Synthese (2021) 199:5719–5746
https://doi.org/10.1007/s11229-021-03043-5

Can happiness measures be calibrated?

Mats Ingelström[1] · Willem van der Deijl[2]

Received: 20 July 2020 / Accepted: 16 January 2021 / Published online: 29 January 2021

Abstract
Measures of happiness are increasingly being used throughout the social sciences. While these measures have attracted numerous types of criticisms, a crucial aspect of these measures has been left largely unexplored—their calibration. Using Eran Tal's recently developed notion of calibration we argue first that the prospect of continued calibration of happiness measures is crucial for the science of happiness, and second, that continued calibration of happiness measures faces a particular problem—The Two Unknowns Problem. The Two Unknowns Problem relies on the claim that individuals are necessarily a part of the measurement apparatus in first person measures of happiness, and the claim that we have no reason to believe that the evaluation standards people employ are invariant across individuals and time. We argue that calibrating happiness measures therefore involves solving an equation with two unknowns—an individual's degree of happiness, and their evaluation standards—which is, generally, not possible. Third, we consider two possible escape routes from this problem and we suggest that the most promising route requires yet unexplored empirical and theoretical work on linking happiness to behavioral or neural evidence.

Keywords Happiness · Calibration · Measurement

1 Introduction

Are our lives happier now than those of our predecessors thirty years ago (e.g. Easterlin 1995; Hagerty and Veenhoven 2003; Easterlin et al. 2010)? Are people in Finland happier than they are in the United States, of in fact, any other country on earth (Helliwell et al. 2019)? To answer these and similar questions researchers

✉ Willem van der Deijl
wjavanderdeijl@gmail.com

Mats Ingelström
mats.ingelstrom@philosophy.su.se

[1] Stockholm University, Stockholm, Sweden

[2] Tilburg University, Tilburg, The Netherlands

and statistical agencies today routinely measure happiness. By asking subjects to rate how happy or satisfied they are with their life, researchers collect data that they use to measure happiness. What can these measures tell us about such research questions?

In the literature, happiness measurement has received various skeptical responses. Some have claimed that the measurements do not capture happiness at all (e.g. Johns and Ormerod 2007). Others argue that even if they capture some aspects of happiness, they grossly misrepresent their targets since happiness is "multidimensional" (see MacKerron 2012 for a discussion, who cites; Annas 2004, p.46 as an example of this concern). Some scholars hold on to the old positivist idea that happiness is too subjective and that therefore comparisons of happiness are meaningless (for a recent discussion, see Kaminitz 2018). A common concern amongst scholars is also that first-person reports in general are problematic for measurement (for a discussion and defence, see Howard 1994). Here we do not address any of these issues directly. We take onboard the assumption that researchers today can measure happiness, and instead we focus on the epistemic status of happiness measurement. If we accept that happiness measurements are about happiness, the next question we should ask is how these measures justify claims about happiness. We should therefore discuss if, and how, happiness measures can be calibrated.

Proponents of happiness measures typically agree that the measures currently are inexact and rough. As a reply, they claim that the measures have passed tests of validity (Diener et al. 2013), and they expect that the measures will improve as the field matures. Paul Dolan and Tessa Peasgood, in an article advocating for the usage of happiness measures in public policy, write: "Measures of subjective evaluation are still being developed, and ongoing improvements in their validity should follow." (2008, p. s25). And, more elaborately, Ed Diener writes:

> "The study of [happiness][1] is growing into a major area in the social sciences. It is imperative that we build this area on a solid measurement foundation. The major message is simply this: [happiness] measures are good, but they can be better. […] [E]xisting measures served surprisingly well during the initial stages of study in this field. But measurement should be an increasingly sophisticated enterprise in any scientific area. (Diener 1994, p. 146)".

Success stories from the history of science indeed suggest that we should expect measures to improve as the scientific field matures. The history of science includes many examples—such as the measurement of time and temperature—where early measurement was inexact, and where the measurement over time were made more exact (Hempel 1952; Chang 2004; van Fraassen 2008).

The important thing to note, however, is that thermometers and clocks improved because they could be—and over time were—increasingly calibrated. Calibration, as the measurement theorist Tal (2017b) recently characterized it, is in its most general sense a process whereby the relationship between the instrument reading and knowledge claims about the parameter being measured is established.

[1] Diener uses the word "subjective well-being" and "happiness" interchangeably. In the quoted passage, he used the former term.

A central epistemic question for happiness research should therefore now be: to what extent can measurements of happiness be calibrated? Our aim is to contribute to answering this question.

The structure of the article is as follows. In Sect. 2, we describe the terminology we adopt, including what we here mean by "happiness". In Sect. 3, we describe Tal's notion of calibration. We also briefly discuss how calibration relates to the psychometric notion of validation, and to Chang's problem of nomic measurement. In Sect. 4, we argue that calibration ought to be seen as a central concern for happiness scholars.

In Sect. 5, we present the *Two Unknowns Problem*. This problem poses a principled challenge for the prospect of continued calibration of happiness measurements. Very briefly put, the problem is that self-reports of happiness depend on at least two mental variables: a person's actual happiness, and what we shall call a person's *evaluation standards*—mental factors that determine how a person rates a specific degree of happiness. This double dependence, we shall argue, undermines establishing a close relationship between the data we get from happiness measurement and happiness itself. In Sect. 6, we present some ways to get around this problem, but suggest that none of them are currently fully satisfactory, primarily due to the lack of theory that connects degrees of happiness to behavioral or neural signs.

Our concluding claim is that continued calibration poses a particularly pernicious challenge in the context of happiness measurement, since it depends on getting around the Two Unknowns problem. At the same time, continued calibration of measurements is central for the epistemic progress in the field. The prospect of further epistemic progress in happiness measurement therefore depends on finding ways out of the Two Unknowns problem.

Before we move on, we should clarify that we are not concerned with the question whether measures of happiness represent degrees of wellbeing (e.g. Nussbaum 2008; Raibley 2012; van der Deijl 2017b), whether happiness is intrinsically valuable in itself or not (e.g. van der Rijt 2013), or whether happiness should be an aim of public policy (e.g. Haybron and Tiberius 2015; Sugden and Teng 2016; and also van der Rijt 2013). While much of the significance of the project to measure happiness depends on these questions, they go beyond the scope of our argument.

Furthermore, the problem we raise is broader than what Fleurbaey and Blanchet calls "the calibration problem" for happiness data (2013, p. 181). Fleurbaey and Blanchet are concerned with "calibrating" the response scale of happiness-questionnaires, so that the meaning of each category is unambiguous. Although calibration in their sense is important and probably crucial, we are not directly concerned with interpreting the meaning of categories of happiness-data. We do not claim that devices for measuring happiness cannot be used—and perhaps even be calibrated—as devices for measuring some interesting quantity. We are concerned with the question to what extent the procedures available for measuring happiness can justify claims about individuals' degree of happiness.

Finally, we use the term "individuals" to refer to those individuals whose happiness is being measured. Whenever we refer to researchers or others measuring happiness, we use the term "researcher".

2 Preliminaries

2.1 A tripartite distinction

In this article, we distinguish between three central concepts relating to happiness measurement. First, there is the concept of happiness (2.2). Second, there is the quantitative evaluation of happiness that individuals may make (2.3). And third, there are devices for measuring happiness (2.4). We will elaborate on these in turn.

2.2 What happiness is

The concept of happiness requires clarification as the term is used in semantically and substantively different senses. First, there are at least two different semantic uses of the term ‘happiness’ (e.g. Haybron 2000). Firstly, the term happiness can be used to denote a positive mental state. On this meaning, happiness refers only to psychological qualities.[2] Alternatively, happiness is sometimes used to refer to a state that goes beyond the merely mental, a more general state of being that is positive for the individual who has it. In particular within Aristotelian usages of the term, the term ‘happiness’ is not limited to our mental states, but describes what other philosophers generally call wellbeing, or welfare. For our argument, not much bears on the usage of these two terms. However, for the purpose of clarity, we will focus on happiness in the simplest and most common sense: happiness as a positive psychological state. But it is good to note that the following discussions will have bearing on the broader usage of the term happiness as well. On this broader notion, it seems plausible that still a significant part of what constitutes happiness would be made up of happiness as a psychological state. Happiness as a psychological state is most plausibly at least one feature of happiness in a broader sense (this is true of objective list theories, such as Fletcher 2013; but also of Bishop 2015′s network theory).[3] Thus, broadening the scope of what happiness can refer to will not eliminate the calibration problem for happiness measures, but will at best suppress the relative importance of happiness as a mental state. Consequently, if measures of happiness as a psychological state cannot be calibrated adequately, moving to a more inclusive notion will not change that.

There are several conceptions about what constitutes happiness as a positive psychological state, but they all hold that happiness is determined by mental states. Prominent theories about happiness in the sense used here include life-satisfaction theory (Sumner 1996), hedonism about happiness (Mill 1871; Tännsjö 2007), and the currently popular emotional state views of happiness (Haybron 2005, 2008; Rossi 2018). On the first view, a person’s happiness is the extent to which that person is satisfied with their life as a whole, which has both affective and cognitive components. On this view, when you are happy, you judge your life to be good,

[2] We use the terms “psychological state” and “mental state” interchangeably.

[3] Though both refer to this broader notion of happiness as “well-being”, as is common in the philosophical literature.

and/or you feel good about your life. On hedonistic accounts, a happy person is a person who has a positive net balance of pleasurable over unpleasurable experiences in their life. And, on emotional state views (Haybron 2005; 2008; Rossi 2018), being happy means being in a positive emotional state, which consists of having emotional dispositions that favor positive affective states, such as joyfulness, elation, confidence. These emotional states need not all be conscious mental states (Haybron 2005), however, conscious mental states, or occurrent mental states, are still necessary and important for happiness. For the purposes of this article, we do not favor a particular view on happiness. What is central is that happiness is a mental state: it is a feature internal to our mind, and is, for at least a significant part, constituted by conscious mental states.

Regardless of the particular view of what happiness is, we consider it a *quantity*. Some may believe it is not. For sake of brevity, we do not try to establish this claim here. A quantity is an attribute of something, by which it is meaningful to compare and order the thing in relation to other things. Length, weight, eye-sight, IQ, blood-pressure and wealth are familiar quantities which we attribute to individuals. Somewhat simplified, quantities are properties that come in degrees. That happiness is a quantity amounts to the claim that persons can be more or less happy. In the terminology often adopted in happiness literature, degree of happiness is the "target concept" of happiness measurements. A target concept is that which a measurement procedure is designed to provide knowledge about.

We furthermore assume that degrees of happiness are both intra- and interpersonally comparable. Interpersonal comparability means that the relation between two persons' degrees of happiness mirrors the relation between how happy they are. If Anna is happy to degree 9 and Bob is happy to degree 6, then Anna is happier than Bob is. Intrapersonal comparability is like interpersonal comparability, but for comparisons within lives rather than across lives. If Anna was happy to degree 10 when she was a PhD-student, and she is happy to degree 9 now, then she was happier when she was a PhD-student than she is now. Economists are often skeptical about interpersonal comparability,[4] but we believe the assumption gains support from how we in everyday life think about happiness. Even though it may often be difficult to know, it seems we can understand what it means to say that one person is happier than another.

2.3 Quantitative evaluations of happiness

How happy we are should be distinguished from how happy we judge that we are. When we judge how happy we are, or how happy a close one is, we make an evaluation of happiness. A quantitative evaluation of happiness is an evaluation that can be expressed in terms of degrees. While how happy we are at a moment—i.e. our degree of happiness at that moment—will typically frequently enter our awareness, we are not constantly quantitatively evaluating our happiness. Quantitatively

[4] Though the concept of interpersonal comparability in economics is typically connected to *utility* rather than happiness.

evaluating our happiness is a cognitive process that we may enter into occasionally, such as when a researcher asks us to do so (Schwarz and Strack 1999). How to make sense of such evaluations may depend on the conception of happiness, discussed above. The distinction between happiness and evaluations of happiness is most straightforward on hedonistic or emotional state conceptions of happiness (see van der Deijl 2017a for a discussion). A quantitative evaluation of our hedonic or emotional state is much like other quantified judgments of observations that we make: we internally observe these states, and quantify them, just as we would if we were asked to quantifiably evaluate the messiness in room, or how warm it feels (Diener, Oishi, and Lucas 2009).

In case of life-satisfaction views of happiness, this may appear to be different. The two concepts—(i) happiness as evaluation of one's life, and (ii) one's evaluation of one's happiness—appear to be very similar in these cases. However, there is an important difference between them. On life-satisfaction views of happiness, people may still be happy even if they are not in the process of evaluating their life. Life satisfaction refers to a general sense of satisfaction with your life that is not wholly cognitive (Sumner 1996). While happiness-constituting life-evaluations are necessarily quantitative, this does not imply that we are making a conscious quantitative judgment whenever we are satisfied about our life in the relevant sense. Quantitative evaluations of happiness, or life satisfaction, however, are conscious quantitative judgments.

Happiness and evaluations of happiness are not merely different kinds of things, they can also come apart. A person can evaluate her happiness higher than another, while not being happier. The possibility of such divergence seems particularly likely when we consider comparisons across persons with different personalities, or with different cultural background.[5] This divergence will have ramifications for how happiness measures can be calibrated, we will return to this in Sect. 5.

2.4 How happiness is measured

The science of happiness has expanded rapidly over the recent decades. The measures of happiness that this field of science uses are typically based on first-person reports of happiness. The procedures involve asking individuals, on a questionnaire format, to express their own happiness, or life-satisfaction, on a scale—for example, on a scale from 0 to 10, or on a Likert scale (Likert 1932), ranging from very unhappy, to very happy.

Many procedures/devices for measuring happiness have been developed. They differ both in how the questions are framed and in how they characterize the quantity of happiness. Broadly, the measurement of wellbeing-pertaining psychological states is called subjective wellbeing measurement. Devices for measuring happiness

[5] Empirical work suggests that differences in reported happiness to a large extent can be explained by differences in personality (Steel et al. 2008). Although personality may impact happiness, this observed co-variance can be due to differences in evaluations of happiness. Similarly, for cross-cultural comparisons, it has been argued that differences in reported happiness are due to cultural response bias (Cummins 2013). We will explain this further in Sect. 5.

come in mainly four varieties: (1) *measures of life satisfaction*—asking individuals to rate their overall satisfaction with their life on a scale, (2) *global measures of happiness*—asking individuals to rank their happiness considering life as a whole on a scale, (3) *local measures of happiness*—asking individuals to rate their momentary experiences of happiness (such as their days or specific moments) on a scale (Kahneman and Krueger 2006), and (4) *combined measures*—using a mix of these measures, such as Ed Diener's Satisfaction with Life Scale (SWLS; Diener et al. 1985).

We here refer to the output of a happiness measurement-device as "happiness reading". The reading is commonly expressed simply as a number. Although some happiness measurements use verbal response scales—such as "very satisfied", "moderately satisfied", etc.—these categories get translated into numbers. Procedures that involve multiple questions, such as the Satisfaction with Life Scale, contain rules for calculating a numerical output based on the answer of each question.

To clarify, we see happiness measurements as a measurement activity where the intended outcome is a knowledge claim about how happy the subject being measured is.[6] We are not discussing nominalist happiness measures, where the indented outcome is a knowledge claim about how someone rates herself in response to happiness related questions.[7] Furthermore, as should be clear from the above distinction between happiness and evaluations of happiness, we are not discussing measures where the intended outcome is a knowledge claim about how happy someone evaluates herself to be.

3 Calibration and nomic measurement

3.1 The epistemic role of calibration

In a series of recent work, Tal (2015, 2017a; b) describes calibration as a process which aims to establish a relation between readings of a measurement device and measurement outcomes. Simply put, scientists measure things to get an outcome, "a knowledge claim attributing … parameter values to the object … being measured" (Tal 2017b, p. 35). The knowledge claim about the object is, however, inferred from the observed reading of the instrument. Having a calibrated measurement device justifies making this inference. Tal gives the example of an ammeter (an instrument used to measure electric current):

[6] There is always some uncertainty involved in measurement. Although outcomes are commonly expressed as numerals, a more proper representation of a measurement outcome includes also a description of the associated uncertainty.

[7] As an anonymous referee pointed out, it may be much less complex to calibrate a measure that instead of the target "happiness" has the target "self-reported happiness". We agree. However, we do believe that a central aim of happiness science is, and should be, to enhance our knowledge of happiness. Calibrating measurements of self-reported happiness may be a first step, as there plausibly is some relation between "self-reported happiness" and "happiness". For the field to take the next step, however, the ambition ought to be to find ways to measure happiness.

> To attain the status of a measurement outcome, a set of values must be abstracted away from its concrete method of production and pertain to some quantity objectively, namely, be attributable to the measured *object* rather than the idiosyncrasies of the measuring instrument, environment and human operators. Consider the ammeter: the outcome of measuring with an ammeter is a knowledge claim about the electric current running through the input wire. The position of the ammeter pointer relative to the dial is a property of the ammeter rather than the wire. When measuring electric current, then, claims concerning the position of the pointer are not candidate measurement outcomes. […] It is only once theoretical and statistical background assumptions are made and tested about the behaviour of the ammeter and its relationship with the wire and environment that one can infer values of electric current from the position of the pointer. The ultimate aim of calibration is to establish such inferences and characterize their uncertainty … (2017b, p. 35, emph. original, footnotes omitted).

Although Tal mostly focuses on measurement in the natural sciences in his discussions and examples, his epistemic account of calibration is general and elaborate enough to be useful and illuminative also for—perhaps especially for—measurement in the social sciences.

To further clarify the role of calibration, it is helpful to elaborate on the distinction Tal makes between *instrument reading* (Tal also uses the term ‘indication’) and *measurement outcome*. An instrument reading is a property of the concrete measurement process, it is the “output” we get when the measurement process is in its final state. A measurement process should here be understood to include the equipment and practical procedure for interacting with a measured object, it also includes mathematical operations and statistical models that may be used to derive the final state of the process. An instrument reading can for example be the height of a mercury bar in a thermometer, or a digital display, or a score derived from answers in a questionnaire. The measurement outcome, on the other hand, is a knowledge claim about the thing being measured (Tal in Mößner et al. 2017, p. 233). A user of a measurement device can typically observe the reading, but not the outcome.

The key measurement epistemological question is how a reading justifies an outcome. Tal’s answer is: through calibration. By relying on an established relation between readings and outcomes, a user of a measurement device can be justified in making inferences about a measured quantity of an object on which the device is used. Calibration is thus central for the epistemology of measurement.

When we in everyday situations use measurement devices, we take for granted that they have been calibrated. Household measurement devices are often designed to conceal the difference between reading and outcome. We effortlessly interpret the symbols displayed on an electronic thermometer as an outcome that describes the temperature of the measured object. But as Tal emphasizes, this relation between reading and outcome is far from guaranteed, nor easy to establish. It is the fruit of much epistemic labor that people today so readily can use scales, timers, thermometers, rangefinders, and decibel meters to successfully ascribe duration, temperature, distance and noise.

It follows from Tal's distinction between reading and outcome that strictly speaking there is no "direct measurement". Even such seemingly direct measurement devices as rulers and measuring cups rely on calibration for their successful uses. To illustrate, let us take the case of measuring length using a rigid rod. The inference that one object is longer than another, based on measuring them both with a measuring rod, is only justified given some theoretical background assumption. A key assumption here is that the length of the rod itself does not vary depending on the object it is measuring. Furthermore, it must be assumed that contextual factors such as where the measurement takes place, by whom, at what time, under what circumstances, etcetera, do not (to a significant degree) affect the relation between reading and outcome. Through calibration, researchers aim to determine whether, when, and to what extent such assumptions are justified.

Consider the following relationship between a reading from a happiness measurement procedure (h_r) and an associated outcome, a justified ascription of happiness to some individual (h_O):

$$h_O = f\left(h_r\right) \tag{1}$$

The function f describes the relationship between the justified ascription of happiness to a measured individual, and the indication the measurement device gives when used on that individual. As the process of calibration aims at establishing the relationship between a reading (h_r) and a knowledge claims about happiness (h_O), the aim is to identify this function.

However, as in the example of measuring rods, the relationship expressed by f depends on contextual factors. An important step in calibration is to identify these factors and estimate how much they impact the relationship between reading and outcome. This involves theorizing. As Tal points out, calibration "does not … constitute an independent test of measurement outcomes, if by 'independent' one means 'free of any theoretical or statistical background assumption'. […] [C]alibration is better viewed as involving … a test for the invariance of measurement outcomes across different instruments and circumstances, and across different ways of applying background theories." (2017b, p. 43).

3.2 Calibration and validation

In psychometrics, the field where most current happiness measurements are being developed, researchers typically talk about validation, rather than calibration. Happiness measures are often claimed to be validated measures (for an overview, see Diener et al. 2013). Validating a measure is considered the main epistemic step in psychometric measurement.[8] Simply stated, validation aims to establish whether a measure is valid, that is if it measures what it is meant to measure (Borsboom et al.

[8] For good overviews in the modern measurement literature of the concept of validation, see Vessonen (2019) and Alexandrova and Haybron (2016).

2009). Although it is often unclear what, more precisely, claims of validity amount to,[9] some of common validation procedures are calibration in Tal's sense.

Some validation exercises are aimed at establishing properties of measurements that are prior to calibration. Tests such as those aiming to establish "face validity" could be interpreted as establishing that the measurement device at least theoretically could provide information about the sought quantity. Tests for internal consistency of multiple items in a questionnaire are also not about calibrating the measurement, but about establishing that it measures one quantity, rather than many.

Other kinds of validation exercises can, or even should be, seen as steps in measurement calibration. Some validation techniques are concerned with how sensitive a measurement procedure is to different circumstances. Tests for construct validity (Cronbach and Meehl 1955), for example, analyze if the numerical representations of the measurement readings correlate with data on other quantities in expected ways. Construct validation occurs by testing both how the measurement readings converge with factors for which there are theoretical reason to believe that they are closely associated with the target concept ("convergent validity"), and how the readings do *not* converge with factors for which there are theoretical reason to believe that they are not closely associated with the target concept ("discriminant validity").

To only mention a few examples of such tests: researchers have investigated how happiness data correlates positively with optimism, hope, self-esteem and positive affect (Abdel-Khalek 2006), whether it varies with the weather (Schwarz and Clore 1983; Connolly 2013), with temporary moods of the test subject (Kozma et al. 2000), with the design and settings of the measurement (Schwarz et al. 1987), or with personality traits (Abbott et al. 2008; Steel and Ones 2002).[10] Put in the terminology of calibration, studies like these "test for the invariance of measurement outcomes across different … circumstances", which Tal (see above) describes as part of calibration.

In theory, construct validation techniques could go a long way towards calibrating happiness measurements. Measurement and theory could, in an iterative epistemic process, reinforce each other by providing more and more accurate predictions and detailed theories that align with those predictions. In practice, there are concerns. In particular, construct validation in practice is not based on such detailed theories, but on what Alexandrova and Haybron describe as "plausible-seeming correlations with relevant-seeming variables" (2016, p. 1103). Through Plausible-seeming correlations between happiness measures (h_r) and other variables, in turn, we can only take small steps in establishing the relationship *f*, because mere correlations cannot discriminate between different forms of *f*. For instance, a correlation between a

[9] The ambiguity of "validity" is acknowledged, not only by measurement theorists (Vessonen 2019, sec. 2.6), but also by researchers working on developing measurements in the field. In *Validity in Educational & Psychological Assessment,* Newton and Shaw state that "The term 'validity' is employed in so many different ways, in so many different contexts, that often it is entirely unclear what the speaker intends to convey" (2014: 2).

[10] It is not at all obvious whether these factors are relevant or irrelevant. Depending on the theory of happiness, temporary moods, personalities and even the weather could be seen as determinators of happiness, rather than noise in the measurement process.

happiness measure and optimism may be due to a causal relationship between happiness and optimism, but is also compatible with optimism affecting the relationship, *f*, between a happiness measures and happiness (with or without also affecting happiness itself).

As we shall argue in the following section, a deep problem for attempts to validate happiness measurement devices is that the function *f* also depends on the evaluation standards of the person who is being measured. The structure of this problem is similar to The Problem of Nomic Measurement.

3.3 Calibration and the problem of nomic measurement

The above characterization of calibration illustrates the following epistemic problem. The function *f* represents a relationship between an instrument reading which researchers can observe, and a knowledge claim about the quantity of interest. How can we establish this relationship unless we already know the quantity value? In practice, calibration of new measurement devices is often carried out by deploying the new device to measure objects where the quantity value is already determined. But how can this be done in situations when we have no pre-determined quantity values? At a general level, this problem is structurally identical to what Hasok Chang calls the problem of nomic measurement:

1. "We want to measure quantity X.
2. Quantity X is not directly observable, so we infer it from another quantity Y, which is directly observable. (…)
3. For this inference we need a law that expresses X as a function of Y, as follows: X = f(Y).
4. The form of this function f cannot be discovered or tested empirically, because that would involve knowing the values of both Y and X, and X is the unknown variable that we are trying to measure." (2004, p. 59)

We see the problem of nomic measurement as a problem of establishing whether a measured quantity Y tracks the sought quantity X.[11] As foundational as this problem appears, it does not undermine all successful measurement. We can be reasonably confident that the quantity we learn about by using thermometers and weight balances do track temperature and weight respectively.[12] The reason for this is that the problem of nomic measurement may be solved by invoking theoretically informed assumptions about quantity X, such as the idea that under normal conditions, water boil and freeze at fixed degrees. Or, in case of weights, that X is additive: placing two objects on a scale should therefore result in the additive degree of their individual weights.

[11] We discuss validation in a bit more detail in Sect. 3.3.

[12] The theoretic concepts of temperature (or weight) might by now be so close to the thing that thermometers (or scales) measure, that they should be considered the same thing.

More generally, a number of philosophers writing on the history of scientific measurement have argued that the validation of measures—by way of escaping the problem of nomic measurement—involves a logic of coherence (Chang 2004; van Fraassen 2008; Tal 2013, 2017b; Alexandrova and Haybron 2016; Hersch forthcoming). While scientific theories are based on measurement, and both the calibration and validation of measures is based on theoretical assumptions, scientific progress is constituted by a process of mutual correction: when our measured quantities and theories do not cohere, either one has to be adjusted.

While we are generally sympathetic to this view of resolving the problem, its success is not guaranteed in all fields of measurement. As Chang writes: the question of whether it does "is a contingent empirical question for each case" (see Chang 2004, pp. 226–227; see also Hersch forthcoming, p. 22). Whether it will be successful relies strongly on the presence of three factors that are contingent to a field of study:

1. *The available measurement procedures.* Some measurement procedures are more difficult to develop and perform than others. This can for example be due to the way the quantity is believed to manifest/interact with the environment, or due to the amount of control the researcher can have over the instruments used in the procedure. It is possible that some quantities are simply too difficult to measure. Notoriously, the presence, or even degree, of sentience in others, may be one for which no measurement procedure can exist (e.g. Block 2002).
2. *The theoretical stability of relationships between different quantities.* One reason for the successful calibration of thermometers is that measured outcomes in terms of temperature can be verified by how well they agree with other measured quantities, through, for example, our physical theories of thermodynamics. When our best theories state stable and universal relationships between quantities, we can use these relationships to calibrate our measurement devices. However, in some fields of sciences—perhaps especially in the social sciences—the relationships that our best available theories state between quantities are not sufficiently stable and universal for such agreement to be expected.
3. *The available theory.* Even if there are sufficiently easy procedures available for measuring a quantity that seem valid at face value, and they covary with other measured quantities, we need to explain these covariations and have some reason to be confident about the expected relationship between the quantities before such coherence result in confidence about the validity of measures. In other words, we need theory. Recently, Alexandrova and Haybron (2016), as well as Leah McClimans (2017), have argued that the lack of theory is a problem for the measurement of happiness.

As the details of these three factors (1, 2, and 3) are likely to differ for each specific quantity, each measurement has to find its own way around the problem of nomic measurement. The problem we raise in Sect. 5 does therefore not automatically generalize to any measurement procedure that involves self-reporting.

4 The importance of calibration for happiness research

As with any field using quantitative data, the epistemic strength of many results from happiness research crucially depends on how well calibrated the measurement devices are. Without adequately calibrated measurement devices, measurement data does not provide strong justification for inferences about comparisons or distributions of happiness. Claims about happiness comparisons are often central in the field. Take for instance the debate surrounding the Easterlin Paradox (see Easterlin 1995; 2013; Hagerty and Veenhoven 2003; Veenhoven and Hagerty 2006; Stevenson and Wolfers 2008; Clark, Frijters, and Shields 2008). The Easterlin Paradox is the (seemingly) paradoxical finding that while economic growth has been substantial throughout recent decades, and while richer countries tend to be happier, happiness among the population in developed countries over time has not increased much. Such a finding, however, can only be inferred from the readings we make of measurement data if the data collected in the 1970s represent happiness in (roughly) the same way as it does in later decades. In other words, we rely on the relationship between the readings and the outcome to justify these kinds of conclusions. Unless we have established a relationship that we have reasons to believe is invariant across time and contexts, we cannot infer from the data that happiness has remained stable over the years. Similar problems arise in many parts of the research field, such as in studies investigating whether parents are happier than other adults (Hansen 2012), or investigations of the happiness costs of being unemployed (Clark and Oswald 1994; Helliwell and Huang 2014). For instance, parents may be less happy (on average) than non-parents, but this can only be inferred if we assume that the happiness measures are well calibrated. Calibrating the devices used to measure happiness is thus of core importance to the field.

To clarify, calibration is not a process that is only successful once *all* uncertainty is removed. All measurement outcomes involve some uncertainty. In the above model, this uncertainty can be accounted for by expressing the outcome h_O as a range rather than as a single value. However, if in a particular situation, significant measurement error is expected, this undermines calibration. This detracts from the justification this measure provides for knowledge claims about its target concept—in our case, happiness. A well-calibrated measurement procedure that measures happiness will provide agreeing outcomes when used on people that are happy to the same degree.

Although calibration does not eliminate uncertainty, evaluating the uncertainty associated with a measurement is, on Tal's account, part of the calibration activity. Calibration is about establishing a relationship between instrument indicator and measurement outcome that holds over iterated measurements. Furthermore, even if a measurement procedure is well calibrated—such as modern-day temperature measurement devices—we may still get errors when using them. A particular device may function badly, a user could occasionally misread its indicators, or make other mistakes in using the measurement instrument. Measurement error is to be expected in almost any measurement practice (Angner 2013).

The upshot of this is that users of well calibrated happiness measurement procedures will be justified in ascribing happiness to the individuals they measure within a uncertainty range, a confidence interval. None of the big studies using happiness data which we cite in this paper do, however, present such uncertainty ranges.

Even if happiness measures are not calibrated in a way that justifies ascribing happiness to individuals within an uncertainty range, could happiness measurements warrant ascribing happiness on a group level? As one anonymous referee pointed out to us, in larger groups, individual differences can often be assumed to even out. We agree that claims about group averages are sometimes justified, even when corresponding claims about individual group members are not. An argument could be made that even if some individuals may change how they evaluate their happiness as they move from one context to another, researchers are safe to assume that these changes will even out on the group level (e.g. Haybron 2007, sec. 5.1).[13]

However, this argument only holds if the researchers are justified to assume that individual changes are evenly distributed, and it may not even hold then. As Bond and Lang (2019) has showed, group ascriptions of happiness are very sensitive to assumptions of individual distributions. If researchers are not justified in making such assumptions, it is therefore scant help to move from individual to group level. To make it even more difficult, there may well be structural factors that affect how individuals evaluate their happiness. For instance, in relation to the Easterlin paradox, people in a particular developed country may be evaluating their happiness differently forty years ago than they do today. Further calibration is thus important to offer stronger justification for claims about happiness comparisons of groups, across time and contexts, based on measurement. Still, progress in the calibration of happiness measures appears to be particularly difficult. A reason for this, we suggest, is the Two Unknowns Problem.

5 The two unknowns problem

In this section, we present a general problem for the calibration of happiness measurements that involve first-person reports. We focus on this type of procedures for two reasons. First, procedures for measuring happiness that involve first-person reports are today by far the most common. Second, given the privileged access that individuals have to their own happiness, it is difficult to see what a satisfactory measure of happiness would look like if it would not be based on first-person reports. However, in the next section, we discuss whether procedures that do not involve first-person reports can contribute to solving the problem.

[13] More precisely, Daniel Haybron suggests in relation to a specific type of measurement error "affective ignorance", or "AI": "… another point to bear in mind is that AI-related errors will often wash out over time, or over large samples. When you are assessing your own happiness, any errors are liable to lead you astray, so AI is much more worrisome. But for researchers studying large populations, many mistakes—e.g., a tendency to report being happier on sunny days—can be set aside as random "noise," as they will tend to cancel each other out." (2007, p. 412).

The Two Unknowns Problem is, in brief, as follows:

(1) The function *f* that describes the relation between a happiness measurement reading and its outcome depends on individuals' evaluation standards.
Therefore,
(2) establishing the function *f* requires having a verified value of either:

- subjects' degrees of happiness,
- or subjects' evaluation standards,

or being justified in assuming that one of them is invariant across the contexts where the measurement devices is intended to be used.
However:
(3) Degrees of happiness are not invariant across these contexts and cannot be verified independently of measuring a person's degree of happiness.
(4) Evaluation standards are not invariant across these contexts and cannot be verified independently of measuring a person's degree of happiness.
Therefore, (follows from 2 to 4).
(5) the relationship described by *f* cannot be established.

In other words, unless at least one of the above premises (1–4) are resisted, happiness measurements cannot be calibrated.

Admittedly, much work in this argument is done by what in premise two counts as "having a verified value of". We take it this is not a binary question, but one that comes in degrees, and may vary depending on how high the bar for having established the relationship is set. Consequently, the prospect of calibrating happiness measurements is a matter of degree. If we have low demands on what counts as establishing the relationship, then calibration is easier. We will return to this issue in Sect. 6.2.

We will go over the five claims of the argument in turn.

(1) First, as we described above, the function *f* from reading (h_r) to outcome (h_O) depends on other factors than these two. Claim (1) maintains that one such factor is the individual's standard of evaluation. To see this, notice that in happiness measures based on first-person reports individuals play a crucial role in the measurement process. As Gualtiero Piccinini observes:

> A subject generating first-person behaviors to fulfill the purposes of a scientific observer is a self-measuring instrument. When a subject generates first-person behaviors, she embodies not only (part of) the experimental materials but also (part of) the measurement apparatus. (2009, p. 11).

The observation that people who are evaluated embody a part of the measurement apparatus is significant for our purposes. This implies that if two people are asked to fill out the same questionnaire, with the same question on happiness, they can still not be said to have been subject to the same happiness measurement device. After all, a crucial part of the apparatus is the person herself. Concretely, when a person answers a happiness question, she evaluates her happiness.[14] This has an important implication for the measurement of happiness.

Evaluations depend on people's standards. For a happiness measure to work at all, the indication we get from this measure must in some way depend on how happy the person being measured is. Whether a person reports a 10, a 7, or a 3 on a happiness response scale should in part depend on how happy she is. However, the report also depends on other factors. Some of these factors can be called a person's "evaluation standard".[15] Our standards include such things as how we interpret the intervals on the response scale, and whether we rank our happiness relative to others, relative to our earlier life, or relative to some fixed standard (See Schwarz and Strack 1999 for a detailed model of how subjects respond to happiness questionnaires). The important thing to note here is that our personal standards will always partly determine how we report our happiness.

There is, as Fleaurbaey and Blanchet (2013) points out in their discussion of using self-report data, no intra-subjective standard readily available to us, which we can use to calibrate our responses. In other words, when people are evaluating happiness, they apply their own evaluation standards. Even if we assume for the present purpose that (A) the meaning of the term happiness, or its more specific cognates, such as pleasure, positive affect, enjoyment, and life-satisfaction, are perfectly clear, (B) people have perfect access to the mental states that constitute happiness, and (C), the semantic meaning of the scale is similarly understood by everyone, this does not imply that people will evaluate their happiness in the same way. After all, someone with higher standards of what it is like to experience a "very happy" state, will evaluate the same level of happiness differently from someone with lower standards, even if they both agree on the meanings of "very" and "happy".

(2) Second, if (1) is true, we have at least three variables in our measurement equation: (i) a person's degree of happiness, (ii) a person's happiness measurement indication (her happiness report), and (iii), a person's happiness evaluation standard. One of these can be verified from observations, namely, a person's happiness measurement indication (ii). However, it is doubtful whether the other two can.

To infer the value of one of the two unknown variables requires that the other is given or that the other remains invariant over a series of repeated tests. Thus,

[14] We are not claiming that she evaluates her life. It is common in the happiness literature to make a distinction between so called "evaluative happiness measures" and "affective happiness measures", where the former are taken to measures life-satisfaction and the latter are taken to measure e.g. pleasure. We are not suggesting that one of these is better than the other.

[15] It is good to note that this is a catchall term. We use the term "evaluation standard" to refer to all mental factors of the individual that may influence the conversion of a person's degree of happiness to a conscious self-rating, which can be invoked in a happiness measurement.

in situations where we have not established the value of at least one of these two variables and where we have no good reasons to believe that one of them remains invariant, estimating these two unknown variables remains fully underdetermined. In such situations, when we observe different reported happiness values, this might be due to a change in happiness, but it could also be due to a change in evaluation standard, or a change in both. From a person's reported happiness, we cannot infer what that person's evaluation standards are unless we are already given their degree of happiness, and vice versa, we cannot infer what a person's happiness is without having established their standard of evaluation.

(3) The third premise we take to be straightforward. A person's degree of happiness is generally not verifiable without measuring it. There might be cases where we as observers, based on how a person behaves or talks, could be reasonably confident in making rough judgments about how happy she is. In such cases, measurement might not be needed to verify imprecise ascriptions of happiness such as "she is not very happy" or "her happiness level is very high". However, to justify more fine-grained ascriptions, and in cases where we cannot make such judgments, we need to measure. After all, to find the value of a sought quantity is why we measure things in the first place. If we could verify how happy someone is without measuring her happiness, then we would have no use for happiness measurement.

The same goes for claims about the invariance of degrees of happiness. It may be that for certain individuals, degrees of happiness remain stable over time and across situations. However, this cannot be assumed a priori. Whether it is so is, in part, an empirical question.

(4) The fourth premise requires a stronger defense. Should we really treat a person's happiness evaluation standard as unknown and variant across contexts? First, note the standards that people apply to their evaluations cannot be inferred from their responses, and they are also themselves determined by a person's mental states. In this, evaluation standards are similar to happiness-constituting states themselves. What researchers can observe is people's expressions of self-evaluations, but researchers can neither observe the conscious mental states that (partly) constitute happiness, nor the evaluation standards that people apply when evaluating them. So, like degrees of happiness, happiness evaluation standards cannot be directly inferred from happiness responses.

Is it plausible that a person's standards for evaluations are stable across time and people, and thus could be considered invariant? This would be highly unlikely. To see this, consider some real happiness scales used in empirical measurement. The numbered scales from 0 or 1 to 10, in which the beginning and end points are considered the "least satisfied" and "most satisfied" (Helliwell 2003), or as “extremely unhappy” to “extremely happy” (Layard 2010). In case of Likert scales, there are more such points, as every point of the scale has such an interpretation—“not too happy”, “very happy”, “not happy, and not unhappy”, etc. If the evaluation standards are invariant, the points of “extremely happy” and “extremely unhappy” must indicate the same degrees of happiness for everyone. And, the point of “extremely satisfied” and “extremely unsatisfied” must indicate the same degree of satisfaction. However, these points rely heavily on people’s imaginative abilities. Moreover, how much happiness we can imagine is quite likely related to the level, and the variety, of happiness we have experienced in our life. Consequently, our conception of “extremely happy” and “most satisfied” are person-dependent. These points may change in at least two ways.

First, how high we evaluate our happiness may depend on how our happiness changes over time. A trivial example to indicate that our calibration points are not fixed points is to consider a person whose life becomes increasingly happier. At first, this person may consider her happiness to be 8, but as her life becomes happier, she will now say that her happiness is 9. After a similar increase, she is now forced to say that her happiness is 10, but as her life becomes even happier, there is no way to express this on the scale. The only way to do so would be to fill in 10, but this 10 now has a different meaning than the previous 10. Psychometricians call this phenomenon “recalibration” (Blome and Augustin 2016, but also Ubel, Peeters, and Smith 2010). While the scale is clearly bounded, it is not obvious whether happiness itself is. To the contrary, while we may think that happiness is not unbounded, it seems hard to imagine that there exists a maximum degree of happiness.

Second, there is no reason to assume that the shifting of standards is limited to fringe cases at either extremes of the scale. The same mechanism that adjusts standards at the fringe, can similarly apply to cases in the middle of the scale (see Haybron 2007). It seems plausible that this would happen. A widespread phenomenon that may illustrate this is the phenomenon of hedonic adaptation. It has been observed that as the conditions of our lives improve, the degree of self-reported happiness tend to go up, but then decreases afterwards, and vice versa for negative events, even though people do not completely adapt (Layard 2005; Dolan, Peasgood, and White 2008; Luhmann et al. 2012). Either this indicates that evaluation standards are not invariant, or happiness is not tied to life conditions in the way we tend to think. There seems to be no reason to disregard the former.

The phenomenon of hedonic adaptation seems at least in part be due to people’s adjustment of aspirational standard (see, for example, van der Deijl 2017a for an argument to this effect). If that is correct, a person’s happiness evaluation standard is not invariant across individuals: people at the same degrees of happiness will evaluate their happiness differently. The effect of differing evaluation standards cannot be inferred from the readings of happiness measurement. The standards are determined by a person’s mental states, and are subject to changes within individuals, which seem likely to be triggered by amongst other things changing degrees in happiness—that is, either significant increases, or significant decreases.

Besides hedonic adaptation, there are further indications that peoples' evaluation standards differ, and may change, depending on circumstances. Scholars in the field have suggested that contextual factors such as personalities (Abbott et al. 2008), age (Costa and Siegler 2003) and culture (Tov and Diener 2009) can impact reported happiness, which in turn could be explained by variant evaluation standards.

(4) From the above premises it follows that the function *f* that describes the relation between happiness measurement reading and measurement outcome in terms of an individual's degree of happiness cannot be established. As claims 1–4 show, for every happiness measurement, the function *f* contains two unknowns: a person's happiness and a person's standard of happiness evaluation. Neither of these two unknowns can be treated as invariant (claim 3 and 4). Jointly, these four claims thus entail that inferences from happiness measurement indications to claims about degrees of happiness are always underdetermined by the subjects' evaluation standards. The relationship between reading and outcome therefore cannot be reliably established, while reliably establishing such a relationship is exactly what calibration aims to achieve.

To summarize: because happiness is a mental concept, the only direct access we have to happiness is through ourselves. For this reason, researchers typically use first-person reports to measure happiness. However, by doing so, the individual becomes part of the measurement apparatus, and their evaluation standards determine the relation between what they report and their happiness. There are no good reasons to believe evaluation standards are invariant over contexts or across people, nor are they externally observable. Therefore, to find out how the impact of evaluation standards vary requires already having determined a person's degree of happiness. Since this cannot be done, the relationship between reading and outcome in happiness measurement cannot be established. Measurement readings that indicate a difference in happiness between or within lives are thus underdetermined. The differences may result from differences in happiness, differences in evaluation standards, or a combination of both.

5.1 Implications for calibration

If the Two Unknowns Problem is sound, this has significant implications for happiness research. The upshot of the Two Unknowns Problem seems to be that calibrating happiness measures is bound to be a frustrating exercise. To calibrate happiness measures, researchers need to find some way to get around at least one of the premises.

As we have pointed out before, calibration is an epistemically central aspect of measurement and crucial for the continued improvement of happiness research. The two unknown problems is a challenge for anyone who attempts to justify claims about how happy a person is based on happiness measurements.

6 Objections and ways around the problem

Is there any way around the Two Unknowns Problem for happiness measurement? In this section, we discuss two routes for escaping the problem. The first route we discuss is to deny claim 4—which states that evaluation standards are not knowable nor invariant across contexts—cannot be established. This route is, we suggest, unconvincing. The second route we discuss is to deny claim 3—which states that there are no measurement independent ways of knowing how happy a person is. This route is more promising, and we believe it illustrates the progress the science of happiness needs to make to have a shot at calibrating measures of happiness successfully.

6.1 Absence of evidence objection

One could object to the speculative nature of premise 4 in the Two Unknowns Problem. We argued that there are reasons to believe that people's evaluation standards are not invariant across different contexts, and that there is little reason to believe that they *are* invariant. However, someone may object that because this suggestion itself is speculation—the argument relies on an untestable hypothesis—the claim should not be accepted. A version of this critique can be found in John Harsanyi's defense of interpersonal comparisons of utility, when he writes that:

> If two objects or human beings show similar behavior in all their relevant aspects open to observation, the assumption of some unobservable hidden difference between them must be regarded as a completely gratuitous hypothesis and one contrary to sound scientific method. (Harsanyi 1980, p. 317).

The difference in evaluation standards, in the context of first-person happiness assessments, qualifies as "some unobservable hidden difference". Harsanyi argues that the denial of this principle would make it permissible to ascribe differences to any unobservable entities, even the phenomenology of other people, and consequently, could lead to solipsism.

It seems plausible that we should not ascribe differences in unobservable entities without having any specific reason for it. However, what is needed to calibrate a measure of happiness is stronger than the absence of the assumption that people's evaluation standards are different. The aim of calibration is to reliably establish a specific relationship between a reading and a measurement outcome. In absence of a good reason to believe that such a relationship can in fact be established, we cannot say that we have reliably *established* the relation. So, while we have presented an argument for believing that it is not, our argument only relies on the weaker claim that there is no good argument for believing that such a stable relationship exists. Premise 4, can thus be rephrased as follows, and the argument will still go through:

(4*) We do not know whether evaluation standards are invariant across these contexts or whether they can be verified independently of measuring a person's degree of happiness.

So, even if it is our intuitive guess that similar readings are indications of similar degrees of happiness even across contexts, this guess is not sufficient. The possibility of calibrating happiness measurements in this way depends on there being positive reasons to believe evaluation standards are invariant across contexts.

6.2 Behavioral and neural measures of happiness

A more promising escape route is to question premise 3. As we stated in the beginning of section IV, the Two Unknowns Problem builds on the assumption that measures of happiness are first-person measures. Our justification for this has been that happiness researchers almost exclusively use happiness measures based on first-person reports, and that, given the internal nature of happiness, it is difficult to imagine what would constitute an adequate alternative. However, while it may be true that the best measurements of happiness require using first-person information, there are reasons to believe that degrees of happiness are related to behavioral and neural signs. If we can measure these signs, this information could thus be used to verify claims about happiness independently of first-person measurements. In other words, behavioral and neural quantities may on their own justify claims about how happy people are. If they do, this could resolve the Two Unknowns Problem by invalidating premise 3. We could then have independent ways of verifying how happy someone is.

Although self-ratings may seem like the most intuitively valid and direct way of measuring happiness, there is nothing in principle problematic with measurement procedures that rely on more indirect manifestations of the quantity being measured. For example, while temperature and distance between tree rings at first sight seem very distinct, distances between tree rings may still be used to measure historic temperatures. Could the same not be true for happiness? For example, could not data on smiles (Ekman et al. 1990) or fMRI scans be related to happiness in ways that can be invoked to justify inferences from these quantities to claims about happiness? Studies of such relations have indeed been used in discussions about the validity of different happiness constructs (Layard 2005; Diener et al. 2009a, b; see also Alexandrova and Haybron 2016).

There is no principled reason not to deny premise 3. To get around it, we just need to find other ways to verify how happy someone is. The problem here is not principled, but specific for the current state of happiness science. If we would have a theory that tells us how degrees of happiness relate to other quantities, for which we in turn already have reasonably well calibrated measurement devices, then measurements of these other quantities could be used to calibrate measures of happiness. If so, researchers could perhaps estimate the differences in evaluation standards that individuals apply and thereby solve one of the unknowns. However, no general such theory is (yet) available in case of happiness.[16] Researchers may have

[16] Leah McClimans (2017) discusses this problem of lack of theory in the case of measuring quality of life. Her arguments apply also in the case of happiness measurement.

well-considered judgments about how happiness relates to other quantities, such as to the amount we smile, or risks of suicide. However, to calibrate measures of happiness, based on such other more easily measurable quantities, we need to have a theory that tell us how smiling, or suicides, relate to happiness in different contexts.[17] If such theories would be established, they could be used to calibrate measures of happiness. However, such theories are currently not available. As long as they are not, behavioral and neural measures of happiness may provide us with some confidence that measures of happiness indeed measure happiness, as previously discussed regarding construct validation, but not that the numbers we derive from happiness measures correspond to more detailed quantities of happiness. In other words, they are only a first step towards calibrating measures of happiness.

This leaves open the possibility that as theory develops, relevant knowledge claims become available that can be used to calibrate measures of happiness. If, as we argued, the calibration of happiness measures requires getting around the Two Unknowns Problem, this is the most promising route.

However, there are some reasons to be pessimistic about the efficacy of this strategy. For one, the way mental phenomena are expressed outwards can be expected to depend on factors that are highly contextual. For example, the amount that we smile is not only dependent on how happy we feel, but also on what we find humorous, or the meaning of smiling in social interactions. For example, in a study of smiles among gold medalist at award ceremony at the Olympic games, Fernández-Dols and Ruiz-Belda (1995) find that they smile very little at moments when we expect them to feel happy. They suggest that this may be due to the fact that smiling and happiness goes together only in specific social situations, or can be suppressed in solemn contextual context, such as the singing of the national anthem. And while this only illustrates that the relationship between smiling and happiness is complex in a very specific social context, it seems plausible that in other social context, many other social factors also affect the amount we smile. If so, we would need a highly extensive and complex quantitative theoretical framework that not merely relies on the measurement of smiling quantity or duration, but also on social factors such as cultural context, social context, etc. This would be a Gargantuan task, if at all feasible.

Neuroimaging may provide an alternative route around premise 3 (see Tanzer and Weyandt 2020 for a recent meta-analysis of neuroimaging studies on happiness). One way to resolve the Two Unknowns problem opens if we can find out how specific brain states relate to specific levels of happiness. While studies so far indicate that there are many neural correlates to happiness, they also indicate that the

[17] In fact, Ekman et al. suggest that duration of smiles are more predictive of positive emotions than the number of smiles: “The duration of facial actions is probably a more accurate index of emotion because duration is sensitive to very long expressions, which may be given little weight if only frequency is considered. However, frequency is less costly to obtain because the precise onset and offset of each action is not required as it is to determine duration, and therefore most investigators have reported frequency data.” (Ekman et al. 1990, p. 346).

relationships are complex, dependent on the specific measure of happiness, and still involve much uncertainty (Tanzer and Weyandt 2020). Currently, neither theory nor empirical results are of the right level of specificity to justify claims about happiness independent of self-reported measures, nor could we find any recent studies that have attempted to arrive at such measures.[18] Would this field develop, however, it would provide a fruitful way out of the Two Unknowns problem.

It is difficult to determine the relationship between self-reported data and happiness, and in order to use behavioral or neural measures to calibrate happiness measurement devices, researchers would need to solve Chang's problem of nomic measurement once again. They would need some way to establish the functional relationship between degrees of happiness and the degrees of these neural and behavioral quantities. While challenging, such steps are necessary to further calibrate happiness measures.

6.3 Scope and proving too much

A final objection is that our argument overgeneralizes and thereby proves too much. The Two Unknowns Problem may appear to affect almost all, or most, mental measures that are based on self-evaluations of mental states (e.g. pain measurement, see Noble et al. 2005 for a historical overview; or pain measurement, see Fried 2017 for a recent overview), while these are used with success throughout the psychological and medical sciences. If our argument would imply that all such measures will struggle to satisfactory establish knowledge claims, the burden of proof on us to defend the argument would be higher than we can offer. A more plausible conclusion may then be that our argument is incorrect, or at the least that the conclusion we draw is exaggerated.[19]

We are not claiming that the Two Unknown Problem generalizes to many other areas. As this section and the previous section have illustrated, the extent to which the Two Unknowns problem hampers calibration of measures that are based on self-evaluations of an individual's mental state depends on the plausibility of non-stable standards, the available theory, and available alternative measures. Whether the Two Unknowns problem is a problem thus depends on multiple issues specific to the field, including which the target concept is and what kinds of knowledge claims the measurement is supposed to justify. The scope of the problem therefore cannot be settled without a detailed investigation of these issues for each measured mental quantity.

[18] As researcher describe the state of the field: "Although subjective well-being has drawn a lot of attention from researchers (…), the precise neural correlates underlying this construct are still largely unknown."(Kong et al. 2015, p. 136) and: "There remains uncertainty as to how happiness as a subjective experience is processed in the brain." (Tanzer and Weyandt 2020, p. 2694).

[19] We thank two anonymous referees for pressing this issue.

7 Conclusion

We have argued that calibration matters to happiness measures, but that the calibration of happiness is made particularly difficult by the Two Unknowns Problem. We have discussed ways around the Two Unknowns Problem, but argued that they, in turn, face their own difficulties.

A first conclusion of our article is that more research is needed on calibrating happiness measurement. The calibration of happiness measures is crucial for the project of studying happiness. Without adequately calibrated measures, inferences based on measured quantities are not well justified, and this undermines the theoretical basis for the quantitative study of happiness. Significantly, without adequately calibrated measurement devices, we cannot, based on measurement, justify comparative claims about how happy persons are, neither across, nor within, lives. Such comparative claims are crucial in many central discussions in the literature, for example the debate surrounding the Easterlin Paradox.

On the other hand, our conclusion is modest. We have, for one, not argued that happiness measures are not *about* happiness, a view that some skeptics have maintained (e.g. Johns and Ormerod 2007). We are not claiming that readings from happiness measures are merely noise, or that they bear no epistemic relation at all to happiness.

Nevertheless, our conclusion poses a problem for the optimistic view that happiness science, in its current state, can establish claims about degrees of happiness. Wodak (2019) has recently suggested that we should represent measures of wellbeing, including measures of happiness, as letters with an ordinal ranking, rather than numbers. If our argument is correct, his solution does not help, as shifts in ordinal ranking may result from shifts in evaluative standards, which we have argued cannot be assumed to be stable.

What then, should be done? If our argument is correct, behavioral or neural measures of happiness can play a crucial role in the calibration of happiness measures. However, at present, our theories nor our validation efforts are fine-tuned enough for this purpose. While the calibration of happiness measures promises to be a challenging endeavor, these steps are crucial for the justification of many scientific claims about degrees of happiness.

Acknowledgments We would like to thank two anonymous referees for their helpful comments, which greatly helped to improve the papers. We also thank Erik Angner for helpful comments as well as the attendees of the workshop on Philosophical Issues in the Science and Policy of Well-being, organized in Copenhagen on December 5 and 6, 2019, in particular: Anna Alexandrova, Valerie Tiberius, Thor Grünbaum, and Victor Lange Nielsen.

References

Abbott, R. A., Croudace, T. J., Ploubidis, G. B., Kuh, D., Richards, M., & Huppert, F. A. (2008). The relationship between early personality and midlife psychological well-being: Evidence from a UK Birth Cohort Study. *Social Psychiatry and Psychiatric Epidemiology, 43*(9), 679. https://doi.org/10.1007/s00127-008-0355-8.

Abdel-Khalek, A. M. (2006). Measuring happiness with a single-item scale. *Social Behavior and Personality: An International Journal, 34*(2), 139–150. https://doi.org/10.2224/sbp.2006.34.2.139.

Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science, 83*(5), 1098–1109.

Angner, E. (2013). Is it possible to measure happiness? *European Journal for Philosophy of Science, 3*(2), 221–240.

Annas, J. (2004). Happiness as achievement. *Daedalus, 133*(2), 44–51.

Bishop, M. A. (2015). *The good life: Unifying the philosophy and psychology of well-being*. Oxford: Oxford University Press.

Block, N. (2002). The harder problem of consciousness. *The Journal of Philosophy, 99*(8), 391–425.

Blome, C., & Augustin, M. (2016). *Measuring change in subjective wellbeing: methods to quantify recall bias and recalibration response shift*. Working Paper 2016/12. HCHE Research Paper. https://www.econstor.eu/handle/10419/145973.

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

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In *The concept of validity: Revisions, new directions and applications, Oct, 2008*. IAP Information Age Publishing.

Chang, H. (2004). *Inventing temperature: measurement and scientific progress*. Oxford: Oxford University Press.

Clark, A. E., Frijters, P., & Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles. *Journal of Economic Literature, 46*(1), 95–144.

Clark, A. E., & Oswald, A. J. (1994). Unhappiness and unemployment. *The Economic Journal, 104*(424), 648–659.

Connolly, M. (2013). Some like it mild and not too wet: The influence of weather on subjective well-being. *Journal of Happiness Studies, 14*(2), 457–473. https://doi.org/10.1007/s10902-012-9338-2.

Costa, P., & Siegler, I. C. (2003). *Recent advances in psychology and aging*. Amsterdam: Elsevier.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Cummins, R. A. (2013). Cultural response bias. *The Encyclopedia of Cross-Cultural Psychology, 24,* 330–333.

Diener, Ed. (1994). Assessing subjective well-being: progress and opportunities. *Social Indicators Research, 31*(2), 103–157.

Diener, Ed., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*(1), 71–75.

Diener, Ed., Inglehart, R., & Tay, L. (2013). Theory and validity of life satisfaction scales. *Social Indicators Research, 112*(3), 497–527. https://doi.org/10.1007/s11205-012-0076-y.

Diener, Ed., Lucas, R., Schimmack, U., & Helliwell, J. F. (2009a). *Well-being for public policy*. New York: Oxford University Press.

Diener, Ed., Oishi, S., & Lucas, R. (2009b). Subjective well-being: The science of happiness and life satisfaction. In S. J. Lopez & C. R. Snyder (Eds.), *The Oxford handbook of positive psychology* (2nd ed.). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195187243.001.0001/oxfordhb-9780195187243-e-017.

Dolan, P., & Peasgood, T. (2008). Measuring well-being for public policy: Preferences or experiences? *The Journal of Legal Studies, 37*(S2), S5–S31.

Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior & Organization, 27*(1), 35–47.

Easterlin, R. A. (2013). Happiness, growth, and public policy†. *Economic Inquiry, 51*(1), 1–15. https://doi.org/10.1111/j.1465-7295.2012.00505.x.

Easterlin, R. A., McVey, L. A., Switek, M., Sawangfa, O., & Zweig, J. S. (2010). The happiness-income paradox revisited. *Proceedings of the National Academy of Sciences, 107*(52), 22463–22468.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. *Journal of Personality and Social Psychology, 58*(2), 342.

Fernández-Dols, J.-M., & Ruiz-Belda, M.-A. (1995). Are smiles a sign of happiness? Gold medal winners at the olympic games. *Journal of Personality and Social Psychology, 69*(6), 1113.

Fletcher, G. (2013). A fresh start for the objective-list theory of well-being. *Utilitas, 25*(02), 206–220.

Fleurbaey, M., & Blanchet, D. (2013). *Beyond GDP: Measuring welfare and assessing sustainability*. New York: Oxford University Press.

Fraassen, B. C. V. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders, 208,* 191–197.

Hagerty, M. R., & Veenhoven, R. (2003). Wealth and happiness revisited-growing national income does go with greater happiness. *Social Indicators Research, 64*(1), 1–27.

Hansen, T. (2012). Parenthood and happiness: A review of folk theories versus empirical evidence. *Social Indicators Research, 108*(1), 29–64.

Harsanyi, J. C. (1980). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. In *Essays on ethics, social behavior, and scientific explanation* (pp. 6–23). Springer. https://doi.org/10.1007/978-94-010-9327-9_2.

Haybron, D. M. (2000). Two philosophical problems in the study of happiness. *Journal of Happiness Studies, 1*(2), 207–225.

Haybron, D. M. (2005). On being happy or unhappy. *Philosophy and Phenomenological Research, 71*(2), 287–317.

Haybron, D. M. (2007). Do we know how happy we are? On some limits of affective introspection and recall. *Nous, 41*(3), 394–428.

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

Haybron, D. M., & Tiberius, V. (2015). Well-being policy: What standard of well-being? *Journal of the American Philosophical Association, 1*(4), 712–733.

Helliwell, J. F., & Huang, H. (2014). New measures of the costs of unemployment: Evidence from the subjective well-being of 3.3 Million Americans. *Economic Inquiry, 52*(4), 1485–1502. https://doi.org/10.1111/ecin.12093.

Helliwell, J. F. (2003). How's life? Combining individual and national variables to explain subjective well-being. *Economic Modelling*, *20*(2), 331–360.

Helliwell, J. F., Layard, R., & Sachs, J. D. (2019). *World happiness report 2019*. Available at: https://worldhappiness.report/ed/2019/. Accessed 25 July 2019.

Hempel, C. G. (1952). *Fundamentals of concept formation in empirical science. International Encyclopedia of Unified Science*. Chicago: University of Chicago Press.

Hersch, G. (forthcoming). Well-being coherentism. *The British Journal for the Philosophy of Science*. http://philsci-archive.pitt.edu/17185/.

Howard, G. S. (1994). Why do people say nasty things about self-reports? *Journal of Organizational Behavior, 15*(5), 399–404.

Johns, H., & Ormerod, P. (2007. Happiness, economics and public policy. *Institute of Economic Affairs, Research Monograph* 62.

Kahneman, D., & Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *The Journal of Economic Perspectives, 20*(1), 3–24.

Kaminitz, S. C. (2018). Happiness studies and the problem of interpersonal comparisons of satisfaction: Two histories, three approaches. *Journal of Happiness Studies, 19*(2), 423–442. https://doi.org/10.1007/s10902-016-9829-7.

Kong, F., Siyuan, Hu., Wang, Xu., Song, Y., & Liu, J. (2015). Neural correlates of the happy life: The amplitude of spontaneous low frequency fluctuations predicts subjective well-being. *NeuroImage, 107,* 136–145. https://doi.org/10.1016/j.neuroimage.2014.11.033.

Kozma, A., Stone, S., & Stones, M. J. (2000). Stability in Components and predictors of subjective well-being (SWB): Implications for SWB structure. In Ed. Diener & D. R. Rahtz (Eds.), *Advances* *in quality of life theory and research. Social indicators research series* (pp. 13–30). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-4291-5_2.

Layard, R. (2005). *Happiness: Lessons form a new science*. London: Allen Lane.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5–55.

MacKerron, G. (2012). Happiness economics from 35,000 feet. *Journal of Economic Surveys, 26*(4), 705–735.

McClimans, L. (2017). Measurement in medicine and beyond: Quality of life, blood pressure and time. In A. Nordmann & N. Mößner (Eds.), *Reasoning in measurement* (pp. 133–146). New York: Routledge.

Mill, J. S. (1871). Utilitarianism. In R. Crisp (Ed.), *Oxford philosophical texts*. London: Oxford University Press.

Mößner, N., Nordmann, A., & Nordmann, A. (2017). Reasoning in measurement. *Routledge*. https://doi.org/10.4324/9781781448717.

Noble, B., Clark, D., Meldrum, M., ten Have, H., Seymour, J., Winslow, M., & Paz, S. (2005). The measurement of pain, 1945–2000. *Journal of Pain and Symptom Management, 29*(1), 14–21. https://doi.org/10.1016/j.jpainsymman.2004.08.007.

Nussbaum, M. C. (2008). Who is the happy warrior? Philosophy poses questions to psychology. *The Journal of Legal Studies, 37*(S2), S81–S113.

Raibley, J. R. (2012). Happiness is not well-being. *Journal of Happiness Studies, 13*(6), 1105–1129.

Rossi, M. (2018). Happiness, pleasures, and emotions. *Philosophical Psychology, 31,* 898–919.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*(3), 513–523. https://doi.org/10.1037/0022-3514.45.3.513.

Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. In *Well-being: The foundations of hedonic psychology* (Vol. 7, pp. 61–84). New York: Russell Sage Foundation. http://profron.net/happiness/files/readings/Schwarz-n-Stack_ReportsOfSubjectiveWellBeing.pdf.

Schwarz, N., Strack, F., Kommer, D., & Wagner, D. (1987). Soccer, rooms, and the quality of your life: Mood effects on judgments of satisfaction with life in general and with specific domains. *European Journal of Social Psychology, 17*(1), 69–79. https://doi.org/10.1002/ejsp.2420170107.

Steel, P., & Ones, D. S. (2002). Personality and happiness: A national-level analysis. *Journal of Personality and Social Psychology, 83*(3), 767–781. https://doi.org/10.1037/0022-3514.83.3.767.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*(1), 138.

Stevenson, B., & Wolfers, J. (2008). Economic growth and subjective well-being: Reassessing the Easterlin paradox. *Brookings Papers on Economic Activity, 1,* 1–87.

Sugden, R., & Teng, J. C.-Y. (2016). Is happiness a matter for governments? In S. Bartolini, E. Bilancini, L. Bruni, & P. L. Porta (Eds.), *Policies for happiness* (pp. 36–57). Oxford: Oxford University Press.

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

Tal, E. (2013). Old and new problems in philosophy of measurement. *Philosophy Compass, 8*(12), 1159–1173. https://doi.org/10.1111/phc3.12089.

Tal, E. (2015). *Measurement in science*, June. https://seop.illc.uva.nl/entries/measurement-science/.

Tal, E. (2017a). A model-based epistemology of measurement. In A. Nordmann & N. Mößner (Eds.), *Reasoning in measurement* (pp. 233–253). New York: Routledge.

Tal, E. (2017b). Calibration: Modelling the measurement process. *Studies in History and Philosophy of Science Part A, 65,* 33–45.

Tännsjö, T. (2007). Narrow hedonism. *Journal of Happiness Studies, 8*(1), 79–98.

Tanzer, J. R., & Weyandt, L. (2020). Imaging happiness: Meta analysis and review. *Journal of Happiness Studies, 21*(7), 2693–2734. https://doi.org/10.1007/s10902-019-00195-7.

Tov, W., & Diener, Ed. (2009). Culture and subjective well-being. In Ed. Diener (Ed.), *Culture and well-Being: The collected works of Ed Diener. Social indicators research series* (pp. 9–41). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-2352-0_2.

Ubel, P. A., Peeters, Y., & Smith, D. (2010). Abandoning the language of 'response shift': A plea for conceptual clarity in distinguishing scale recalibration from true changes in quality of life. *Quality* *of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 19*(4), 465–471. https://doi.org/10.1007/s11136-010-9592-x.

van der Deijl, W. J. A. (2017a). Which problem of adaptation? *Utilitas, 29*(4), 474–492.

van der Deijl, W. J. A. (2017b). Are measures of well-being philosophically adequate? *Philosophy of the Social Sciences, 47*(3), 209–234. https://doi.org/10.1177/0048393116683249.

van der Rijt, J.-W. (2013). Public policy and the conditional value of happiness. *Economics & Philosophy, 29*(3), 381–408.

Veenhoven, R., & Hagerty, M. (2006). Rising happiness in nations 1946–2004: A reply to Easterlin. *Social Indicators Research, 79*(3), 421–436.

Vessonen, E. S. M. (2019). *Representing and constructing. Psychometrics from the perspectives of measurement theory and concept formation*. Thesis, University of Cambridge. https://doi.org/10.17863/CAM.36695.

Wodak, D. (2019). What if well-being measurements are non-linear? *Australasian Journal of Philosophy, 97*(1), 29–45.