientists.

[7] Haybron (2008), Feldman (2010), Nussbaum (2000) and Kristjánsson (2013) among many others.

[8] Tiberius and Hall (2010), Prinzing (2020) and Vessonen (2021b).

2.3 Strategy Three: Seek political legitimacy

Suppose you refuse to turn a thick term into a technical term, and you lack the assurance to make the value judgment yourself. What more can you do? The third strategy, as we see it, is to fill out the thick content by a legitimate political process. The motivation behind this option is simple: if the practice of science requires making value judgments about essential aspects of life such as wellbeing, and if this knowledge is sometimes close to power and therefore potentially coercion, then these judgments should be subject to a legitimacy requirement. In political theory, legitimacy is a property – whose nature is widely debated – that justifies the power of state or institutions over citizens (Peter, 2017). In our case, legitimacy would be a constraint on the epistemic process, that is a constraint on the way in which thick concepts are approached by scientists and researchers. The purpose of such a constraint is to give this knowledge an additional layer of security: to the familiar scientific process covered in textbooks on measurement – more on that in Section II.2 – Strategy Three adds a new political requirement.

Exactly what this requirement demands is a big question, which we begin to answer in Part II. For now a minimal definition is sufficient: Strategy Three requires that the process of specifying the content of a thick concept takes into account the relevant value judgments of those to whose lives stand to be affected by this research. This is the sense in which Strategy Three calls for democratisation. How exactly? Full electoral competition, representative parliaments, and other large scale democratic exercises are typically ill-suited to the meticulous and niche process of measurement. So what options are there?

Recent decades have seen a rise of public participation in science – a diverse movement that takes many forms from citizen science, to public consultations, to simple outreach.[9] We take cue from one strand of this movement, namely stakeholder engagement. Stakeholders are individuals and communities who are outside the scientific process but who have a genuine interest at stake in a given scientific or healthcare project.[10] In our case, the stakeholders have an interest in how social scientists operationalise thick concepts, because these thick concepts may be used to rearrange their lives through new policies and institutions. So Strategy Three invites scientists to share the power and the responsibility of this task with the full of range of potential users of these concepts and those who stand to lose or benefit from them.

In the case of wellbeing, this strategy calls for researchers to learn whether their preconceptions about wellbeing line up with the views held by the people whose wellbeing they are trying to measure and to study, namely the stakeholders. Crucially, the demand is not just to learn about the wellbeing of the stakeholder, but to learn what the stakeholders think about how to gauge their

[9] See Douglas (2005) for an early overview of the efforts and their rationale, and Schrögel and Kolleck (2019) and Elliott (2017) (chapter 7) for more recent surveys.

[10] Rolin (2009), Brugha and Varvasovsky (2000) and Abelson et al. (2016).

wellbeing and to take this information into account. So it is a meta-demand to reflect the values of your stakeholders in the methodology of your research. Strategy Three shares with Strategy One the idea that concepts sometimes have to be engineered for purposes of research, rather than inherited, and it shares with Strategy Two the desire to preserve their evaluative thickness. But they have to be engineered responsibly.[11]

In today's landscape we see two kinds of attempts to implement Strategy Three: by the letter and by the spirit. Scientists follow the letter of this option when they gesture towards democratic legitimacy of their measures without actually going through any *process* of legitimation. For example, we showed above how proponents of life satisfaction sometimes defend it as the most democratic definition of wellbeing because it enables people to "assess their lives on the basis of whatever they consider important to themselves". Similarly, when Martha Nussbaum formulates the capabilities approach she too make an argument to the effect that promoting capabilities is the best way to respect citizens' autonomy (Nussbaum, 2000). These claims certainly count as attempts to give a political justification for a respective measure, rather than to make it into a technical term or to keep the value judgement in house. But arguably they do not live up to the *spirit* of Strategy Three. In case of life satisfaction, nobody asks stakeholders whether this concept is a fair representation of their views about wellbeing, whether 1–10 scales accurately measure those views, or even what determines their life satisfaction. And there is certainly no attempt here to make room for a challenge by the stakeholders of the experts. In case of capabilities, Nussbaum's self-generated list of ten has been criticised for sidestepping consultation and many capability theorists work towards implementing participatory methods for filling out the content of this approach (Robeyns, 2006).

So how could wellbeing measurement live up to the spirit, and not just the letter, of Strategy Three? This will likely differ by context. Recent efforts by capabilities theorists to democratise the operationalisation of their paradigm have often involved coproducing capabilities surveys through extensive interviewing of and discussions with communities.[12] Such an approach may be unwieldy at large scale. At national levels there have instead been consultations soliciting citizen input into what official statistics should reflect if they are to represent wellbeing of these citizens.[13]In the field of healthcare, the scale of analysis can often be a single patient, and indeed, involving patients in the production of scales representing their quality of life is increasingly standard practice.[14] Some settings may call for a mixed approach. For example, scholars of educational guidance and counselling have also recently trialled what they call a 'stakeholder-responsive approach

[11] Conceptual engineering is a familiar proposal in philosophy of wellbeing, but its advocates do not typically consider the need for stakeholder input, see Prinzing (2020) and Tiberius and Hall (2010).

[12] See, for example, Yap and Yu (2016) and Greco et al. (2015).

[13] These have taken place in the UK, New Zealand, Germany, among others. See this FOI press release by the UK's Office of National Statistics about the process they follow: https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/uknationalwellbeingindex

[14] Baron et al. (2021), Abelson et al. (2016), Harvard et al. (2020) and Degeling et al. (2015).

to researching wellbeing' (Daniels et al., 2018). These practices can differ a great deal in their scale and scope – they can be interviews, surveys, consultations, focus groups, or citizen fora – but they share an intention to democratise well-being in one way or another.

2.4 Choosing between the three strategies

We submit that these strategies likely exhaust the presently available options of dealing with thick concepts and that, strictly speaking, they are mutually exclusive. If you reject the technical term approach of Strategy One, then you have to make a decision about the source of evaluative content in your thick concepts. One source can be the intellectual decision taken by yourself (or perhaps your immediate research community) – Strategy Two – and another source can be a political process involving more than just the experts – Strategy Three. In reality it might be difficult to classify each instance of actual research as falling into one and only one of the three spaces. We have found that the same project can mix the rhetoric of two or three of our strategies, because researchers will not always invest the resources needed for formulating their strategy carefully and with full consideration. It is not uncommon to claim both that life satisfaction is a technical term while also making a brief appeal to its democratic credentials. This stance is logically possible, but strictly speaking, one or the other reason has to be a primary justification for the use of a given concept.

Now we are in a position to evaluate their strengths and weaknesses. Each of the three strategies has a long history. As a result, each is well integrated into existing practices that researchers presumably regard as well motivated and useful. So it would be unwise to take an uncompromising approach presenting one strategy as uniquely superior always and everywhere, while debunking all others. We submit that there may be good reasons to pursue any of the three strategies depending on circumstances. However, we shall present what we see as serious short-comings of Strategies One and Two for research close to policy and law. In those cases, treating thick concepts as technical terms amounts to abrogating responsibility that scientists have to anticipate and forestall misuse of their work. Imposing researchers' own value judgments as per Strategy Two raises dangers of coercion. These issues will not always trump all considerations, but they are substantial weaknesses nevertheless. Let us see why in more detail.

Strategy One seeks to rid science of thick concepts altogether. This strategy stakes the authority of science in its ability to live up to the ideal of value-freedom, or rather a specific sub-ideal of it – neutrality (Lacey, 2004). Neutrality demands that claims of science neither presuppose nor imply moral, political, or aesthetic judgments. Thick concepts fail the test of neutrality and are therefore illegitimate.[15] This harsh stance is frequently justified by empirical claims that failures of neutrality are

[15] Advocates of using subjective well-being measures in public policy seem animated by these concerns. Diener and Seligman (2004, p. 1-2), for example, argue that: "we believe that measures of well-being are—and must be— exactly as neutral politically as are economic indicators.

dangerous and will undermine public trust in science (Arneson, 2019; Haack, 2007). We are not convinced. Empirical studies show that public trust in science responds to many different factors (Rutjens et al., 2018). Scientists' refusal to handle concepts that are meaningful and significant to the public could plausibly undermine this trust as well.

The key consideration we are marshalling here comes from two sources: a general responsibility of scientists to the communities that support them and a specific responsibility generated by thick concepts. The first kind arises out of what Heather Douglas calls "the moral terrain of science", that is the network of duties scientists acquire due to their status as producers of powerful and valuable knowledge within the constraints of broader societal good (Douglas, 2014). The second source is Max Weber's demand that social scientists have a responsibility to investigate phenomena that are 'significant' to people, where significance reflects a subjective dimension of communal living (Weber, 1949). Because of this responsibility, social scientists do not have the freedom to convert concepts into technical terms (he thought this was a contrast with natural scientists who do have such a freedom, but we need not follow Weber in this thought). This is not the only responsibility social scientists have and there may be other responsibilities that will conflict with this one. However, the general idea stands - other things being equal, it is good for science to study phenomena that are significant to communities that enable their work.

If we accept this constraint, we can ask what it means for scientists to fulfil this responsibility. Does it mean they get to pick a significant phenomenon such as well-being and define it as an expert would? This brings us to what is wrong with Strategy Two. Defining a thick term takes conceptual and empirical work – what is well-being? How does it relate to being good or being healthy? How can we know when we are well? Answering these questions has been the province of philosophy, literary fiction, religion, personal reflection, psychotherapy, and more recently science. But there are no uncontroversial answers to these questions, and there is thus no definition of wellbeing that is obviously and uniquely superior to all else (Alexandrova, 2017). So it takes some hubris for scientists to pursue Strategy Two. Scientists who keep value judgments in-house may be doing so for reasons of convenience and speed, but they should not be doing so because they take themselves to be the sole and the best experts about well-being. This expertise is in fact distributed.

A proponent of Strategy Two might retort in two ways. First, responsible scientists do their homework and do not just consult their untutored intuitions when picking a definition of well-being. Secondly, they may argue that adopting a given conception of wellbeing does not reflect a conviction that it is the correct one, but just a belief that it is a significant conception for science to investigate. Neither of these replies justify Strategy Two. Scientists can be very thoughtful about the conceptions they adopt: Kahneman cites Bentham as his intellectual inspiration for 'objective happiness', the capabilities theorists cite Aristotle, and life satisfaction advocates too have their standard list of references (Tatarkiewicz, 1976; Sumner, 1996). But it is one thing to identify a lineage for your favourite theory and it is another to show that your choice has legitimacy in the public sphere. For the latter task, lineage, no matter how eminent, is not enough. There is still a danger that the chosen theory does not reflect the values of the people you study. Nor does the judgment of significance made in-house, to which the second reply appeals, has the legitimacy it could have if it was made inclusively.

This is why Strategy Three emerges as most attractive when the research in question is close to action. Sometimes the benefits of Strategies One or Two outweigh their costs. For example, Strategy One is acceptable when the study is highly theoretical, exploring uncharted areas, and far from applications, while Strategy Two can conceivably be justified when the precise definition of wellbeing does not matter because, say, the empirical effect is so robust that it holds on any definition of wellbeing. But outside these contexts, Strategy Three has a prima facie advantage of being upfront about the evaluative content (unlike Strategy One) and being responsible about the limits of scientific judgment (unlike Strategy Two).[16]

What does it take to implement Strategy Three for measurement?

3 Part II: Implementing legitimacy

Our goal in this section is to articulate a plausible and a realistic ideal of participatory measurement, for this is a way to implement the spirit not just the letter of Strategy Three. We start on the basis of an account of measurement built specially for social and medical sciences and then build a participatory element into this account.

3.1 A theory of measurement for thick concepts

An influential account of measurement by Norman Bradburn, Nancy Cartwright, and Jonathan Fuller requires that the process of constructing and justifying measures, especially in sciences of policy and healthcare, fulfils three desiderata:

1. We define the concept or quantity, identifying its boundaries, fixing which features belong to it and which do not (**characterization**).
2. We define a metrical system that appropriately represents the quantity or concept (**representation**).
3. We formulate rules for applying the metrical system to tokens to produce the measurement results (**procedures**). (Bradburn et al., 2017, p.3)

This account is a good starting point for us because it pulls together ingredients of measurement that are normally treated separately. It also treats all three requirements as equal, in contrast to the earlier theories that focused on representation almost uniquely (Suppes, 1998). This account is consistent with other influential views of measurement such as the model-based account, which conceives of measurement as a coupling between two ingredients: 1) a concrete process of

[16] Our argument here is similar to Haybron and Tiberius (2015) who argue that in the context of public policy researchers should adopt a subjective conception of wellbeing, steering maximally close to citizen values. We are taking this line of thought further, recommending that citizens also should be able to vet the construct and measures researchers adopt.

interaction between an instrument and the environment and 2) an abstract model that represents this process (Tal, 2020, Section 7). Bradburn et al.'s theory is helpful because it unpacks more deeply the stages of construction of such a process and the corresponding model and it does so in a way that is recognisable to social and medical scientists. It is thus unsurprising that this three-stage account is also consistent with the textbook recommendations for measure development, validation, and implementation (de Vet et al., 2011).

Let us now see how the three-part framework applies to measurement of wellbeing. To fulfil characterisation, wellbeing needs to be defined first as a concept. The questions to ask at this stage include: is wellbeing predicated of an individual or a community? Does it encompass just welfare or also justice? Is the wellbeing in question all-things-considered or focused only on a specific context, like the wellbeing of newborns? Secondly, researchers need to decide what states or processes in the world realise this concept: are they people's aggregated subjective states and if so which states exactly? Or are they the states that describe objective features of their lives and if so which features? Or are they some combination of subjective and objective indicators? Or perhaps they are not states at all but processes (McClimans & Browne, 2012). This is the point at which heavy-duty theorising must take place and the various philosophical theories of wellbeing play an essential role.

Moving to the second stage of representation, the wellbeing states or processes identified as relevant at the stage of characterisation must be connected to observable indicators whose values should fall along a scale. There are agreed upon conventions about the nature of these scales: they can be ordinal, interval, or ratio. In wellbeing it is rare to see fully interval scales, let alone ratio scales, and ordinal scales are most common. The indicators making up these scales can be subjective reports of, for example, happiness or life satisfaction, objective indicators of quality of life, or some combination of the above, provided there is a credible story about how variation in the value of these indicators enables their comparison. This is the stage at which the numerical structure of the indicators needs to be shown to correspond to the structure of wellbeing as specified at the stage of characterisation. This is normally accomplished by techniques such as representation theorems, or Rasch modelling, or more controversially construct validation (Vessonen, 2020; Alexandrova, 2017 chapter 5). This stage is usually considered the business of psychometrics or metrology more generally.

At the final stage, measurement requires clear and comprehensive procedures. For example, if wellbeing is characterised by a certain class of mental states represented by self-reports, how are those self-reports to be collected and collated into usable data, by whom and under what circumstances?

Now that we see the overall shape of measurement, we can ask what it would mean to make this process participatory in spirit, not just the letter. Making sure that measures of wellbeing respond to people's priorities takes more than just using subjective and maximally open indicators such as life satisfaction. Stakeholders also need to have a real say about the survey items and how their answers are used to ascribe to them a particular level of wellbeing. This input needs to fit in with the above three-stage theory of measurement. To flesh out how this is supposed to work we turn to the concept of *co-production* because it is uniquely attuned to the necessity of attending to different kinds of expertise in different stages of scientific research.

3.2 Joining measurement and co-production

Co-production is a term with fuzzy meaning used in several fields often to mean different things. In the hands of public policy, public administration, healthcare, and technology scholars it describes a model of governance, care, and service provision that involves users in all aspects of design, delivery, and evaluation (Osborne et al., 2016). In science and technology studies, co-production captures the fact that scientific theories, instruments, and other products emerge from a complex interplay of nature, researchers, users, institutions, audiences (Jasanoff, 2004). These uses converge on the ambition of bottom-up collaborative work, whether in science, policy, or design. Our focus on measurement of thick concepts necessitates a bespoke definition of co-production, hereafter *co-production**, based on these existing ones. We are neither producing a service or a policy, nor making an empirical claim about the nature of the scientific process. Rather we are looking for a normative account of responsible measurement when phenomena are denoted by thick concepts. Hence we propose the following definition:

> **Co-production*** is an arrangement for sharing power and responsibility in the process of defining thick concepts and developing their measures. This arrangement requires, first, recognising different types of expertise that each group of stakeholders have about these concepts and their measurement and, second, ensuring that the final products meet, to the extent that it is possible, the demands stemming from each type of expertise.

Let us unpack each element of this definition for our example of wellbeing. When a project adopts a definition of wellbeing, the power resides in the possibility of using this definition to alter people's lives through policy, healthcare, and services. For example, recent work in happiness economics in the UK identifies mental illness as the strongest determinant of life satisfaction and urges provision of cognitive behavioural therapy as the most cost effective policy (Clark et al., 2018). Such a policy recommendation naturally comes with all the attendant consequences – redirection of welfare spending, redesign of services, and possibly even coercion, such as when CBT becomes a condition of unemployment benefits (Friedli & Stearn, 2015). In this case, scholarly responsibility requires thinking through the consequences of one's research once its results enter into the public sphere and policy discourse. When researchers produce knowledge about wellbeing, it is on them to watch out for unintended harmful consequences of this knowledge, at least to the extent that it is foreseeable. These are well known and uncontroversial constraints on science, whether it concerns physics of weapons, biology of viruses, or determinants of wellbeing (Douglas, 2003, 2014). A measurement process 'shares' this power and responsibility when it is organised in a way that distributes them among all stakeholders. All stakeholders should have a say in the conceptualisation and measurement of wellbeing to the extent that their distinctive expertise allows. And if they have a say, they acquire a responsibility for consequences of this definition.

The next crucial clarification is who counts as a stakeholder in projects that involve thick concepts. We follow a definition of a stakeholder for contexts of research rather than for corporate or management contexts. Such definitions generally identify stakeholders with individuals or organisations that stand to benefit or to be harmed by a research project to a reasonably foreseeable extent.[17] Although these discussions often draw a distinction between scientists and stakeholders, for purposes of measurement of wellbeing such a distinction is unsuitable. Scientists are stakeholders – it matters to them that wellbeing be measured as well as it could be – and so are people and organisations outside academia. So we propose three very general classes of stakeholders for our particular focus:

a) Members of the public, especially service users
b) Policy makers and service providers
c) Scholarly researchers

This is a natural division within contemporary evidence-based policy. Academic researchers are supposed to produce knowledge that gets translated into practice by policy makers and service providers, with the goal of improving outcomes for members of the public (Marmot, 2004). Of course, sometimes the researcher is also the policy maker and a member of the public. So this distinction is between *roles* different groups occupy, not between stable categories in which they belong.

Co-production must recognise that, when it comes to thick concepts, people playing these three roles bring distinctive expertise, as we summarise in Table 1. Members of the public are typically the ones whose wellbeing is being studied and their perspective on their own wellbeing is clearly of unique significance. In this role people have what is sometimes called 'lived expertise', in the sense that their knowledge of wellbeing comes from navigating daily tasks of life often from the vantage point of their own circumstances such as disability, poverty, or another source of perspective (Park, 2020). This is in contrast to the role of scholarly researchers for whom wellbeing and measurement are objects of technical study undertaken at universities or think tanks. Their expertise covers existing definitions of wellbeing from scholarly literatures, the standard measures used in different disciplines, and how these measures are tested and validated. Finally, policy makers and service providers represent a distinctive *professional* expertise about how the world of politics and science gets translated into actual institutions, therapies, and initiatives on the ground. This expertise includes an understanding of implementation and the nitty gritty of applying thick concepts in real world policy.

[17] See this HEFCE guide to stakeholder analysis in the UK https://www.vitae.ac.uk/doing-research/leadership-development-for-principal-investigators-pis/leading-a-research-project/applying-for-research-funding/research-project-stakeholders. In the environmental and climate research the definitions of stakeholders are developed specifically for conservation and waste management projects and involve any consideration from economic, to health risks, and aesthetic appreciation of nature (Burger, 2011, p9).

Table 1 Stakeholders and expertise about thick concepts

Stakeholder role	Distinctive expertise
Members of public	Lived expertise
Scholarly researchers	Technical expertise about theories and measurement
Policy makers and service-providers	Professional expertise about delivery and implementation

Each type of expertise is relevant to measurement and a good measure of a thick concept is one that emerges when the three sets of experts learn from each other in an equal and productive arrangement, where no expertise dominates another.

Let us now extend this model to measurement. The key is to show how our three kinds of expertise contribute to the three demands on measurement, that is characterisation, representation, and procedures. Table 2 captures the challenge:

We add the rightmost column to show that experts in each of the three roles contribute to each of the three elements of measurement. However, experts in different roles are likely to have different levels of investment into these elements and their contribution will be distinctive at each level. We bolded those elements of measurement that different experts are likely to attend to more than others in virtue of their knowledge, but without implying that they cannot also make distinctive contributions at all three stages.

Lived experience gives members of the public a unique purchase on characterisation of whatever thick concept is in question. This experience is essential for articulating the content and the boundaries of the concept as characterisation requires. However, this lived expertise does not typically extend to representation. Representation demands quantification that is not normally present in daily life. Procedures, on the other hand, are likely to be more visible to those members of the public that are on the receiving end of measurement. They are the ones who will be filling out the surveys and pondering how to reflect their views within the constraints of questionnaire items.

Table 2 Stakeholders, expertise, and measurement

Stakeholder role	Distinctive expertise	Contribution to measurement
Members of public	Lived expertise	**Characterisation** Representation **Procedures**
Scholarly researchers	Technical expertise about theories and measurement	**Characterisation** **Representation** Procedures
Policy makers and service-providers	Professional expertise about delivery and implementation	Characterisation Representation **Procedures**

Scholarly researchers are likely to have a lot to say about characterisation and representation. In the wellbeing sciences, they will be familiar with different theoretical approaches such as hedonism, subjectivism, and eudaimonism, since these are typically the starting points of all the existing constructs. Academics are also supposed to have a grip on representation, the most technical and esoteric aspect of measurement. Although they might have views on the third element, that is measurement procedures, unless they regularly administer surveys themselves, they do not have a first-hand experience of this. Academics do not typically spend a lot of their time and attention on what happens to their questionnaires once they get deployed in the world outside of research.

Finally, we hypothesise that professional expertise gives a special purchase on the procedures and less so on characterisation and representation. Service providers and policy makers are on the implementation end of things and they invest energies into delivery platforms of surveys and their operation. They are especially attuned to clarity of survey items, their lengthiness, and the ways they might alienate people. They would be aware, for example, of whether qualitative or quantitative measures would be more useful to service providers.

The idea behind co-production* it is to bring out different types of expertise as they map onto the different elements of measurement. Since no group of experts is in the driver's seat, all can contribute everywhere. But the point of recognising different types of expertise is to allow that some of us know more about some aspects of measurement than others. Even when we lack expertise about characterisation, representation, or procedures, it is good to have oversight from people playing different roles. The hope is that when the process of co-production is organised and managed well, the impact of each expertise is maximised. There is a learning process in all directions. The emerging measure consequently has the best chance of meeting all three demands: the phenomenon is well characterised, faithfully represented, and there are effective procedures for gauging it. Such a learning process may well show that there are trade-offs between characterisation, representation, and procedures. True wellbeing may not be quantifiable, or a true quantity may not be measurable through realistically available procedures. Coproduction* may turn up a measure that is deeply compromised but nonetheless fit for its context-specific purpose, or no measure at all.[18] Our point is that, if such participatory measurement is at all theoretically justifiable, it should have the shape we have described here. As it happens we do believe this ideal is realistic and we now move on to illustrate this.

3.3 Case study of ongoing project with Turn2us

The theory above is informed by our experiences collaborating with Turn2us, a national anti-poverty charity in the UK. Turn2us has a wide range of activities that all fall under the banner of helping people who come upon hard times financially. Their work includes issuing emergency grants that enable people to cover bills,

[18] Philosophers of measurement have explored such conflicts and trade-offs in Larroulet-Philippi (2021) and Vessonen (2021a) among other places.

helping people with the often confusing and stressful process of applying for welfare benefits (this is accomplished through an online-platform called the Benefits Calculator), and campaigning for policy reforms that would reduce poverty. Turn2us has a wealth of experience with coproduction of their services and they invited us to participate in the development of a concept and measure of 'thriving'. They were interested in what thriving means in the context of financial hardship and how they could monitor the impact of their activities on the thriving of their clients. A close relative of 'wellbeing' and 'flourishing', thriving is a thick concept with a temporal dimension – it is an effort to learn to live well over time. How can such a concept be coproduced?

In conversation with Turn2us we developed a blueprint of the process with the following key stages:

Survey 1 ➔ Working group ➔ Workshop ➔ Survey 2

Survey 1 was distributed using Turn2us' fortnightly newsletter and received 1550 responses from users of Turn2us' services. It asked them about their conception of thriving. Alongside an open ended question about what thriving means to you, this survey elicited respondents' attitudes to classic theories of wellbeing. It also posed some conventional questions about what aspects of wellbeing respondents' valued relatively more (such as feelings of purpose or good mood). But we were especially keen to hear what they feel others misunderstand about thriving of people in their circumstances. We brought the results of this survey to the working group to give it an initial steer and inform its deliberations.

The working group was selected to represent equally the three groups of stakeholders of this exercise: 1) people whose thriving is or was undermined by sudden financial insecurity, 2) the Turn2us employees, and 3) scholars who study thriving and poverty. These three groups represent three corresponding types of expertise: lived expertise, professional expertise, and technical expertise. The remit of the working group was to develop a measure of thriving in an intense and equitable deliberative process. The group thus had to be small enough to build a trusting rapport and to enable in-depth discussion and one-to-one interviews, but big enough so that each expertise is sufficiently represented. In a series of meetings chaired by the Turn2us coproduction lead Abby Meadows, the working group accomplished the following tasks:

- examined the results of the initial survey to get clear on the priorities of the users of Turn2us.
- set out the terms of the interviews wherein each participants interviewed at least one member of each expert group to which they do not belong, focusing on what thriving means to them. We borrowed ideas from the practice of "relational interviewing" for this process, which emphasises genuine power-sharing and two-way learning between participants and eliminates the distinction between interviewee and interviewer (Fujii, 2017; Hydén, 2014).

- After 23 interviews were conducted, the academics on the team performed qualitative analysis of the themes and presented these themes to the working group as a whole.
- The group then worked towards systematising these themes and ensuring they conformed with the lived experience of the service users and the practical needs of Turn2us practitioners. This involved the academic group presenting their thematic analyses and organising theories to the rest of the group for debate and refinement.

Once we had a consensus within the working group on a preliminary theory of thriving, we took it first to Turn2us' board of directors for input and then to a larger workshop. Here the working group was joined by an additional 12 lived experts who scrutinised it and offered suggestions for improvement. In the event, most of these concerned the language of the theory and its presentation, rather than elements of the theory itself. These suggestions were incorporated into a final report that was then approved by workshop participants. That report was then again presented to an online survey through Turn2us' newsletter for endorsement. This methodology was designed to balance, at least to some extent, the high logistical demands of engaging in depth with expert groups to formulate a rich and context-sensitive theory of thriving, and the need for the theory to be representative. The working group and workshop processes provided the depth, while the surveys at either end enhanced representativeness.

Such was the process. The substantive theory and measure of thriving developed in this process is available on the Turn2us webpage.[19] Here we report only enough to illustrate the practical implementation of the model of coproduction* proposed in Section II.2. As the model recommends, we identified different types of expertise corresponding to the different roles of stakeholders. Turn2us had a wealth of experience with coproduction and they recruited lived and professional experts who had the experience and the availability to engage in the lengthy and detailed discussions. Our model of coproduction* also specifies that each type of expert knowledge be accorded respect and equality vis-à-vis others. To ensure healthy power dynamics in the working group, the chair compiled a coproduction social contract that enforced norms of respect and forestalled dominating behaviour by any members of the group. Substantive grant funds were dedicated to providing payments (the hourly equivalent to London living wage) to the coproduction partners for the time they gave the exercise. The published outputs on thriving are planned so that the coproduction partners get credit as co-authors on reports and articles. Together these actions help to create a sense of trust and partnership and enable genuine learning in all directions: lived experts to professional experts, professional experts to academics, and so on.

[19] The detailed description of the coproduced theory and our methodology can be found in Fabian et al. 2021, while the public facing report and visual aids are available on the website of Turn2us: https://www.turn2us.org.uk/About-Us/Media-Centre/Research-and-Insights/Thriving

The other key aspect of our model of coproduction* is the distinction between characterisation, representation, and procedures we inherit from Bradburn, Cartwright and Fuller. How is this distinction reflected in our work with Turn2us? It is fair to say that the exercise as conducted so far has focused mostly on characterisation of thriving, some on procedures, and less on representation. While Turn2us is interested in measuring thriving to track their effectiveness, we quickly realised that there should not be one such measure for all aspects of their work. Instead, different activities of this charity call for different levels of quantification and varieties of appraisal. The specific application of each measure bears heavily on how it should be formulated. Indeed, Turn2us has found that off-the-shelf measures developed by academics are unsuitable to its operations. In particular, capabilities surveys are too onerous to impose on someone desperately seeking financial help, and the charity has found that subjective wellbeing questions are insufficiently sensitive to changes in respondent circumstances as a result of Turn2us interventions. So it seems bespoke measures are required, but Turn2us wants these to emerge organically as it goes about applying the theory of thriving in its operations.

So at the time of writing this article the working group had developed a construct of thriving under financial insecurity with some indications about how it can be gauged, but without yet a fully validated scale of it. We have devised potential questionnaire items and formulated ways in which these items can be integrated into the activities of Turn2us, but this does not yet meet the standard of representation and procedures as formulated in the Bradburn et al theory. Still, even recognising these limits, our experience with Turn2us serves as evidence that coproduction of thick concepts such as thriving is possible.

4 Part III: Discussion and Conclusion

In Part I we argued that it is desirable to co-produce measures of phenomena denoted by thick concepts. In Part II we showed, using the example of thriving under financial hardship, that with due support and preparation, it is feasible to implement a process of coproduction that meets the spirit as well as the letter of the theory in Part I. In conclusion we comment on the limits of our proposal and put it in a wider context of participatory methods and wellbeing sciences.

Our primary focus has been the production of a measure for a specific context.[20] Its value is in focusing on the distinctive needs of Turn2us, which enabled a deep deliberative engagement across all stakeholders. The conception of thriving we were able to articulate is more detailed and in line with what Alexandrova (2017) calls mid-level theories of wellbeing: theories geared to a specific group of people in a specific context, rather than the general homo sapiens. This grounded nature might even be what makes this concept thick rather than thin (Abend, 2019). However, we concede that such contextuality will not always be possible or indeed desirable.

[20] See Scott and Bell (2013) and Sollis et al. (2021) on other participatory efforts to develop local indicators of wellbeing.

Sometimes stakeholders are a far bigger and more diverse population and the purposes of the measures are less specific. This is the case for national or international efforts to develop wellbeing statistics. In those cases indicators are validated through country-wide consultations and expert input. Without necessarily endorsing these initiatives, we nevertheless acknowledge that coproduction* may not be right for these purposes. At the same time our approach taking a general thick concept and converting it to a locally legitimate measure should be implementable far beyond thriving or wellbeing.

Another potential weakness of our proposal applies to all participatory approaches. They can easily turn into box ticking exercises that reify their public without recognising their variability and fluidity (Chilvers & Kearnes, 2020). Worse even idealistic pursuits like citizen science can be highjacked by special interests and play the role of public relations, providing its initiators a show of legitimacy where in fact there is none (Blacker et al., 2021). There are no simple fixes to these problems. Co-production* will only safeguard legitimacy of thick concepts if the process is implemented with care and due respect for the expertise of all involved. Our theoretical model and our case study with Turn2us is a bona fide attempt to do so.

Philosophers of science will see other limits in our model. Coproduction* presumes that it will be possible to safeguard the high scholarly standards of measurement while opening it up for lay participation. Measurement and validation are some of the most technical areas of science. Judging whether or not a given measure performs at all ends of the scale and meets the long list of validities that metrology demands takes intricate expertise. How realistic is it to expect all stakeholders to engage with these questions? Aren't we opening the door to the possibility of coproduced measures of poor technical validity?

Here too we gladly acknowledge that our model, in allowing stakeholder input at all levels of measurement, does potentially invite compromises. But we think such compromises are worth considering if we are to avoid giving one group of experts – namely metrologists and psychometricians – undue authority. In the health sciences it is common for patient groups to contribute to the initial stage of scale design. However, psychometric validation, by virtue of coming last in the process of measure construction, often overrides the judgments of patients with lived experience. The patients may believe that a certain ability is crucial to their quality of life with their medical condition, but if the item representing this ability does not have the right statistical properties, it can be dumped.[21] This practice may sometimes be appropriate but it is hard to defend universally. Historians and philosophers of measurement have shown time and again the many unformalizable and controversial judgment calls that enter this process (Chang, 2004; McClimans, 2017; Stegenga, 2015b). Our view is that, when it comes to thick concepts and life-changing policies, it is a good idea to open up these judgment calls to a wider set of experts, including people themselves.

[21] See De Vet et al. (2011) for the textbook description of this process and Alexandrova (2017, chapter 6) for a critique.

Acknowledgements Funding for this research from ESRC/AHRC under grant number ES/T005556/1 is gratefully acknowledged, as is the assistance of What Works Centre Wellbeing and our colleagues at the Bennett Institute for Public Policy. We also thank Turn@us and coproduction partners for their invaluable collaboration. This paper benefited from advice of two exceptionally generous referees for this journal and numerous comments from members of audiences in the seminars in which we presented earlier versions.

Declarations

Ethical statement The work reported was funded by the Economic and Social Research Council Grant ES/T005556/1 "The Many Dimensions of Wellbeing". There are no conflicts of interests to report. Ethical approval was granted by University of Cambridge.

References

Abend, G. (2019). Thick concepts and sociological research. *Sociological Theory, 37*(3), 209–233.

Abelson, J., Wagner, F., DeJean, D., Boesveld, S., Gauvin, F., Bean, S., et al. (2016). Public and patient involvement in health technology assessment. *International Journal of Technology Assessment in Health Care, 32*(4), 256–264. https://doi.org/10.1017/S0266462316000362

Adler, M., & Fleurbaey, M. (2016). *The Oxford handbook of well-being and public policy*. Oxford University Press.

Alexandrova, A. (2017). *A philosophy for the science of well-being*. Oxford University Press.

Alkire, S., Foster, J. E., Seth, S., Santos, M. E., Roche, J. M., & Ballon, P. (2015). *Multidimensional poverty measurement and analysis: A counting approach.* Oxford University Press.

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia, 19*(1), 1–24.

Anderson, E. (2002). Situated knowledge and the interplay of value judgments and evidence in scientific inquiry. In *In the scope of logic, methodology and philosophy of science* (pp. 497–517). Springer.

Arneson, D. (2019). Comments on Anna Alexandrova, a philosophy for the science of well-being. *Res Philosophica, 96*(4), 513–520.

Baron, M., Riva, M., Fletcher, C., Lynch, M., Lyonnais, M.-C., & Laouan Sidi, E. A. (2021). Conceptualisation and operationalisation of a holistic indicator of health for older inuit: Results of a sequential mixed-methods project. *Social Indicators Research, 155,* 47–72.

Blacker, S., Kimura, A. H., & Kinchy, A. (2021). When citizen science is public relations. *Social Studies of Science, 51*(5), 780–796.

Bradburn, N. M., Cartwright, N., & Fuller, J. (2017). A theory of measurement. In L. McClimans (Ed.), *Measurement in medicine: Philosophical essays on assessment and evaluation* (pp. 73–88). Rowman and Littlefield International.

Brown, M. J. (2020). *Science and moral imagination: A new ideal for values in science*. University of Pittsburgh Press.

Brugha, R., & Varvasovszky, Z. (2000). Stakeholder analysis: A review. *Health Policy and Planning, 15*(3), 239–246. https://doi.org/10.1093/heapol/15.3.239

Burger, J. (Ed.) (2011). *Stakeholders and scientists: Achieving implementable solutions to energy and environmental issues*. Springer Science & Business Media.

Carnap, R. (1950). Empiricism, semantics, and ontology. *Revue Internationale De Philosophie, 4*(11), 20–40.

Chang, H. (2004). *Inventing temperature*. Oxford University Press.

Chilvers, J., & Kearnes, M. (2020). Remaking participation in science and democracy. *Science, Technology, & Human Values, 45*(3), 347–380.

Clark, A., Flèche, S., Layard, R., Powdthavee, N., & Ward, G. (2018). *The origins of happiness*. Princeton University Press.

Cohen Kaminitz, S. (2018). Happiness studies and the problem of interpersonal comparisons of satisfaction: Two histories, three approaches. *Journal of Happiness Studies, 19*(1), 423–442.

Daniels, K., Connolly, S., Ogbonnaya, C., Tregaskis, O., Bryan, M. L., Robinson-Pant, A., & Street, J. (2018). Democratisation of wellbeing: Stakeholder perspectives on policy priorities for improving national wellbeing through paid employment and adult learning. *British Journal of Guidance & Counselling, 46*(4), 492–511.

De Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge university press.

Degeling, C., Carter, S. M., & Rychetnik, L. (2015). Which public and why deliberate?–A scoping review of public deliberation in public health and health policy research. *Social Science & Medicine, 131*, 114–121.

Deringer, W. (2018). *Calculated values: Finance, politics, and the quantitative age*. Harvard University Press.

Djordjevic, C., & Herfeld, C. (2021). Thick concepts in economics: The case of Becker and Murphy's theory of rational addiction. *Philosophy of the Social Sciences, 51*(4), 371–399. https://doi.org/10.1177/00483931211008541.

Diener, E., & Seligman, M. (2004). Beyond money: Towards an economy of well-being. *Psychological Science in the Public Interest, 5*(1), 1–31.

Diener, E., Lucas, R., Schimmack, U., & Helliwell, J. (2009). *Well-Being for Public Policy*. Oxford University Press.

Dolan, P., & Peasgood, T. (2008). Measuring well-being for public policy: Preferences or experiences? *Journal of Legal Studies, 37,* 5–31.

Dolan, P., & White, M. (2007). How can measures of subjective well-being be used to inform public policy? *Perspectives on Psychological Science, 2*(1), 71–85.

Douglas, H. E. (2003). The moral responsibilities of scientists (tensions between autonomy and responsibility). *American Philosophical Quarterly, 40*(1), 59–68.

Douglas, H.E (2005) "Inserting the Public into Science" in *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making, Sociology of the Sciences vol. 24*, Sabine Maasen and Peter Weingart (eds.), Springer, pp. 153-169.

Douglas, H. (2011). Facts, Values. In J. I. Objectivity & J. Zamora Bonilla (Eds.), *The SAGE Handbook of Philosophy of Social Science* (pp. 513–529). SAGE Publications.

Douglas, H. (2014). The moral terrain of science. *Erkenntnis, 79*(5), 961–979.

Dupré, J. (2007). Fact and value. In H. Kincaid, J. Dupré, & A. Wylie (Eds.), *Value-Free Science? Ideals and Illusions* (pp. 27–41). Oxford University Press.

Elliott, K. C. (2017). *A tapestry of values: An introduction to values in science*. Oxford University Press.

Fabian, M., Alexandrova, A., Cinamon-Nair, Y., & Turn2us (2021). A coproduced theory of 'thriving'for people experiencing financial hardship. Bennett Institute for Public Policy Working Paper. https://www.bennettinstitute.cam.ac.uk/media/uploads/files/Working_paper_A_Coproduced_Theory_of_Thriving.pdf. Accessed 5 Jan 2022.

Feldman, F. (2010) *What is this thing called happiness?*. OUP Oxford.

Friedli, L., & Stearn, R. (2015). Positive affect as coercive strategy: Conditionality, activation and the role of psychology in UK government workfare programmes. *Medical Humanities, 41*(1), 40–47.

Friedman, M. (2001). *Dynamics of reason*. Csli Publications.

Frijters, P., Clark, A. E., Krekel, C., & Layard, R. (2020). A happy choice: Wellbeing as the goal of government. *Behavioural Public Policy, 4*(2), 126–165.

Fujii, L. (2017). *Interviewing in social science research: A relational approach*. Routledge.

Greco, G., Skordis-Worrall, J., Mkandawire, B., & Mills, A. (2015). What is a good life? Selecting capabilities to assess women's quality of life in rural Malawi. *Social Science & Medicine, 130*, 69–78.

Haack, S. (2007). The integrity of science: What it means, why it matters. *Contrastes: Revista Internacional de Filosofia (Spain), 12*, 5–26.

Harvard, S., Werker, G. R., & Silva, D. S. (2020). Social, ethical, and other value judgments in health economics modelling. *Social Science & Medicine, 253*, 112975.

Hawthorne, S. (2013). *Accidental intolerance: How we stigmatize ADHD and how we can stop*. Oxford University Press.

Haybron, D. M. (2008). *The pursuit of unhappiness: The elusive psychology of well-being*. Oxford University Press.

Haybron, D. M., & Tiberius, V. (2015). Well-being policy: What standard of well-being? *Journal of the American Philosophical Association, 1*(4), 712.

Hydén, M. (2014). The teller-focused interview: Interviewing as relational practice. *Qualitative Social Work, 13*(6), 795–812.

Jasanoff, S. (Ed.) (2004). *States of knowledge: The co-production of science and the social order*. Routledge.

Kahneman, D. (1999). Objective happiness. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 3–25). Russell Sage Foundation.

Kingma, E. (2014). Naturalism about health and disease: Adding nuance for progress. *Journal of Medicine and Philosophy, 39*(6), 590–608.

Kirchin, S. (Ed.). (2013). *Thick concepts*. Oxford University Press.

Kristjánsson, K. (2013). *Virtues and vices in positive psychology*. Cambridge University Press.

Lacey, H. (2004). *Is science value free?: Values and scientific understanding*. Psychology Press.

Larroulet-Philippi, C. (2021). Valid for what? On the very idea of unconditional validity. *Philosophy of the Social Sciences, 51*(2), 151–175. https://doi.org/10.1177/0048393120971169

Marmot, M. (2004). Evidence based policy or policy based evidence? *BMJ, 328*(1), 906–907.

McClimans, L., & Browne, J. P. (2012). Quality of life is a process not an outcome. *Theoretical medicine and bioethics, 33*(4), 279–292. https://doi.org/10.1007/s11017-012-9227-z

McClimans, Leah (2017) ed. *Measurement in medicine: Philosophical essays on assessment and evaluation*. Rowman & Littlefield.

Nagel, Ernst. *The structure of science* (1961).

Nussbaum, M. C. (2000). *Women and Human Development: The Capabilities Approach*. Cambridge University Press.

Oishi, S., Kushlev, K., & Schimmack, U. (2018). Progressive taxation, income inequality, and happiness. *American Psychologist, 73*(2), 157–168. https://doi.org/10.1037/amp0000166

Osborne, S. P., Radnor, Z., & Strokosch, K. (2016). Co-production and the co-creation of value in public services: a suitable case for treatment? *Public management review, 18*(5), 639–653.

Park, S. E. (2020). Representative bureaucracy through staff with lived experience: Peer coproduction in the field of substance use disorder treatment. *The American Review of Public Administration, 50*(8), 880–897.

Peter, F. (2017). Political legitimacy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2017 Edition). https://plato.stanford.edu/archives/sum2017/entries/legitimacy/. Accessed 5 Jan 2022.

Prinzing, M. (2020). Positive psychology is value-laden: It's time to embrace it. *Journal of Positive Psychology*. https://doi.org/10.1080/17439760.2020.1716049

Putnam, H. (2004). *The collapse of the fact/value dichotomy and other essays*. Harvard University Press.

Reiss, J. (2010). *Error in economics: Towards a more evidence-based methodology*. Routledge.

Robeyns, I. (2006). The capability approach in practice. *Journal of political philosophy, 14*(3), 351–376.

Rolin, K. (2009). Scientific knowledge: A stakeholder theory. In J. Van Bouwel (Ed.), *The social sciences and democracy*. Palgrave Macmillan. https://doi.org/10.1057/9780230246867_4.

Root, M. (2007). Social problems. In H. Kincaid, J. Dupré, & A. Wylie (Eds.), *Value-free science? Ideals and illusions* (pp. 42–57). Oxford University Press.

Rutjens, B. T., Sutton, R. M., & van der Lee, R. (2018). Not all skepticism is equal: Exploring the ideological antecedents of science acceptance and rejection. *Personality and Social Psychology Bulletin., 44*(3), 384–405. https://doi.org/10.1177/0146167217741314.

Scott, K., & Bell, D. (2013). Trying to measure local wellbeing: indicator development as a site of discursive struggles. *Environment and Planning C: Government and Policy, 31*, 522–539.

Schrögel, P., & Kolleck, A. (2019). The many faces of participation in science: Literature review and proposal for a three-dimensional framework. *Science & Technology Studies, 32*(2), 77–99. https://doi.org/10.23987/sts.59519.

Seligman, M. E. (2012). *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster.

Sollis, K., Yap, M., Campbell, P., & Biddle, N. (2021). Conceptualisations of wellbeing and quality of life: A systematic review of participatory studies. https://doi.org/10.31235/osf.io/rfegt

Stapleford, T. A. (2009). *The cost of living in America: A political history of economic statistics, 1880-2000*. Cambridge University Press.

Stegenga, J. (2015a). Effectiveness of medical interventions. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 54*, 34–44.

Stegenga, J. (2015b). Measuring effectiveness. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 54*, 62–71.

Sumner, L. W. (1996). *Welfare, happiness, and ethics*. Clarendon Press.

Suppes, P. (1998). Measurement, theory of. In *The Routledge Encyclopedia of Philosophy*. Taylor and Francis. Retrieved 5 Jan 2022, from https://www.rep.routledge.com/articles/thematic/measurement-theory-of/v-1. https://doi.org/10.4324/9780415249126-Q066-1

Tal, E. (2020). Measurement in Science. E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/measurement-science/. Accessed 5 Jan 2022.

Tatarkiewicz, W. (1976). Analysis of happiness.

Tiberius, V. (2013). Well-being, wisdom, and thick-theorizing: On the division of labour between moral philosophy and positive psychology. In S. Kirchin (Ed.), *Thick concepts*. Oxford University Press.

Tiberius, V., & Hall, A. (2010). Normative theory and psychological research: Hedonism, eudaimonism, and why it matters. *The Journal of Positive Psychology, 5*(3), 212–225.

Yap, M., & Yu, E. (2016). Operationalising the capability approach: Developing culturally relevant indicators of indigenous well-being—an Australian example. *Oxford Development Studies, 44*(3), 315–331.

Väyrynen, P. (2013). *The lewd, the rude and the nasty: A study of thick concepts in ethics*. Oxford University Press.

Vessonen, E. (2021a). Respectful operationalism. *Theory & Psychology., 31*(1), 84–105. https://doi.org/10.1177/0959354320945036

Vessonen, E. (2021b). Conceptual engineering and operationalism in psychology. *Synthese*. https://doi.org/10.1007/s11229-021-03261-x

Vessonen, E. (2020). The complementarity of psychometrics and the representational theory of measurement. *The British Journal for the Philosophy of Science, 71*(2), 415–442.

Weber, M. (1949). *The methodology of the social sciences*. Edward A. Shils, & Henry A. Finch (Trans/Eds.). Free Press.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.