Sewaybricker, L. E. (2026). Bridging the gap between critical and applied approaches in the science of well-being: An analysis of evaluative subjective well-being measures. *International Journal of Wellbeing*, *16*(4), 6457, 1-24. https://doi.org/10.5502/ijw.v16i4.6457

ARTICLE

Bridging the gap between critical and applied approaches in the science of well-being: An analysis of evaluative subjective well-being measures

Luciano E. Sewaybricker

Abstract: This article examines three recurring arguments for evaluative subjective well-being (swb) measures: instruments that ask respondents to broadly assess their lives, such as the Cantril Ladder and general life satisfaction questions. To organise the debate and support better dialogue between critics and supporters of these instruments, we draw on two conceptual tools from the philosophy of well-being: the distinction between constitutive and determinative questions, and the characterisation of well-being as a thick concept. The first argument we analyse holds that leaving the object undefined is desirable (a "non-paternalistic" approach). We argue that this stance legitimately declines the constitutive question about well-being, but that in delegating it to respondents it makes engagement with that question a precondition for interpreting the resulting scores. The second argument holds that the absence of a clear definition is not disqualifying, given large samples, respondents' familiarity with the relevant expressions, and the instruments' psychometric properties. We argue that each support admits an alternative explanation grounded in the social rules governing well-being talk, and that no thin, purely descriptive construct lies beneath the everyday expressions these measures rely on. The third argument holds that the measures are useful regardless of these limitations. Here we distinguish an agnostic stance, for which scores are a means of predicting other outcomes, from a score-as-target stance, for which raising the scores is itself the end. The agnostic stance declines constitutive questions successfully, at the price of abandoning well-being as the central object of research; the score-as-target stance cannot, since it must still say why this quantity rather than another deserves to be raised. From these analyses, we propose two practices to bridge critical and applied approaches: making constitutive commitments explicit, and exercising greater rigour in the use of words and expressions.

Keywords: measurement; well-being; evaluative subjective well-being; life satisfaction; construct validity; thick concept

1. Introduction

Assessment instruments and the measurements derived from their use inevitably tolerate a certain degree of imprecision. For instance, there is little discussion about discrepancies in time measurement across different clocks. This lack of concern is understandable: for the intended use of clocks, such variations are usually irrelevant, just a few seconds of difference per week or month among distinct instruments.

We can apply the same logic to assessments of well-being: measurement imprecision must be weighed against the utility of the results for their intended purpose. Despite their imperfections,

Luciano E. Sewaybricker
São Camilo University Center
luciano.sewaybricker@prof.saocamilo-sp.br

Copyright belongs to the author(s)
www.internationaljournalofwellbeing.org

how useful are these measurements? This question should guide researchers' efforts, whether theoretical or empirical.

Large-scale applications of these instruments are now standard across psychology, economics, and related fields (O'Connor, 2023; Frijters et al., 2024; Helliwell et al., 2026). Many researchers argue that the field and its measurement tools have matured sufficiently to yield results that are highly valuable for both individual and collective decision-making, ranging from choosing activities and managing time to donating money, as well as evaluating public policies and government performance (Adler & Seligman, 2016). Thanks to this advancement, we are said to be experiencing a "happiness revolution", as Easterlin and O'Connor (2025) put it.

Nevertheless, the enthusiasm for these instruments has been accompanied by criticisms both of their theoretical consistency (e.g., Alexandrova & Haybron, 2016; Alexandrova, 2017) and of the difficulties involved in translating such a complex object into cardinal evaluation scales, particularly regarding heterogeneity in scale use and problems of linearity (e.g., Bond & Lang, 2019; Fabian et al., 2026). These two bodies of work, the applied and the critical, largely move in parallel. Without a clear map of what the conceptual objections target, defences of the instruments and critiques of them can pass one another by. A defence may address a weaker version of the objection than the one being pressed, while an objection may attribute to practitioners a disregard for theory that their own work does not display.

This article aims to examine, in greater detail, the relationship between these two bodies of work, bringing advocates of large-scale measurement "into conversation" with those calling for deeper conceptual work (e.g., Flanagan et al., 2023; Alexandrova, 2017; Fabian et al., 2024; Alexandrova & Haybron, 2016). The gain we seek is a practical one: specifying precisely where the disagreement lies, so that applied researchers can see which of their commitments are genuinely at stake and critics can direct their objections at positions actually held.

Several of the individual critiques we engage with are well established in the literature: concerns about construct validity (Alexandrova & Haybron, 2016), about the linguistic complexity of well-being and happiness (Cromby, 2011), about the possibility of evaluating life as a whole (White, 2006; Kahneman & Riis, 2005), and about scale use and linearity (Bond & Lang, 2019; Fabian, 2022a; Fabian et al., 2026). We do not claim novelty for these critiques taken individually, nor do we propose a new framework or theory. Our contribution here is analytical. The constitutive/determinative distinction is standardly applied to well-being itself, and, to our knowledge, has not been applied a second time to the construct a given instrument assesses, treated as a distinguishable object of constitutive inquiry in its own right. This recursive application is what allows us to organise three recurring supporting arguments for evaluative swb measures, interpret the interests and motivations underlying each, and show where each argument's relation to a constitutive question breaks down.

1.1 The object of analysis: Evaluative subjective well-being measures

This article focuses on evaluative swb measures: instruments composed, usually, of a single question that asks respondents to broadly and subjectively assess their lives, such as the Cantril Self-Anchoring Striving Scale (Cantril Ladder), used in the Gallup World Poll, and general life satisfaction questions, as found in the European Values Study (EVS). These instruments are central to national and international monitoring initiatives and have been used to guide public policy (see UK HM Treasury, 2022).

These measures draw on two distinct lineages. The first is the broad social survey tradition, in which Cantril's ladder was developed as a self-anchoring device for comparing populations. The second is the research programme in psychology that owes considerably to Diener (1984), in

which subjective well-being is treated as comprising two distinct components: an evaluative (or cognitive) component, expressed in global judgements about one's life, and affective components, expressed in the frequency and intensity of positive and negative feelings (Diener et al., 1999; Kahneman & Riis, 2005). Within that programme, cognitive evaluation is a central piece of subjective well-being, but only a piece. The instruments we examine here are therefore somewhat distanced from Diener's proposal: for pragmatic reasons as much as theoretical ones, they invest in the cognitive evaluation alone. This restriction is often defended on the grounds that life evaluations encompass affects and better "tell us about the relative importance of different life circumstances" (Helliwell et al., 2012, p. 22).

Throughout the article, we use "evaluative swb measures" for these instruments, and not as a reference to Diener's (1984) subjective well-being proposal. On terminology more broadly, our exposition defaults to "well-being"; where a cited author's term is "happiness" specifically, or where the historical or linguistic distinction between the two is itself at issue, we preserve it. The definitions of each, and their relationship, are complex and have led to substantial divergence within the field (Lomas et al., 2024; Author, 2022): we treat neither as equivalent to the other, and adopt no particular theory of either. This definitional plurality is, in fact, part of the subject matter of this article, and we will revisit it later.

1.1 Two kinds of questions and a thick concept

To organise the debate, we will rely on two conceptual tools developed within the philosophy of the science of well-being. The first is the distinction between constitutive and determinative questions. Constitutive questions concern what something is. Determinative questions concern what affects or changes it: which factors are associated with higher or lower reported evaluations, without themselves being part of what well-being is. This distinction has been drawn, in closely related forms, by Poller (2026) and by Fabian (2022b), who separates the constitutive aspects of well-being from its determinants or indicators. A large share of empirical research in the field operates at the determinative level, and legitimately so. Our interest is in what happens at the constitutive level, and in particular in the relationship between two constitutive questions that are not always distinguished: the question of what well-being itself is, and the question of what the construct assessed by a given instrument is. As we shall see, much of the debate between critics and proponents of evaluative swb measures turns on whether these two questions can be answered (or declined) independently of one another.

The second tool is the characterisation of well-being as a thick concept (Alexandrova, 2017). Thick concepts, such as cruelty or generosity, fuse descriptive and evaluative content: to describe an act as cruel is at once to characterise and to evaluate it. Alexandrova (2017) argues that well-being behaves this way: any attempt to specify descriptively what well-being is already involves taking a position on what makes a life go well. If this is correct, it has a direct consequence for measurement, and it is a central thesis of this article: for evaluative swb measures, the constitutive question about well-being cannot be separated from the constitutive question about the measured construct. Thickness alone does not explain this collapse; a further consideration about where the evaluative judgement is made will be needed to close the gap. Why this is so, and why the strategy of theorising a thin descriptive construct while bracketing the value-laden question is unavailable in the case of well-being, is the goal of our analysis of the three arguments.

2. First argument: The non-definition of the object is desirable

The first argument we will explore is commonly referred to as the "non-paternalism" of evaluative swb measures. This argument highlights the desirability of leaving the object of evaluation deliberately broad. As O'Connor (2023, p. 3) explains, "In this way, respondents can evaluate what is relevant in their life, implicitly assign importance to each aspect, and convey one summary value." In cases where the subject of inquiry is widely debated and lacks consensus, this non-paternalistic approach becomes particularly appealing (Sumner, 1996). To prevent the ultimate object of evaluation from reflecting what well-being means to the researcher, who may have developed a particular theory or instrument, rather than to the respondent, the evaluation parameters are intentionally left vague. In this sense, the instrument's response is taken as more meaningful in relation to the respondent's own life. As Diener et al. (1998, p. 35) made the case:

> With the increasing democratization and equality in the world, it is no longer defensible to rely only on kings and mandarins to prescribe the desirable life. (...) Thus, we allow people to decide whether their lives are satisfying based on their individual values, goals, and life circumstances.

This strategy of non-definition also supports the broad applicability of a given tool. Assuming that each of the many definitions of, say, happiness is more relevant to some individuals than to others, a researcher who commits to only one such definition must continuously justify that choice over others and accept that the broader the instrument's application, particularly across cultural boundaries, the less likely the chosen definition is to resonate with respondents (see Sewaybricker & Massola, 2025; Joshanloo, 2025; Sollis et al., 2022).

To avoid these complexities and incorporate such "non-paternalism", evaluative swb measures tend to use broad phrasing, relying on commonly used and assumed-to-be-universal expressions: the Cantril Ladder's "best and worst possible life", the EVS's "satisfied with your life as a whole". These refer to seemingly ordinary ideas, presumably familiar to nearly all participants, and are rarely treated as concepts in need of further definition, unlike more historically loaded terms such as "happiness" or "well-being". The strategy also relies on open-ended instructions (e.g., "all things considered"), allowing participants to weigh whatever they find most relevant: cognitive, emotional, individual, collective, historical, or present-day aspects (Mahoney, 2023, p. 12).

As we can see, the strategy of not defining the object of evaluation greatly enhances the usefulness of these instruments: they are easier for participants to understand, avoid endless theoretical justifications and cultural divergences, and are therefore more practical for large-scale applications. In the vocabulary introduced before, non-paternalism is the deliberate declining of a constitutive question. We now ask which constitutive question is being declined, and whether declining it is as innocuous as it appears.

2.1 A tacit concept is unavoidable

Even if the object of evaluation is not explicitly named in the wording of the instrument, we must assume that some conceptual understanding of it, however vague or tacit, underpins the instrument's design and the entire investigative strategy (Tal, 2020). This tacit concept must, in turn, be the basis for a construct: without one, it would be impossible to justify the relevance of an instrument ("Why is it important to measure X?") or the rationale behind its specific structure ("Why were certain formats, phrasings, or attitudinal items chosen?"). Thus, even if respondents are unaware that they are evaluating, for instance, their well-being (since this is not named in

either the items or the instrument title), the scientific discussion of the results should return to this underlying concept.

However, in the literature on evaluative swb measures, aside from brief definitions or nominal references, it is not easy to find well-developed articulations of the concept being assessed. One example is Loewenstein's (2009) assertion that subjective well-being is "that which makes life worthwhile". Another statement is Diener's (2020) that "Subjective well-being is thinking and feeling that your life is going well, not badly". To some extent, non-paternalism appears to extend to the researchers themselves (see Alexandrova, 2018; Flanagan et al., 2023).

This lack of conceptual clarity does not seem to pose a problem for many researchers, who indicate they are more focused on applying instruments and analysing the data produced, likely assuming that the concept is already sufficiently understood (see Adler & Seligman, 2016; Diener et al., 1999). For example, in 1984, Ed Diener, even as he was compiling a fragmented research field, already emphasised the importance of measurement (Diener, 1984, p. 543). This was a perfectly reasonable stance, though somewhat ironic given the lack of agreement on the field's core concept (see Linton et al., 2016; Margolis et al., 2021).

The emphasis on measurement, by itself, is not problematic. What is potentially problematic is the transition from one kind of unclarity to another: if researchers cannot agree on what well-being conceptually is, it becomes difficult to know how well a construct represents it, and whether items succeed in eliciting its assessment.

2.2 When can the concept-level question be omitted?

The strongest version of the non-paternalist argument can be stated as a question about levels. Declining the constitutive question at the level of the concept (refusing to say what well-being is) need not entail declining it at the level of the construct. A research tradition may leave the concept deliberately open, out of respect for the diversity of values (Sumner, 1996), while theorising rigorously about the psychological state its instruments capture, specifying that construct descriptively and submitting it to test. If that division of labour is reasonable, non-paternalism costs very little: rigour is secured at the level where measurement actually operates, and critiques aimed at the absence of a theory of well-being miss their target, since the target never claimed to have one.

Intelligence research is a standard illustration that the division can hold. For much of its development, that field offered no settled definition of its object (Sternberg & Detterman, 1986; Neisser et al., 1996). Yet at the construct level the field was and still is prolific. Factor-analytic work produced well-understood and replicable structures (Carroll, 1993), and the tests' internal consistency and predictive validity, for educational attainment, occupational performance, health, and longevity, are well documented in psychology (Schmidt & Hunter, 1998; Deary, 2012).

It is worth asking what licenses the omission in that case. Some conception of intelligence is of course at work when a test is designed, in the selection of tasks and formats; the point is that it need not be settled, shared, or even stated for the results to be interpreted. The operationalisation is answerable to psychometric criteria, not to consensus about what intelligence is. It is not answerable to the respondent either. What a participant believes intelligence to be makes no difference to the datum. Since no interpretive step requires knowing what the respondent thought the test was about, the researcher may keep the entire presentation of results at the descriptive level without loss.

The analogy is in any case less secure than it appears: "intelligence" is arguably as thick as well-being, and the history of testing offers well-known examples of value judgements entering under a descriptive guise (Gould, 1981). But the point holds nonetheless.

The intelligence case, however, cuts against the non-paternalism defence rather than for it. The licence just described is the stance we examine under the third argument as agnostic: the researcher declines constitutive questions and secures the instrument's standing entirely through prediction and correlation, treating the score as a means to an end. That position is coherent, but it is not the position of the non-paternalist. The agnostic needs only that the score predict; the non-paternalist needs it to express each respondent's own evaluation of their life, since that is what the deference to respondents' values amounts to. What they decline is the obligation to say what well-being is, not the claim that the score captures it.

The difference from the intelligence case lies in where the evaluation sits within the measurement act: in the well-being case, the respondent's own conception of a good life, which is thick, is not incidental to the datum, it generates it. This has a consequence the non-paternalism argument cannot absorb. If the content of the number was fixed by the respondent's constitutive judgement, the researcher cannot interpret it without some account of what was being assessed (the concept-level question becomes a precondition of using the data at all). Non-paternalism thus turns out to be self-undermining: the strategy adopted to avoid taking a position on what well-being is delegates that very question to respondents, and in doing so makes engagement with it unavoidable for anyone who wants to know what the resulting scores mean.

Still, a defender might resist by rejecting the premise, holding that respondents merely report an attitude or a felt state, so that their conception of well-being is no more an input than in the intelligence case. We address this reply under the second argument, where the linguistic and cognitive-interview material is examined.

Two clarifications are important regarding our claim. First, we do not conclude that rigorous measurement of well-being is impossible. Alexandrova (2017, 2018) argues, and we agree, that claims fusing descriptive and evaluative content can be handled scientifically, provided the value presuppositions at work are made explicit, examined where they are contested, and settled with the participation of those the measures concern rather than by the researcher's decision alone. What blocks the attempt to secure rigour by bracketing the concept, in the case of evaluative swb measures, is not thickness alone but where the thick judgement sits: because it is the respondent's judgement that generates the datum, no thin construct can be carved off to stand in its place. Second, our claim is not that evaluative swb measures are invalid measures of well-being. Construct validity requires a construct specified richly enough to generate testable hypotheses (McClimans, 2013); that specification is what thickness makes unavailable here, so the question of whether these instruments measure what they purport to measure cannot yet be settled.

2.3 The "ultimate goal" and "satisfied with life" characterisations

Two characterisations of the object are common in the literature and illustrate the difficulty just described. The first presents it as the ultimate goal of individuals:

> We mean the thing that ultimately matters: your inner subjective state – the quality of your life as you experience it, how happy you are. (...) Well-being, this book argues, is the overarching good and other goods (like health, family and friends, income and so on) are good because of how they contribute to our well-being. (Layard & De Neve, 2023, p. 1)

This definition, which draws on intellectual traditions as diverse as Aristotelian ethics and utilitarian philosophy, calls for a depth of inquiry rarely pursued. To state that well-being is the ultimate end (that to which all other things aim or contribute) easily becomes vacuous or tautological (Duncan, 2019, p. 86). Beyond the assertion that well-being is "good" by definition,

little is said about the object itself, and the claim risks circularity: well-being is the ultimate end of human life; the ultimate end is called well-being. The contrast with the traditions invoked is instructive. Aristotle (1934) identified the ultimate goal with the practice of virtue and elaborated extensively on what virtues are, how they are cultivated, and how they connect with a particular account of human beings; Bentham (1823) proceeded similarly from pleasure and pain. Both committed to such an account, an answer to the constitutive question, before stating what the "ultimate goal" is. Simply naming well-being as the ultimate goal contributes little to a meaningful understanding of the concept.

The second characterisation presents the object in the form of "life satisfaction", "how well he likes the life he leads" (Veenhoven, 1984, p. 22), or "a person's perception that life is going well" (Diener, 2020). Given the general familiarity with expressions like "satisfaction" or "doing well", this position might seem to avoid the need for theoretical clarification. After all, we talk to one another using such terms, saying we are "satisfied with something", that we are "doing well", or that life "is good". This linguistic familiarity will be analysed later.

In this context, Veenhoven's (1984, 2017) reasoning in favour of adopting this perspective is illustrative of how the field arrives at a supposedly thin object. Veenhoven's goal was to arrive at a workable concept amid the "discouraging confusion of tongues" regarding the object of interest, which he named happiness (1984, p. 4). To this end, he recognised the need to make a motivated choice for a narrower version of it; otherwise, his research "would probably get no further than repeating the old and unsolvable problems" of the happiness debate (Veenhoven, 1984, pp. 6-18). He concluded that among the various understandings of happiness available, the one that could best be operationalised was "the degree to which an individual judges the overall quality of his life-as-a-whole favorably" (1984, p. 22). It is not that Veenhoven declined the constitutive question; rather, he settled it by the criterion of measurability, collapsing the concept into the construct.

It is commendable that Veenhoven explicitly stated his interest in an operationalizable definition and acknowledged the inevitable cost of narrowing the concept. This seems to be the aim of many researchers in the field, though few are so explicit: whatever the theoretical cost, a measurable object should be achieved. Note, however, that even if this choice gives us a clearer operationalisation of well-being, three problems remain: at the level of the concept, whether the fundamental understanding of the object is adequate, given that there are divergences about what constitutes "happiness"; at the transition from concept to construct, whether the result of such collapse (as it considers only one aspect of the concept) can be regarded as a good representation of it; and at the transition from construct to items, whether instruments developed from this narrowing are capable of eliciting the expression of the object (the supposed latent content). In the end, the criterion of "usefulness" may diverge from that of "theoretical rigour".

It is worth noting that these three problems are available for examination precisely because Veenhoven made his choice explicit. Explicitness of this kind does not dissolve the difficulties; it makes them visible and debatable, which is more than can be said for the same move performed tacitly.

2.4 Researcher-side common sense in instrument validation

If the concept is left underspecified and theory is scarce, how are instruments evaluated in practice? Here the idea of common sense plays a central role (Haybron, 2024), especially given the strong desire to evaluate an object with democratic contours, as opposed to one defined and understood only by "specialists". There are many aspects of human life where there is a broad consensus on whether they are good or bad, requiring little or no justification of their value. For

example, it is hard to imagine a scenario in which we would need to argue why being healthy is good in someone's life. Accordingly, health is widely treated as essential for well-being. The same could be said for having a place to live, a stable income, freedom from food insecurity, or having people one can rely on. These consensuses are used by researchers both to characterise the concept and to assess the adequacy of instruments. Easterlin and O'Connor (2025, p. 5) point this out:

> It turns out that the happiness of most people everywhere – in rich and poor countries, democratic and authoritarian – depends mainly on the same three broad personal concerns – first of all, their economic situation, and then, family life and health.

Common sense, in this role, points toward relevant elements of the object of interest during instrument development. Researchers must define which observable variables are assumed to correlate with the construct, and understanding common sense is crucial at this stage to evaluate whether, in the instrument's early uses, the construct behaves as expected (Alexandrova, 2017). For example, if a person's health worsens, their well-being is expected to decline (and if it doesn't, the instrument may need to be revised). The development of the Satisfaction with Life Scale (SWLS), as summarised in Diener et al. (1999, pp. 277-278), illustrates this well. The instrument was validated based on its correlation with (among other things) absence of depression and fewer suicide attempts, two elements with little controversy regarding their relationship to well-being.

Notice what kind of consensus is being recruited here. The previous quotation states what well-being depends on; the SWLS validation rests on what tends to accompany it. Both are determinative claims. What common sense supplies, in other words, is agreement about determinants and indicators, and this is then used to certify an instrument whose constitutive content was never articulated. The distinction Fabian (2022b) draws between the constitutive aspects of well-being and its determinants or indicators is pertinent here, because common sense does not respect it. Consider health: is it part of what well-being consists of, or a condition that promotes it? Common sense registers only that health is good and that it goes with living well; it is silent on which role health plays. Researchers can therefore recruit health as validation evidence without ever settling the constitutive question at the concept level.

Still, the issue is not only whether a particular common-sense notion is relevant to well-being. It also concerns recognising that such notions do not apply to everyone alike, and that they are generic: they must be translated into the specifics of someone's life during the reflective process of evaluating well-being (Alexandrova, 2017). "Being healthy" may mean retaining the physical capacity to maintain one's hygiene, or being able to run a half-marathon, and it means something different for a child than for an old person. Without this level of detail, the consensus that health matters to well-being is too generic to guide either the respondent's judgement or the researcher's assessment of the instrument.

Even with these points, it is entirely understandable that common sense is taken into account during the initial development of instruments intended to measure a complex object. Common sense and measurement attempts can help better understand both concept and construct through an iterative process (Rodriguez Duque et al., 2024). In the case of SWLS validation, however, it would be important for common-sense indicators such as the absence of depression to be explicitly derived from a richer account of what well-being consists of, or to serve to enrich that account. Otherwise we know that suicide incidence should be inversely related to SWLS scores without understanding why. Ahmed (2023) points to a further risk in treating common sense uncritically: left unarticulated, it conceals the value judgments at play, making their origins and effects harder to examine. Why these forms of common sense rather than others? Is it common

sense for whom? Defending the non-paternalism of evaluative swb measures on common-sense grounds does not imply an absence of value judgements; it merely avoids discussing them (see Rodriguez Duque et al., 2024, p. 3). In the terms of our framework, common sense here functions as an implicit answer to a constitutive question at the level of the concept, an answer all the more influential for being unexamined.

Before moving on, a boundary should be marked. The role of common sense discussed in this section is researcher-side: it concerns the criteria researchers use to develop and validate instruments in the absence of explicit theory. A distinct, respondent-side appeal to everyday familiarity, the claim that respondents' ordinary use of words like "satisfied" shows they know what they are evaluating, will be addressed in the next argument.

3. Second argument: The lack of a clear definition is not disqualifying

The second argument can be understood as a response to the difficulties raised above about the underspecification of the concept. Whereas the first argument holds that declining the concept-level question is a virtue, the second concedes that it may be a gap but holds it harmless: the absence of a precise definition of the concept is not a problem, or at least not a disqualifying one, because other properties of the instruments and the data they produce carry the weight. In this line of reasoning, the psychometric properties of the instruments are considered strong enough, as illustrated by Diener et al. (1999):

> These measures do possess adequate psychometric properties, exhibiting good internal consistency, moderate stability, and appropriate sensitivity to changing life circumstances. Furthermore, global reports show a moderate level of convergence with daily mood reports, informant reports, spouse reports, and recall for positive versus negative life events. People who score high on global life satisfaction are less likely to attempt suicide and to become depressed in the future. (Diener et al., 1999, pp. 277-278)

This argument rests on three supporting elements, each addressing the underspecification at a different level: statistical washout, tacit respondent familiarity, and construct-level performance.

The first concerns large samples (Easterlin & O'Connor, 2025): variation in what respondents consider would be mitigated by sample size, yielding a reliable average. This logic underpins much cross-cultural research, justifying comparisons between vastly different countries, as in the annual ranking published in the World Happiness Report (Helliwell et al., 2026). We might not know the concept's meaning for any one person, but we know it well enough for people in general.

The second support is the respondent-side appeal deferred above: people commonly make everyday-life evaluations similar to those required by evaluative swb measures. People ask themselves and ask each other whether life is going well, whether they are satisfied with something, or even whether they are happy (Kahneman & Riis, 2005). If this is a common occurrence, it would suggest that people are sufficiently familiar with the concept to assess their relationship with it in an instrument.

Finally, the third supporting element refers, as highlighted in the quote from Diener et al. (1999), to the psychometric properties of the instruments themselves. High response rates (Veenhoven, 2017), strong test-retest scores, and the alignment between self-assessments and the assessments of others who know the respondent (such as partners and friends; Schneider & Schimmack, 2009) are commonly cited as evidence that respondents know what they are evaluating (Frijters et al., 2024, p. 4).

All three of these supports are consistent in their own right and widely endorsed within the field. However, each admits an alternative explanation that deserves closer attention. We examine them in turn.

3.1 What large samples can and cannot fix

We should begin by stating plainly what the appeal to large samples can establish. Averaging over many respondents mitigates random error: idiosyncratic variation that is unsystematic will tend to cancel out, yielding stable means. This is a matter of reliability. But the question at issue here is one of validity: whether respondents' answers express the same construct, and whether that construct tracks well-being. Large samples are silent on this question. If different respondents are answering about systematically different things, no increase in sample size converts their responses into measurements of a common object; it merely produces a precise average of heterogeneous quantities. The support, in other words, answers a different question from the one the critic is asking. An instrument can yield highly stable numbers while it remains unsettled what those numbers are stable measurements of.

With this distinction in place, two further counterpoints can be made: "ignoring outliers" and "simplification without meaning". Panasiuk et al. (2025) and Barrington-Leigh (2025) offer interesting examples of the former. By analysing databases from studies that used single-question well-being assessments, both found that the distributions of responses differ significantly from a normal curve. More specifically, Barrington-Leigh (2025) noted that focal-value rounding is more pronounced in some regions than in others. In European and Anglophone countries, where most theories and measures are developed, focal-value rounding and variation in response distributions across instruments are considerably smaller than in other regions. This pattern may be explained by differences in familiarity with the semantic field related to well-being (which we will address shortly) and in familiarity with numerical evaluations of subjective experiences (such as school grading or hospital pain scales), a practice that is uncommon for many communities, such as some Indigenous peoples (see Sewaybricker & Massola, 2025; Da Silva Sinha et al., 2017). These studies reinforce a point that is often overlooked yet obvious: discussing results based on averages may create a false impression of consensus.

For specific purposes, such as some public policy evaluations, working with averages and setting deviations aside may be justified by practical utility, and finds support in item-response theory, which treats deviation that is random rather than systematic as "noise". However, with a contentious object of inquiry we lack a basis for discerning latent content from noise. Careful attention to response heterogeneity and outliers is then especially relevant for advancing theory, particularly with diverse populations (Truijens et al., 2023).

The second counterpoint, "simplification without meaning", concerns a potential consequence of treating the object as simple. The simpler the object is presumed to be, the easier it becomes to develop a "rigorous" instrument for its evaluation, and this is almost a necessity if an instrument is to be applied to a large population. Everything participants might wish to express is narrowed to a few numerical or verbal gradations, and it is assumed that people navigate the scale in similar ways, so that the "5" one person chooses is comparable to another's (Fabian et al., 2026). This simplification is not accidental; it follows a long process of psychometric testing (Diener et al., 1985; Alexandrova & Haybron, 2016). Still, its degree is striking for a historically complex subject (McMahon, 2006): both the EVS question and the Cantril Ladder use a single attitudinal item, usually answered in under ten seconds (Liu & Netzer, 2023; Samuelsson et al., 2023). Such measures appear to elicit a shortened, heuristic form of reflection that may bear little or only superficial relation to the object of interest, even when the resulting data display

good psychometric properties. Noticing oversimplification is itself difficult, moreover, since the concept is nebulous. Are people evaluating well-being, a simplified version of it, or another concept altogether? Answering requires greater attention to theoretical foundations.

3.2 Linguistic familiarity involves rule-following

The second support holds that the everyday currency of the words used in evaluative swb measures shows people are familiar both with the object addressed and with the reflective process the instruments require. Regarding familiarity with the object, Cromby (2011) offers an alternative explanation grounded in linguistics. Cromby argues that regular use of a word should not be interpreted as evidence that individuals possess sufficient and stable knowledge of the object to which the word refers. Instead, regular use of a word primarily indicates that the person knows the social rules governing its use, independent of any stable or deep understanding of the object itself. For example, when we wish someone a "happy birthday", we do not need to rely on a specific definition of happiness. Nor does the person receiving the wish need to have a formed idea of what happiness is or to understand the particular perspective of the one offering the wish. We know that "happiness" is something positive, desirable, and appropriate to mention to someone on their birthday, and that suffices.

The birthday case shows that an exchange organised around "happiness" can succeed by mastering social rules, without either party evaluating a life. This is enough to defeat the inference from everyday fluency to competence at life-evaluation specifically. What respondents are practised in may instead be solving the social problem of using expressions like "happiness" appropriately, so that their reflective process is guided by thoughts like: "Given the social context I am in, would it be appropriate to say I am happy, and with what intensity?" or "How will others react to my answer?"

This is a form of assessment, but different in kind from what an evaluative swb measure asks for: judging what a situation calls for is bounded by a context that fixes, more or less, what counts as an adequate answer, while an evaluative swb measure offers no such context to bound the judgement it asks for. Knowing how to deploy "happy" appropriately in social exchanges is one competence; performing a global appraisal of one's life and mapping its result onto a numerical scale is another.

This suggests a different reading of what happens when someone answers an evaluative swb item. It is not that respondents fail to consult their own sense of how their life is going and consult a social script instead. It is that both are available, both are practised, and the instrument gives no way to tell them apart. Part of what any respondent knows about "satisfaction" and "a good life" they know partly through the social-rule-following that everyday usage trains: what counts as a satisfactory life, what one is expected to say about one's own, which reports invite concern and which close a conversation. When they mark a 7 rather than an 8, they may be reporting a judgement about their life, a judgement about what it is appropriate to claim, or some combination of the two. How much each contributes is not something the instrument records, and not something the researcher can recover from the number.

The difficulty this creates is compounded by a further one: it is unclear that the appraisal the instruments request is one anyone can perform. Kahneman (Kahneman & Riis, 2005) investigated the effects of memory on self-assessment, illustrating how volatile retrospective evaluations of life can be. In the field of philosophy, White (2006) argued that treating "life as a whole" as a unit of analysis is logically impossible, since it would require considering everything about life at once, and new elements and perspectives will always emerge. If the requested act is unstable or incoherent, and what respondents supply in its place is partly a judgement of social

appropriateness, the resulting number is doubly difficult to interpret: unstable in what it targets, and unclear in what it reports.

The cognitive-interview evidence reported by Fabian et al. (2026) is consistent with this picture. Respondents differ in the considerations they weigh, the heuristics they employ, and their interpretation of what the question asks. This heterogeneity is what one would expect if answering involved drawing, in varying proportions, on more than one kind of assessment: different respondents settle the ambiguity differently, and nothing in the instrument constrains how they settle it. What the evidence does not support is the assumption on which the second support rests, namely that everyday familiarity guarantees a common act of appraisal underlying the responses.

3.3 Strong psychometrics, blurred concept

In light of the considerations above, let us revisit the third supporting element: the claim that the instruments' psychometric properties indicate that participants know what they are evaluating. The numbers for these psychometric qualities are indeed significant and represent potentially strong evidence for the instruments. However, once again, there are alternative hypotheses that explain why these numbers are so significant without requiring respondents to have (explicit or tacit, conscious or unconscious) knowledge of the object of interest.

One commonly presented point is the high response rate. According to Veenhoven (2017), fewer than 1% of participants fail to respond to the questions or select alternatives such as "don't know". It is assumed that if the vast majority of people complete the tests adequately, it is because they understand what is being asked of them. However, for such a highly complex object, which is hard for even specialists to grasp, the high response rate may reflect people's familiarity with its expressions rather than with the concept (Cromby, 2011). The diversity of definitions of well-being (see McMahon, 2006; Haybron, 2007; Sewaybricker & Massola, 2025) and the general difficulty people have in presenting a coherent and stable reasoning for the assessment of well-being seem to support this hypothesis (Fabian et al., 2026).

In any case, a defender of evaluative swb measures might insist that the high response rate stems from tacit knowledge of the object: people filling out the instruments would know what the object is, even if they cannot explain it or justify the score. This could also explain why response time for evaluative swb measures is usually less than ten seconds. However, this is a weak argument for supporting the rigour of an instrument, as we would have no way to verify it. If researchers do not clearly know what the object is, and participants cannot explain it, how can we ensure that the knowledge elicited by the instrument concerns well-being?

The same problem extends to test–retest scores and to the alignment between self-assessments and assessments by others who know the respondent. Given the complexity of the object, and perhaps the impossibility of evaluating life as a whole, the stability of these numbers specifically might be better explained by common sense and the rules governing the use of expressions than by successful engagement with latent content. Whatever mixture of genuine evaluation and social judgement generates a given response, it is the common-sense and rule-governed component that is more stable and more available in everyday exchange, and that would better explain the cohesion found among individuals and within individuals over time. We are more familiar with calling people and situations "happy" than with defining happiness. Note that this is the respondent-side counterpart of the common sense discussed under the first argument: in both cases, what stabilises the practice is a shared stock of consensual associations and rules of use, rather than any shared concept.

3.4 Closing the loop: The thin residue does not exist

The evidence assembled under this second argument helps answer a question we partially addressed earlier: whether "life satisfaction" might be a thin descriptive construct, carved off in order to escape the contested constitutive questions. If there were such a construct beneath the everyday expressions, two things would be expected. First, that respondents' answers could in principle be resolved into an act of evaluating that construct, however imperfectly performed; instead, what the evidence shows is a mixture (part evaluation of a life, part judgement of appropriateness) in a ratio the instrument cannot record and the researcher cannot recover (Cromby, 2011; Fabian et al., 2026). Second, that the appraisal of life as a whole was a stable, performable operation, since a thin construct must still be something respondents can be asked to assess; instead, there are serious reasons to think it volatile at best and incoherent at worst (Kahneman & Riis, 2005; White, 2006).

Taken together, these observations suggest that the everyday expressions on which evaluative swb measures rely refer to a thick concept, and not to a stable thin construct that could be theorised descriptively while the value-laden questions are bracketed. The terms are norm-governed and semantically unsettled, the requested evaluative act is of doubtful coherence, and even where genuine evaluation occurs, it arrives inseparable from judgements the thin-construct hypothesis has no place for. What the instruments record may therefore be less a thin construct awaiting better specification than responses generated by conceptions the measures never asked respondents to state.

4. Third argument: Despite their limitations, evaluative swb measures are useful

The third and final argument for evaluative swb measures concerns their usefulness. That is, the limitations pointed out above are taken to be relevant only insofar as they affect the usefulness of these assessments. In the terms of our framework, this argument proposes to operate entirely at the determinative level, setting constitutive questions aside. As we shall see, it comprises two stances, and the proposal succeeds for one of them but not the other.

The usefulness of evaluative swb measures is conceived mainly in terms of their predictive power and the strength of their correlations with variables of interest. The underlying thinking is roughly as follows: if instruments can predict behaviours and show strong correlations with relevant variables, then there is no need to delve into the details of what well-being actually is, or even to question whether the instruments genuinely measure it.

By focusing on usefulness, research efforts can be viewed as twofold. First, the treatment of data derived from different instruments to make their results comparable (e.g., Kaiser & Oswald, 2022; Kaiser & Lepinteur, 2025; Frijters et al., 2024). The second effort involves investigating the respondents' reporting function, which is a necessary step for adjusting instruments and research design to promote greater comparability both across participants and within the same participant over time (see Prati & Senik, 2025).

These efforts are not free from disagreement. Divergences emerge especially in the discussion of the reporting function, since it involves a complex "affective, cognitive, social, and linguistic process that maps subjective judgements to numerical response categories on a scale question" (Fabian et al., 2026). Bond and Lang (2019), Fabian (2022a), and Fabian et al. (2026) have pressed this line most incisively, emphasising how much the reporting function varies. The deeper difficulties, though, arise from the nature of the object itself (Flanagan et al., 2023; Alexandrova, 2017; Angner, 2013; Rodriguez Duque et al., 2024). Unlike a ruler, which measures length using consistent units, well-being (an abstract and highly contested object) is measured using integers, a property of an entirely different kind. A score of 5 may mean a great deal or very little to

different people, or even to the same individual at different times. Thus, score increases may not correspond to linear increments of the underlying concept: the jump from 5 to 6 may feel larger than from 8 to 9. This problem is not unique to the cardinal evaluation of well-being, but it is more acute in this case, given the object's complexity.

Note, however, where these critiques locate the difficulty. For Bond and Lang, Fabian, and others, heterogeneity in scale use follows from the nature of the object itself: if well-being is not a cardinal quantity, no adjustment to the reporting function will make it one. The difficulty is constitutive, and its diagnosis depends on a view of what well-being is. Research oriented by the third argument locates the same phenomena elsewhere, in the mapping between judgement and scale, and treats them as properties of the measurement process rather than of the thing measured. They become technical problems with technical solutions, and no view about the object is needed to address them. Some researchers assume that responses can be read directly as expressions of well-being (see Easterlin, 1974); others do not assume this, but treat variation in the reporting function as noise obscuring a signal they take the instruments to carry (see Prati & Senik, 2025), and set about identifying and reducing it. Strategies include adjusting for participants' response styles (Kaiser, 2022), using anchoring vignettes (King et al., 2004; Weiss & Roberts, 2018), and calibrating a respondent's current score against their own retrospective evaluation of an earlier one (Prati & Senik, 2025).

Regarding linearity, evidence suggests that the impact of the difficulties identified by some critics is far less significant than expected. People seem to use well-being scales in a cardinal manner (Layard & De Neve, 2023; Prati & Senik, 2022), especially in scales with more than three response items (Kaiser & Vendrik, 2020). As Kaiser and Lepinteur (2025, p. 1) wrote, "respondents use survey scales in ways that deviate from linearity, but only mildly so." Nonetheless, it is essential to reiterate the more critical stance: for critics not primarily focused on the usefulness of instruments, the evidence of "quasi-linearity" might be interpreted not as a validation of evaluative swb measures but as an indication that they capture something other than well-being, which these critics take to be a contested concept without cardinal structure (Angner, 2013). The question of what is actually being expressed in evaluative swb measures would therefore remain open.

4.1 Two stances towards usefulness: Agnostic and score-as-target

Various studies indicate that, whatever one concludes about construct validity, evaluative swb measures predict behaviour well. Kaiser and Oswald (2022) found that they anticipate "exit behaviour" or "get-me-out-of-here" actions, such as changing jobs, moving house, or ending a marriage; Ward et al. (2021) and Liberini et al. (2017) found that they predict voting patterns. Given this predictive power, it is unsurprising that the measures are used by individuals and organisations to guide decisions (for work, see De Neve & Ward, 2025; for public policy, see UK HM Treasury, 2022): companies may act to improve employee retention and politicians to adjust campaign strategies.

Within this usefulness-oriented research, however, two stances should be distinguished, and the distinction concerns what the score is for: whether it serves as a means of acting on something else, or whether moving the score is itself the target of action.

Let us call "agnostic" the stance in which, without engaging in conceptual debates, one supports the use of evaluative swb measures for decision-making because of what the scores predict. The decision-maker need not understand well-being, or accept researchers' definitions; it suffices that they trust the statistical analyses and see value in the predicted behaviours and correlations. Scores are used as a means to an end (e.g., employee retention or anticipating

unemployment) and the score is an instrument for forecasting it. In our framework's terms, the agnostic stance is a purely determinative research programme.

There is another stance that seeks to influence actions based on evaluative swb measures, but for which the scores themselves are the point. We will call this the score-as-target stance. Unlike the agnostic stance, it is interested in the actual scores of evaluative swb measures: the self-declared nominal level of well-being (Frijters et al., 2024; Plant et al., 2025). Specifically, it is concerned with how to raise these scores, as can be seen in the focus of specific institutions (e.g., Action for Happiness and the Happier Lives Institute) and in the work of those who argue that going "beyond GDP" requires placing the promotion of well-being at the centre of policy (Weijers & Jarden, 2013; Frey & Stutzer, 2017).

4.2 The agnostic stance: Well-being exits the stage

It is hard to find arguments to refute the agnostic stance's claims, as it bypasses constitutive questions altogether. A caveat, however, is worth noting: in this stance, well-being ceases to be the focus of research, which shifts to unemployment, turnover, or votes. The agnostic stance therefore requires no conceptual deepening and need not engage in multidisciplinary or theoretical debate, and studies interested in why the predicted behaviours occur operate under a very different logic. Cultural particularities that might bring certain groups closer to or further from the response patterns typically observed are not of direct concern to agnostics, except insofar as they affect predictive power or correlation strength.

The reporting function does not stand still. If the mapping from inner states to scale points drifts across cohorts, cultures, or survey waves, as the scale-norming evidence suggests it can (Fabian, 2022a), predictive models calibrated on one population or period will degrade on another. The agnostic can treat this as a technical problem of recalibration, and can do so on their own terms: a correction succeeds if predictive performance recovers, and nothing further need be asked. Whether the adjustment has distorted what the responses originally conveyed is not a question the stance is equipped to raise, since it assigns no value to that content. This is consistent with operating purely at the determinative level, but it shows how far the object has receded. The agnostic stance is internally consistent to the extent that it has quietly changed the subject.

4.3 The score-as-target stance and the magic wand

The score-as-target stance might appear able to decline constitutive questions as successfully as the agnostic one does. Indeed, the strongest defence available to it can be put as follows. Suppose someone possessed a magic wand: waving it over a person yields a single number, and nothing is known about how that number is generated or what it stands for. Suppose further that the number reliably tracks several things we already have independent reason to value: lower rates of depression and illness, longer lives, higher incomes. Would it not be reasonable to treat that number as worth raising, purely on the strength of what it predicts, and without any account of what it is? If so, then this stance needs no theory of well-being either: the value of raising the score is secured entirely by the score's correlational profile. On this defence, the score-as-target stance is just as constitutively abstinent as the agnostic one, only with a different practical orientation.

The defence is instructive, but it does not deliver the immunity it promises, for the wand-waver must still answer a question that the agnostic never faces: why intervene on this number rather than on any other quantity with a similar correlational profile? Many variables correlate with that same bundle; income itself, for instance, is measurable, movable by policy, and predictive of the rest. To single out the evaluative swb score as the target, rather than treating it

as one indicator among many, requires a reason, and the reasons actually given in this literature are revealing. The score is said to capture "the thing that ultimately matters" (Layard & De Neve, 2023, p. 1); it is said to register the quality of life "as you experience it". Every such reason is a constitutive commitment: a claim about what the number is, and about why that content, rather than its correlates, deserves to be moved. The magic wand relocates theory from the question "what does this instrument measure?" to "why is this quantity, rather than another, the proper end of action?". The answers to the relocated question are usually supplied by unexamined common sense about well-being; the same researcher-side common sense whose role in instrument validation we examined under the first argument, which now does its work at the level of policy justification instead.

Two pragmatic answers to the relocated question deserve mention, since they appear to avoid constitutive commitments and do not. The first appeals to incremental predictive power: the score should be targeted because it forecasts the bundle of valued outcomes better than any single member of the bundle does. But incremental prediction justifies using the score as an indicator, which is the agnostic stance; it does not justify treating the score as the end, which is what distinguishes the score-as-target stance. The second appeals to democratic warrant: the score should be targeted because it reflects what people themselves report caring about. Yet this assumes that responses express something respondents value, which is a claim about what the responses are, and thus a constitutive commitment by another route.

Moreover, this stance faces a difficulty the agnostic escapes, since no one is rewarded for a forecast. Once a measure becomes a target, pressures arise that degrade it as a measure - a pattern familiar from Campbell's (1979) analysis of social indicators and commonly summarised as Goodhart's law. Those rewarded for higher scores, such as a government claiming to have improved people's lives, acquire reasons to produce higher scores by whatever pathway is cheapest. And where the object is contested, accounts of what genuine improvement would consist in are available but unsettled (Duncan, 2010), so a claim to have raised well-being rather than merely the number cannot be adjudicated without first settling which account applies.

The score-as-target stance, therefore, cannot avoid grappling with the conceptual complexity of well-being. The possibility, highlighted by critics, that what these measures capture may not be well-being poses a serious challenge to the entire stance, in a way that it does not for agnostics. Among other things, it requires reflecting on whether "corrections" to participants' scores, such as those proposed by Prati and Senik (2025), might in fact distort whatever content the scores are taken to carry, on which the justification for intervening depends. Adjustments to the reporting function made for the sake of comparability and linearity are unproblematic for an agnostic, for whom the score is a forecasting device; for those who act in order to move the score because of what it is taken to mean, such adjustments create as many problems as they solve.

5. Discussion

Concepts and their assessment tools will never be perfect in the sense of definitively resolving all disagreement within a field. There is no "holy grail" in the assessment of well-being, and there never will be; when it comes to mental states, instruments are further from any such thing than in the natural sciences, where measuring an external, directly perceptible object facilitates consensus. That said, the particular complexities of well-being do not mean we cannot strive for greater rigour and develop useful constructs and instruments at the same time. What is essential is to acknowledge the limitations of the approaches adopted, so that results are not extrapolated beyond what they rigorously indicate. This is a particular concern where instruments developed in one linguistic and cultural setting are commonly used to rank populations for whom the

response format itself might be unfamiliar (see Barrington-Leigh, 2025; Da Silva Sinha et al., 2017; Ahmed, 2023). In what follows, we first consider whether these difficulties extend to affective measures, before proposing two practices to help bridge critical and applied approaches: making constitutive commitments explicit, and exercising greater rigour in the use of terms and expressions.

5.1 The scope of our critique: Do these problems extend to affective measures?

Although our analysis has focused on evaluative swb measures, it is fair to ask whether the concerns raised extend to measures of the affective components of subjective well-being, which the instruments we have examined set aside. Several of the difficulties discussed here are not specific to evaluative swb measures. The norm-governed character of emotion vocabulary, heterogeneity in scale use, and questions about the comparability and linearity of responses plausibly apply to affective self-reports as well. Other difficulties, however, are specific to the evaluative swb measures: the requirement to appraise "life as a whole", with its attendant logical and mnemonic problems (White, 2006; Kahneman & Riis, 2005), has no direct analogue in reports of recent or current feelings. Our claim is narrower: evaluative swb measures face the constitutive problem in its most acute form, since they request the most global judgement the field asks for, phrased in vocabulary that is thickly loaded and about an object that remains contested. A systematic comparison of how far the present analysis carries over to affective measurement would be a valuable extension of this work.

5.2 Recommendation 1: Making constitutive commitments explicit

One obstacle to achieving rigour in research is that most uses of evaluative swb measures are not agnostic in that they treat the score as informative about, or worth acting on for the sake of, how someone's life is going; and any such use incurs a constitutive commitment, whether or not it is acknowledged. And this commitment is rarely examined, probably because of the cost imposed by thickness: doing so requires taking a position on what makes a life go well, which most presentations of results avoid. Once a use claims or presupposes that the score represents well-being, that claim cannot simply be added to a purely predictive, agnostic use without cost, since the agnostic's silence about content and the non-agnostic's reliance on it cannot both describe the same score. Researchers and institutions sometimes move between them without registering the shift. WELLBY advocacy is a case in point: the target is stated without hesitation, while the underlying constitutive commitment (what raising that number is meant to represent) is not. Without clarity on these commitments, how is it possible to construct a meaningful critique? In such cases, the dialogue is likely to be fruitful only when limited to methods and results.

Our first recommendation, then, is that this explicitness (already standard at the level of research design and data analysis) become standard at the constitutive level as well, for anyone whose use of a score is not purely predictive. These assumptions may be embedded in the theoretical frameworks adopted, the instruments selected, how questions are framed, and what researchers suppose when testing an instrument's quality, as we saw in the role of researcher-side common sense in validation. Where such claims are made, the answers they presuppose should be available for inspection.

This recognition opens the door for theoretical-conceptual research to be presented as complementary rather than antagonistic to practice-oriented research, and it is continuous with what is proposed by Alexandrova (2017, 2018), for whom the value presuppositions of such uses should be unearthed and checked rather than denied. Our analyses above indicate where, in the case of evaluative swb measures, some of these presuppositions are likely to hide. One location

is the frameworks, instruments, and validation criteria just discussed; a second, distinct one is the respondent's own act of judgement, which non-paternalism defers to without requiring examination. One direct response to this second source is to involve respondents in specifying what is being measured in the first place and in its analysis, which is the aim of participatory and co-produced approaches (Sollis et al., 2022; Yap & Yu, 2016; Alexandrova & Fabian, 2022).

Explicitness about the researcher's own assumptions and participation by those whose lives are being evaluated are complementary, not competing, strategies. Together, though, they do not guarantee an adequate measure; a constitutive commitment can be made explicit and still be indefensible. What they address is a narrower aim: that researchers be able to say what constitutive commitment their project encodes, and that this commitment be examined.

5.3 Recommendation 2: Rigour in the use of terms and expressions

Our second recommendation pertains to a more rigorous use of terms and expressions. We acknowledge the difficulty of this recommendation, since word choice involves precision as well as a desire to be understood. Moreover, we are dealing with a particularly confusing semantic field in which word choice often requires consideration of translation, particularly into English. For example, due to the linguistic and conceptual complexities of the object, Easterlin and O'Connor (2025, p. 12) proposed a practical solution: "We refer to them both by the less cumbersome and more user-friendly term, 'happiness.'" However, we might ask: for whom should it be less cumbersome and more user-friendly? While such a choice may be convenient, it may hinder dialogues grounded in conceptual rigour.

In this sense, it seems important to consider words or expressions in their original contexts, respecting (where possible) the word used in the local language of the research, either by retaining the original term or noting the translation. For instance, while the English word *happiness* underwent a process of trivialisation over the 20th century (Oishi et al., 2013; Wierzbicka, 2004; Sewaybricker & Massola, 2023), this does not mean that words from other languages commonly translated as *happiness* have gone through the same process (as Wierzbicka notes in the case of the French *bonheur*). So, even when words are part of the same semantic field, they should not be treated as "the same". This also applies to terms used by historical authors (e.g., Aristotle and *eudaimonia*, Epicurus and *makarios*, Jeremy Bentham and *happiness*). Moreover, investigating the etymology and historical transformation of words can help ensure more precise usage. Such inquiries can help avoid inadvertently replicating moves such as Ed Diener's, who introduced *subjective well-being* rather than *happiness* to gain credibility in the scientific community, without theoretical justification (Disabato et al., 2025, p. 2; Lyubomirsky, 2008, p. 316).

Given that the object is complex to define, polysemic, and contested in the literature, it is still possible to strive for internal consistency within each research study. A straightforward way to ensure internal rigour is to state clearly how expressions will be used (e.g., whether happiness and well-being will be treated as synonyms). However, it is valuable to go further. It is common for papers to cite many other studies that do not treat the object of interest the same way or use the same terminology. Thus, it is essential to double-check how expressions are used and justify the chosen conceptual approximations. For instance, it would be careless to treat happiness and well-being as synonyms without acknowledging whether cited studies do the same.

Neither of these two recommendations is novel in principle. Their value lies in highlighting practices that, despite being widely endorsed, are still inconsistently applied, and in showing, through the analyses above, what is at stake when they are neglected.

6. Conclusion

The growing relevance of evaluative swb measures and the expanding body of research employing them reflect the value placed on assessing well-being. However, this is a particularly complex object, which has led many researchers to define it in terms of what can be measured (e.g., Veenhoven, 1984). In this case, the instrument ultimately determines the object. If well-being exceeds what is being measured, this is not considered a problem for many researchers.

Our aim in this article has been to interpret the interests and intentions at play in this practice. Of the positions examined, only the agnostic stance declines the constitutive question successfully, and it does so at a price: the object of research is no longer well-being but whatever the scores forecast. Every other position is, in the sense developed in our first recommendation, non-agnostic and each incurs a constitutive commitment it does not fully discharge. Non-paternalism declines the concept-level question, but in delegating it to respondents it makes engagement with that question a precondition for interpreting the resulting scores. The supports offered by the second argument (large samples, everyday familiarity, and confidence in psychometric performance) address reliability and construct-level behaviour rather than the underspecification of the concept, and each admits an alternative explanation that leaves that underspecification intact. And the score-as-target stance must still explain why this quantity rather than another deserves to be raised, which cannot be answered without constitutive commitments.

This is what we take Alexandrova and Haybron (2016, p. 1103, emphasis in original) to mean when they write that construct validation in well-being science is "*theory avoidant*, sacrificing valid theoretical knowledge for statistics for no good reason". Such a posture hinders broad discussions of scientific rigour by excluding the object's complexity from the "discussion table". This might make sense for those adopting an agnostic stance, who have given up well-being as their object, but not for others primarily interested in it. The practice recalls the anecdote of someone who searches for a lost item only where the light is shining (Angner, 2011): construct-level rigour is where the light is, and so the search continues there, even though the difficulty lies at the level of the concept. And without a clear picture of what is being sought, there is no way to tell whether what turns up under the light is the right thing at all, which is exactly the risk Ahmed (2007, p. 9) identified: advancing without clarity about the object tends to confirm the searcher's own expectations, finding well-being precisely where one had already assumed it would be.

It is essential for the development of the field, and for the development of the instruments themselves, that we place greater emphasis on the level of concept and theory; that we bridge the gap between measurement efforts and critical discussion. If we fail to do so, we risk continually simplifying the object of interest, reducing it to the most tangible and measurable elements, even if these are not the most meaningful, rigorous, or ethically responsible. The two practices we have proposed, making constitutive commitments explicit and exercising greater care with terms and expressions, are offered in that spirit. They will not resolve the intrinsic conceptual difficulties of well-being. What they can do is make the exchange between those who apply these instruments and those who question them more productive. The current moment, referred to as a "happiness revolution" by Easterlin and O'Connor (2025), could either become a revolution aligned with scientific rigour, or merely a discursive one, a likely risk, according to Flanagan et al. (2023).

Author

Luciano Sewaybricker
Psychology Department, Centro Universitário São Camilo, Brazil
https://orcid.org/0000-0001-6917-7275
luciano.sewaybricker@prof.saocamilo-sp.br

Author contribution statement

LES conceived the research, conducted the investigation, wrote the original draft, and reviewed and edited the final manuscript.

Conflict of interest statement

The author declares no conflict of interest that could have influenced this paper.

Funding

This study was partly financed by the São Paulo Research Foundation (FAPESP), Brazil. Process Number 2025/17168-9.

AI statement

The author used ChatGPT (GPT-5) for the initial translation of the manuscript from Portuguese to English; Claude (Opus 5) to critique the manuscript as a simulated reviewer and to help prepare the response to reviewers; and Grammarly to improve readability. The author reviewed, edited, and verified all AI-assisted text and references, and takes full responsibility for the publication's content.

Acknowledgements

I thank the members of the International Society for Quality-of-Life Studies (ISQOLS) for their feedback on this research presentation at the 2025 ISQOLS conference. I am particularly grateful to Parviz Dabir-Alai, Mark Fabian, Kelsey J. O'Connor, and Joonas Uotinen for taking the time to read earlier versions of this work and offer valuable suggestions. I also thank two anonymous reviewers, whose comments improved the manuscript considerably; the comparison with intelligence research and the magic-wand example in the discussion of the third argument were suggested by Reviewer 1.

Publishing Timeline

Received 30 March 2026
Revised version received 14 August 2026
Accepted 10 October 2026
Published 12 November 2026

References

Adler, A., & Seligman, M. (2016). Using well-being for public policy: Theory, measurement, and recommendations. *International Journal of Wellbeing, 6*(1), 1–35. https://doi.org/10.5502/ijw.v6i1.429

Ahmed, S. (2007). The happiness turn. *New Formations, 63*, 7–14. https://journals.lwbooks.co.uk/newformations/vol-2007-issue-63/

Ahmed, S. (2023). *The feminist killjoy handbook: The radical potential of getting in the way*. Hachette UK.

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

Alexandrova, A. (2018). Can the science of well-being be objective? *The British Journal for the Philosophy of Science, 69*(2), 421–445. https://doi.org/10.1093/bjps/axw027

Alexandrova, A., & Fabian, M. (2022). Democratising measurement: Or why thick concepts call for coproduction. *European Journal for the Philosophy of Science, 12*(7), 1–23.

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

Angner, E. (2011). Are subjective measures of well-being 'direct'? *Australasian Journal of Philosophy, 89*(1), 115–130. https://doi.org/10.1080/00048400903401665

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

Aristotle. (1934). *Aristotle in 23 Volumes* (transl. H. Rackham). Harvard University Press.

Barrington-Leigh, C. (2025). Are international happiness rankings reliable? *arXiv preprint*. https://arxiv.org/html/2509.06867v1

Bentham, J. (1823). *An introduction to the principles of morals and legislation*. Clarendon Press.

Bond, T. N., & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy, 127*(4), 1629–1640.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67–90.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Cromby, J. (2011). The greatest gift? Happiness, governance and psychology. *Social and Personality Psychology Compass, 5*(11), 840–852. https://doi.org/10.1111/j.1751-9004.2011.00398.x

Da Silva Sinha, V., Sinha, C., & de Araújo Sampaio, W. B. (2017). The many ways to count the world: Counting terms in indigenous languages and cultures of Rondônia, Brazil. *Brief Encounters, 1*(1). https://doi.org/10.24134/be.v1i1.26

De Neve, J. E., & Ward, G. (2025). *Why workplace wellbeing matters: The science behind employee happiness and organizational performance*. Harvard Business Press.

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology, 63*, 453–482. https://doi.org/10.1146/annurev-psych-120710-100353

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*(3), 542–575. https://doi.org/10.1037/0033-2909.95.3.542

Diener, E. (2020). Happiness: The science of subjective well-being. In R. Biswas-Diener & E. Diener (Eds.), *Noba textbook series: Psychology*. DEF Publishers. http://noba.to/qnw7g32t

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment, 49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Diener, E., Sapyta, J. J., & Suh, E. (1998). Subjective well-being is essential to well-being. *Psychological Inquiry, 9*(1), 33–37. https://doi.org/10.1207/s15327965pli0901_3

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin, 125*(2), 276–302. https://doi.org/10.1037/0033-2909.125.2.276

Disabato, D. J., Goodman, F. R., & Kashdan, T. B. (2025). The hierarchical framework of wellbeing (HiFWB). *Frontiers in Psychology, 16*, 1515423. https://doi.org/10.3389/fpsyg.2025.1515423

Duncan, G. (2010). Should happiness-maximization be the goal of government? *Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being, 11*(2), 163–178. https://doi.org/10.1007/s10902-008-9129-y

Duncan, G. (2019). Happiness and the new politicization of subjectivity. In N. Hill, S. Brinkmann, & A. Petersen (Eds.), *Critical happiness studies* (pp. 83–97). Routledge.

Easterlin, R. A. (1974). Does economic growth improve the human lot? Some empirical evidence. In P. A. David & M. W. Reder (Eds.), *Nations and households in economic growth* (pp. 89–125). Academic Press.

Easterlin, R. A., & O'Connor, K. J. (2025). *The happiness revolution in Europe*. Cambridge University Press.

Fabian, M. (2022a). Scale norming undermines the use of life satisfaction data for welfare analysis. *Journal of Happiness Studies, 23*(4), 1509–1541. https://doi.org/10.1007/s10902-021-00460-8

Fabian, M. (2022b). *A theory of subjective wellbeing*. Oxford University Press.

Fabian, M., Alexandrova, A., Coyle, D., Agarwala, M., & Felici, M. (2024). Wellbeing public policy needs more theory. https://doi.org/10.17863/CAM.112205

Fabian, M., Kaiser, C., Panasiuk, S., Funk, S., Pountney, L., & Brett, C. (2026). What cognitive interviews reveal about life satisfaction measurement. Preprint. https://doi.org/10.31234/osf.io/r67g9_v1

Flanagan, O., LeDoux, J. E., Bingle, B., Haybron, D. M., Mesquita, B., Moody-Adams, M., ... & Wilson, Y. Y. (2023). *Against happiness*. Columbia University Press.

Frey, B. S., & Stutzer, A. (2017). *Public choice and happiness* (No. 2017-03). CREMA Working Paper. Retrieved November 19, 2024, from https://hdl.handle.net/10419/214596

Frijters, P., Krekel, C., Sanchis, R., & Santini, Z. I. (2024). The WELLBY: A new measure of social value and progress. *Humanities and Social Sciences Communications, 11*(1), 1–12. https://doi.org/10.1057/s41599-024-03229-52

Gould, S. J. (1981). *The mismeasure of man*. W. W. Norton & Company.

Haybron, D. M. (2007). Do we know how happy we are? On some limits of affective introspection and recall. *Nous, 41*(3), 394–428. https://doi.org/10.1111/j.1468-0068.2007.00653.x

Haybron, D. M. (2024). Well-being policy: Consensus hallmarks and cultural variation. In J. D. Rooney & P. Zoll (Eds.), *Beyond classical liberalism* (pp. 205–221). Routledge India.

Helliwell, J. F., Layard, R., & Sachs, J. D. (Eds.). (2012). *World Happiness Report 2012*. UN Sustainable Development Solutions Network.

Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (Eds.). (2026). *World Happiness Report 2026*. University of Oxford: Wellbeing Research Centre.

Joshanloo, M. (2025). A person-centered exploration of happiness conceptions and their relation to hedonic and eudaimonic well-being. *Journal of Happiness Studies, 26*(7), 118. https://doi.org/10.1007/s10902-025-00953-w

Kahneman, D. (1992). Reference points, anchors, norms and mixed feelings. *Organizational Behavior and Human Decision Processes, 51*, 296–312. https://doi.org/10.1016/0749-5978(92)90015-Y

Kahneman, D., & Riis, J. (2005). Living, and thinking about it: Two perspectives on life. In F. A. Huppert, N. Baylis, & B. Keverne (Eds.), *The science of well-being* (pp. 285–304). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198567523.003.0011

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

Kaiser, C., & Lepinteur, A. (2025). Measuring the unmeasurable? Systematic evidence on scale transformations in subjective survey data. University of Oxford Wellbeing Research Centre. Working Paper 2503. https://doi.org/10.5287/ora-vyq7o5eqb

Kaiser, C., & Oswald, A. J. (2022). The scientific value of numerical measures of human feelings. *Proceedings of the National Academy of Sciences, 119*(42), e2210412119. https://doi.org/10.1073/pnas.2210412119

Kaiser, C., & Vendrik, M. C. M. (2020). How threatening are transformations of happiness scales to subjective wellbeing research? IZA Discussion Paper No. 13905.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191–207. https://www.jstor.org/stable/3593024

Layard, R., & De Neve, J. E. (2023). *Wellbeing science and policy*. Cambridge University Press.

Liberini, F., Redoano, M., & Proto, E. (2017). Happy voters. *Journal of Public Economics, 146*, 41–57. http://dx.doi.org/10.1016/j.jpubeco.2016.11.013

Linton, M.-J., Dieppe, P., & Medina-Lara, A. (2016). Review of 99 self-report measures for assessing well-being in adults: Exploring dimensions of well-being and developments over time. *BMJ Open, 6*(7), e010641. https://doi.org/10.1136/bmjopen-2015-010641

Liu, S., & Netzer, N. (2023). Happy times: Measuring happiness using response times. *American Economic Review, 113*(12), 3289–3322. https://doi.org/10.1257/aer.20211051

Loewenstein, G. (2009). That which makes life worthwhile. In A. B. Krueger (Ed.), *Measuring the subjective well-being of nations* (pp. 87–106). National Bureau of Economic Research.

Lomas, T., Pawelski, J. O., & VanderWeele, T. J. (2024). A flexible map of flourishing: The dynamics and drivers of flourishing, well-being, health, and happiness. *International Journal of Wellbeing, 13*(4), 3665, 1–38. https://doi.org/10.5502/ijw.v13i4.3665

Lyubomirsky, S. (2008). *The how of happiness: A scientific approach to getting the life you want*. Penguin Press.

Mahoney, J. (2023). Subjective well-being measurement: Current practice and new frontiers. Working Paper No. 17. OECD Papers on Well-being and Inequalities. https://doi.org/10.1787/4e180f51-en

Margolis, S., Schwitzgebel, E., Ozer, D. J., & Lyubomirsky, S. (2021). Empirical relationships among five types of well-being. In M. T. Lee, L. D. Kubzansky, & T. J. VanderWeele (Eds.), *Measuring well-being: Interdisciplinary perspectives from the social sciences and the humanities* (pp. 377–406). Oxford University Press. https://doi.org/10.1093/oso/9780197512531.003.0014

McClimans, L. (2013). The role of measurement in establishing evidence. *Journal of Medicine and Philosophy, 38*(5), 520–538. https://doi.org/10.1093/jmp/jht041

McMahon, D. M. (2006). *Happiness: A history*. Grove Press.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77

O'Connor, K. J. (2023). Measuring societal progress. Working Paper No. 1. International Society for Quality of Life Studies. https://isqols.org/Workingpaperseries

Oishi, S., Graham, J., Kesebir, S., & Galinha, I. C. (2013). Concepts of happiness across time and cultures. *Personality and Social Psychology Bulletin, 39*(5), 559–577. https://doi.org/10.1177/0146167213480042

Panasiuk, S. L., McCanny, A., & Cheung, F. (2025). Methods reflect values: Evaluating the shortcomings of the average for measuring population well-being. *Journal of Personality and Social Psychology, 128*(6), 1355–1370. https://doi.org/10.1037/pspp0000549

Plant, M., McGuire, J., Dupret, S., Dwyer, R., & Stewart, B. (2025). Giving to others: How to convert your money into greater happiness for others. In J. F. Helliwell, R. Layard, J. D. Sachs, J. E. De Neve, L. B. Aknin, & S. Wang (Eds.), *World Happiness Report 2025*. https://doi.org/10.18724/whr-e0cy-0r69

Poller, C. C. (2026). Thin Metrics for a Thick Concept: Trade-offs in Global Well-Being Measurement. Unpublished manuscript, accepted for publication in the *PSA 2026 Contributed Paper Proceedings*.

Prati, A., & Senik, C. (2022). Feeling good is feeling better. *Psychological Science, 33*(11), 1828–1841. https://doi.org/10.1177/09567976221096158

Prati, A., & Senik, C. (2025). Is it possible to raise national happiness? Centre for Economic Performance. Discussion Paper 2068. https://cep.lse.ac.uk/pubs/download/dp2068.pdf

Rodriguez Duque, S., Tal, E., & Barbic, S. P. (2024). The role of ethical and social values in psychosocial measurement. *Measurement, 225*, 113993. https://doi.org/10.1016/j.measurement.2023.113993

Samuelsson, C., Dupret, S., Plant, M., & Kaiser, C. (2023). Can we trust wellbeing surveys? A pilot study of comparability, linearity, and neutrality. Happier Lives Institute. Retrieved November 19, 2024, from https://www.happierlivesinstitute.org/comparability-linearity-and-neutrality-pilot-report/

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schneider, L., & Schimmack, U. (2009). Self-informant agreement in well-being ratings: A meta-analysis. *Social Indicators Research, 94*(3), 363–376. https://doi.org/10.1007/s11205-009-9440-y

Sollis, K., Yap, M., Campbell, P., & Biddle, N. (2022). Conceptualisations of wellbeing and quality of life: A systematic review of participatory studies. *World Development, 160*, 1–39. https://doi.org/10.1016/j.worlddev.2022.106073

Sternberg, R. J., & Detterman, D. K. (Eds.). (1986). *What is intelligence? Contemporary viewpoints on its nature and definition*. Ablex Publishing.

Sewaybricker, L. E., & Massola, G. M. (2023). Against well-being: A critique of positive psychology. *History of the Human Sciences*, *36*(1), 131-148. https://doi.org/10.1177/09526951221114733

Sewaybricker, L. E., & Massola, G. M. (2025). Dialogues on happiness with the Guarani Mbyá in South America and the limits of the science of well-being. *International Journal of Wellbeing, 15*(2), 2759, 1-23, https://doi.org/10.5502/ijw.v15i2.2759

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

Tal, E. (2020). Measurement in science. *The Stanford Encyclopedia of Philosophy*. Retrieved November 19, 2024, from https://plato.stanford.edu/archives/fall2020/entries/measurement-science

Truijens, F., De Smet, M. M., Vandervoorde, M., Desmet, M., & Meganck, R. (2023). What is it like to be the object of research? On meaning making in self-report measurement and validity of data in psychotherapy research. *Methods in Psychology, 8*, 1–9. https://doi.org/10.1016/j.metip.2023.100118

UK HM Treasury. (2022). *The Green Book – Central government guidance on appraisal and evaluation*. Retrieved November 19, 2024, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1063330/Green_Book_2022.pdf

Veenhoven, R. (1984). *Conditions of happiness*. Springer. http://dx.doi.org/10.1007/978-94-009-6432-7

Veenhoven, R. (2017). Measures of happiness: Which to choose? In G. Brulé & F. Maggino (Eds.), *Metrics of subjective well-being: Limits and improvements* (pp. 65–84). Springer International Publishing. https://doi.org/10.1007/978-3-319-61810-4_4

Ward, G., De Neve, J.-E., Ungar, L. H., & Eichstaedt, J. C. (2021). (Un)happiness and voting in U.S. presidential elections. *Journal of Personality and Social Psychology, 120*(2), 370–383. https://doi.org/10.1037/pspi0000249

Weijers, D., & Jarden, A. (2013). The science of happiness for policymakers: An overview. *Journal of Social Research & Policy, 4*(2), 21–40. https://repository.openpolytechnic.ac.nz/handle/11072/1629

Weiss, S., & Roberts, R. D. (2018). Using anchoring vignettes to adjust self-reported personality: A comparison between countries. *Frontiers in Psychology, 9*, 325. https://doi.org/10.3389/fpsyg.2018.00325

White, N. (2006). *A brief history of happiness*. Blackwell Publishing. https://doi.org/10.1002/9780470690888

Wierzbicka, A. (2004). "Happiness" in cross-linguistic & cross-cultural perspective. *Daedalus, 133*(2), 34–43. https://doi.org/10.1162/001152604323049370

Yap, M., & Yu, E. (2016). Operationalising the capability approach: Developing culturally relevant indicators of indigenous wellbeing – an Australian example. *Oxford Development Studies, 44*(3), 315–331. https://doi.org/10.1080/13600818.2016.1178223