Erkenntnis
https://doi.org/10.1007/s10670-019-00156-9

ORIGINAL RESEARCH

Developing Dark Pessimism Towards the Justificatory Role of Introspective Reports

Elizabeth Irvine[1]

Received: 10 May 2018 / Accepted: 22 July 2019

Abstract
This paper argues for a position of 'dark pessimism' towards introspective reports playing a strong justificatory role in consciousness science, based on the application of frameworks and concepts of measurement. I first show that treating introspective reports as measurements fits well within current discussions of the reliability of introspection, and argue that introspective reports must satisfy at least a minimal definition of measurement in order to play a justificatory role in consciousness science. I then show how treating introspective reports as measurements makes it possible to identify the foundational methodological problems that underlie much of the current philosophical and scientific debate about the status of introspective evidence in studying consciousness. I argue that these problems prevent introspective reports from playing a strong justificatory role and resolving long-standing debates in consciousness science, both in contemporary work and in the future.

1 Introduction

Consciousness science does not currently make much use of introspective reports in a strong justificatory role; in particular they are not used to inform or provide empirical confirmation for theories of consciousness. For various researchers, both empirical and philosophical, this is a poor state of affairs, and they argue that introspective evidence can provide necessary, unique, and potentially revolutionary data, poised to resolve long-standing debates in consciousness science concerning the boundaries and contents of subjective experience (Hurlburt 2011; Jack and Roepstorff 2002; Kriegel 2013; Olivares et al. 2015; Overgaard et al. 2006a, b; Petitmengin 2006). This is set against a background where both supporters and detractors of introspection are aware of the problems associated with gathering veridical introspective reports. Supporters of introspection claim that these problems must be, and can be, overcome, while detractors are more sceptical.

✉ Elizabeth Irvine
irvinee@cardiff.ac.uk

[1] Cardiff University, Cardiff, UK

Published online: 08 August 2019

This paper argues for a position of 'dark pessimism'[1] towards introspective reports playing a strong justificatory role in consciousness science, based on the application of frameworks and concepts of measurement. Introspective reports are not usually described as 'measurements', but these reports are after all attempts to veridically and publically record the properties of experience, sometimes using pre-set rating scales or response categories. Discussions of introspective evidence often focus on questions of variability, validity, and accuracy, all of which are basic features of any measure. Discussions of introspective reports also often mention training or calibration of introspective participants, validation of methods, and so on. Linking introspective reports with frameworks of measurement is therefore not totally alien.

At the same time however, the use of this kind of vocabulary is distanced from any rigorous evaluation of introspective reports as measurements. Attempts at calibration or validation are often local, as are resolutions to problematic instances of introspective variability or inaccuracy. As I argue below, the idea of introspection as measurement can in fact be used to illustrate and identify the deep methodological problems that underlie much of the current debate about introspective evidence in consciousness science. These problems most obviously apply to the current state of affairs with respect to introspection, but I further argue that these problems prevent introspective reports from ever playing a strong justificatory and decisive role in consciousness science.

The argument rests on evaluating the methodologies available to researchers to validate introspective procedures and reports. Compared to other uses of verbal reports in cognitive science, introspective reports about the nature of consciousness raise specific methodological challenges, related in particular to how unknown, how unpredictable, and how complex and sensitive the generation of introspective reports is, compared to other sources of evidence about consciousness. Analysing the steps required to use either 'bottom-up' bootstrapping methods or comparative techniques for validating introspective procedures or evidence shows that there are a number of reasons why introspective evidence cannot carry significant justificatory weight. Instead of introspective reports being able to resolve long-standing debates in consciousness science, the methodology presented below suggests that these debates would have to be largely resolved before introspective reports could be appropriately validated. By this point though, introspective reports would no longer be able to provide an independent source of justification for theoretical claims about the nature of consciousness.

Below, Sect. 2 briefly identifies what introspective evidence is supposed to provide evidence about, and reviews recent arguments in favour of a strong justificatory role of introspective reports in consciousness science. This includes a discussion of the justificatory role of verbal reports in other areas of cognitive science, where I show that the methods used there are not transferrable to the case of consciousness. Section 3 motivates treating introspective reports as measurements and reviews

[1] A term used by Schwitzgebel in his (2011). The pessimism argued for here is in some ways darker than Schwitzgebel's, and is motivated in a radically different way.

existing discussions of introspection in terms of measurement. In Sect. 4 a 'bottom-up' process for developing measurement procedures is outlined, and problems in applying this to introspection are identified. In Sect. 5 the possibility of cross-validating introspective procedures is outlined and evaluated. Section 6 considers the cross-validation of sets of introspective evidence, and the evidentiary and justificatory status of introspective reports that result within this framework. An objection to the scope of the argument is considered in Sect. 7, and Sect. 8 concludes.

2 The Role of Introspective Reports

Within consciousness science there is currently a lot of interest in the role of subjective data, and introspective reports in particular, in understanding the nature of experience. For the purposes of this paper, this mostly concerns identifying where the boundary between conscious (experienced) and unconscious (not experienced) perception is, and what its contents are (e.g. is subjective experience 'rich' or sparse' in detail, Block 2007; Kouider et al. 2010). Many, though not all, of the positive proposals for introspective reports playing a strong justificatory role for claims about these features of experience have come from philosophers and cognitive scientists influenced by the phenomenological tradition within philosophy (e.g. Gallagher and Sørensen 2006; Kriegel 2013; Lutz and Thompson 2003; Petitmengin 2006). However, there is still a general scepticism from cognitive scientists about the possible evidentiary roles of introspective reports (e.g. Schooler and Schreiber 2004), and there are philosophical sceptics too (Bayne and Spener 2010; Dennett 2003, 2007; Schwitzgebel 2008, 2012).

The basis of the positive proposals is that introspective evidence is simply a necessary form of evidence for developing theories of consciousness (subjective experience), so methods with appropriate safeguards must be developed to try to use it. There is an accompanying optimism about the ability of such methods to uncover at least some veridical introspective reports, and there are concrete proposals for what these methods might look like.

A clear statement of this approach comes from Kriegel (2013), who puts forward an argument for the 'epistemic indispensability' of introspective reports. He argues that it would be strange to conduct a study of zebras based entirely on indirect observations of zebras (zebra tracks, droppings, etc.) if it was possible to directly observe them. Similarly, given that we have direct (observational) access to our own experiences via introspection, it would be bizarre to develop theories of consciousness without making use of this evidence. A further claim is that since introspective reports are likely to be more veridical than not, in at least some circumstances (under normal conditions, from normal subjects, not aimed at 'elusive phenomenologies'), we should be making use of introspective reports.

Kriegel further specifies that the role that introspective reports should play is a *justificatory* role, rather than a role in discovery. Researchers regularly use introspection, both applied to themselves, and introspective reports from subjects, to develop hypotheses and non-introspective experiments to test these hypotheses.

Kriegel's claim is that introspective reports, adequately gathered and properly controlled, should play a justificatory role too.

On the face of it, this claim seems entirely reasonable. After all, verbal reports of inner states regularly play a justificatory role in some areas of cognitive science, and while not without critics, they can be well validated and informative. Despite the oft cited Nisbett and Wilson (Nisbett and Wilson 1977) study suggesting that subjects' reports about decision making should not be trusted, Ericsson and Simon (1980, 1993) generated a framework that is still being used (with modification) for using verbal reports to gain data about thought processes in problem solving. The main two procedures are 'thinking aloud' and 'talking aloud' during the completion of a cognitive task, but the application of these procedures are limited to avoid methodological pitfalls (as found in the Nisbett and Wilson study).

First, these procedures are limited to cases where one can theoretically motivate the idea that task performance relies on complex thought that takes a form similar to inner speech. In this case verbal reports are outward expressions of existing (verbalised) thoughts; they make public, in a fairly straightforward way, what was private. In cases where the verbal report goes beyond this task-related inner speech, such as when subjects are asked to carefully explain their task strategy, such reports are likely to distract attention away from the main task and as a result subjects are less likely to generate accurate reports (e.g. they may confabulate). Second and relatedly, the general validity of the procedures can be tested by seeing whether generating a verbal report during a task changes task performance or task-related memory. If generating a verbal report alters task performance, this shows that to produce the report, subjects are engaging in additional cognitive processing to that directed at the main task. This in turn threatens the accuracy of the verbal report. Third, these procedures are limited to tasks where there is a right answer (e.g. mental arithmetic). The validity of subjects' verbal reports about their thought processes can be tested by comparing their reports with features of their task performance (accuracy, reaction times, eye movements) and sets of alternative possible cognitive strategies (via task analysis). Mismatches between reports and task performance suggest that the reports are inaccurate.

These three constraints seem to rule out safely using verbal reports to learn about the properties of experience. In particular, the focus of these constraints is on avoiding subjects engaging in 'reactive' cognitive processing, where the task of generating verbal reports changes the processes or states under investigation, and usually leads to the generation of inaccurate reports of these targets. This is a very familiar problem from the case of introspection: here there is a significant question of whether engaging in introspection changes the experience itself, and so whether the introspective report accurately reflects experience. Accordingly, Ericsson (2003) rejects the justificatory value of studies where the possibility of reactive reports cannot be ruled out, and where 'open-ended' introspection is used. He is however supportive of these reports in playing a role in generating testable hypotheses:

> …introspecting subjects are instructed to engage in additional observation and noticing. Which aspects of the cognitive processes and for how long these aspects are observed are likely to be unpredictable to both the subjects and the

> experimenter. As a consequence, the traditional methods of validation….will not be available as tools for the analysis of open-ended introspective reports. This type of introspective report can still provide valuable opinions and ideas that might lead to the generation of interesting and more targeted hypotheses. (Ericsson 2003, p. 16)

Therefore this approach, while supportive of the justificatory role of verbal reports in some limited contexts, works against optimism about introspective reports being able to play a methodologically robust role in justifying claims in consciousness science. Indeed, it is this inability to use standard methods of validation, combined with a lack of knowledge of how introspective reports about consciousness are generated that causes endless problems in consciousness research.

Similarly, in reply to Kreigel's optimistic claims about the widespread veridicality of introspection, Schwitzgebel (2013; see also Spener 2013) questions in just what conditions we can take introspective reports to be more likely veridical than not, and what counts as an elusive phenomenology. One worry is that most instances of introspection will not generate reports that are more likely to be veridical than not, since most instances of introspection are not done under 'normal' conditions (which need specification), and because many features of our experiences are elusive. A related worry is that we (as yet) have little idea under what conditions introspective reports are likely to be veridical, but that these conditions may be both very complex and very limiting. This paper is an attempt to rigorously address these worries and to see how substantial they are, by further analysing introspection as a form of measurement.

3 Introspection as Measurement

First, despite the preceding discussion it may initially sound odd to treat introspection about conscious experiences as a form of measurement, and introspective reports as measurements of properties of conscious states. While introspective reports may form one way of 'capturing' and making public certain aspects of experience, it is not immediately clear that this amounts to measurement, or what it is that is being measured in particular cases.

However, measurement can be thought of in a minimal way. At heart, measurement is a method for identifying some aspect of an object or event, and labelling it according to some (publicly shared) format or metric. This minimal account does not make great demands of introspective reports about consciousness in order to be treated as measurements. And if introspective reports are to play a justificatory role in consciousness science, then they will have to satisfy this definition. Introspective reports that do not successfully identify some aspect of a conscious state, and which cannot be put in the form of a publicly shared format or metric, are of no scientific use. That is because if introspective reports do not tell us about the target phenomenon in some kind of publicly shared format, then they cannot be used to make comparisons and generalisations about instances of experience across subjects, and from here it is not clear what theoretical claims they could be used to justify or challenge.

In this case, if introspective reports are not measurements in this minimal sense, they cannot play a justificatory role in consciousness science.

Indeed, introspective reports of some kinds can be treated fairly straightforwardly as measurements of properties of conscious states. Introspective reports have been used to measure and compare properties of experience like the clarity, visibility, and brightness of different visual experiences, often using graded response scales or pre-set response categories. This involves treating some target property of an experience as falling along a (linear) scale, where introspective reports label an experience as falling somewhere along this scale, with the hope that the reports can be used to order these experiences in terms of how well they exhibit the target property. The more methodologically rigorous experiments that use these kinds of introspective reports are also mindful of potential problems with the variability, calibration, and validity of the introspective reports that are generated.

Further, one of the very general discussions about introspective reports falls under the umbrella of measurement. This discussion stems from the unexpected degree of variation in introspective reports about conscious experiences across subjects, across time in the same subject, even under what one might consider controlled conditions, and what this means for the 'trustworthiness' of introspective reports.

For example, the variation in subjects' reports for fairly simple perceptual tasks is well known in psychophysics (for classic treatment see Swets and Green 1966; for more recent framework see King and Dehaene 2014). In addition, using Hurlburt's 'Descriptive experience sampling' technique (e.g. Hurlburt and Schwitzgebel 2007), which includes a fair amount of post-experiment debriefing to control for possible subjective bias, there can still be a significant degree of variation in subjects' reports about their experiences (Schwitzgebel 2007). Finally, Dennett (e.g. 1993, 2002) and Schwitzgebel (e.g. 2002a, b, 2008, 2012) have argued that we are sometimes radically inconsistent, and sometimes just plain wrong, when we report what things are like for us. This is argued to be true for a variety of phenomenal states, including visual experience, dreams, imagery, and emotional states, among others (for other examples see e.g. Bayne and Spener 2010).

Findings like these underlie attitudes of varying degrees of skepticism towards the veracity of introspective reports about consciousness across both philosophy and consciousness science. This is because they generate what Hohwy (2011) has called the 'Argument from Variability':

> AV1: There is evidence of introspective variability across conditions and across subjects.
> AV2: Introspective variability across conditions and subjects is best explained by introspection's being unreliable.
> So, by inference to the best explanation, introspection is unreliable. (p. 265)

Before proceeding however, and in order to make the core terms more clear, it helps to follow a standard framework for assessing measurement that includes reliability, validity, and accuracy (e.g. Carmines and Zeller 1979). Here, *reliability* refers to the stability or repeatability of a measure when it is applied to (what we take to be) the same phenomenon under the same conditions. *Validity* refers to how well (or if) a measure targets what it is meant to, and not some other property

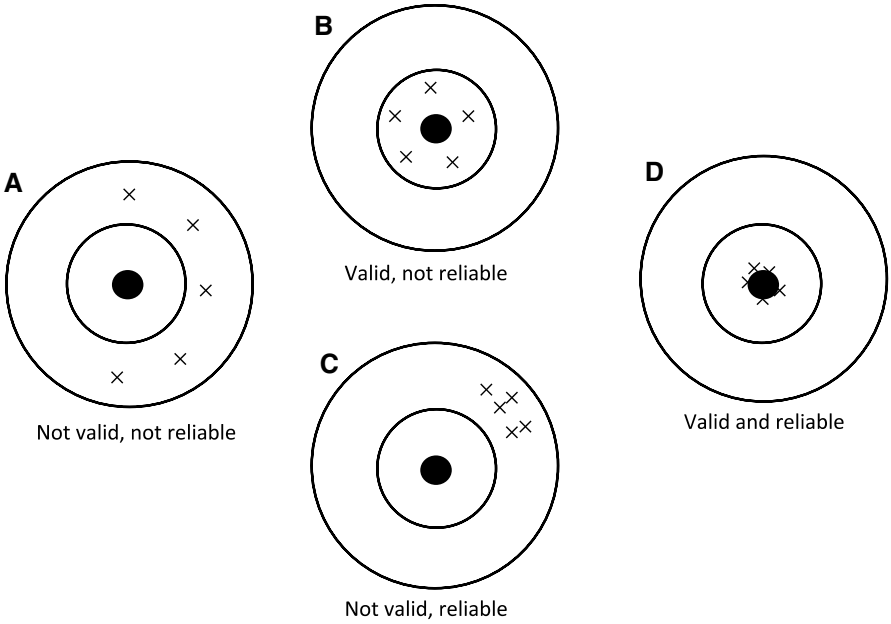

Fig. 1 This figure shows possible combinations of validity and reliability. Each 'target' contains a bulls-eye, which represents the real value of the measured property. A measure is reliable if repeated measures (×) fall within the same small area on the target. A measure is valid if repeated measures are clustered around the bulls-eye. A measure is more accurate the closer the x's are to the bulls-eye. **a** Shows an unreliable and also invalid measure. **b** Shows an unreliable but valid measure. **c** Shows a reliable but invalid measure. **d** Shows a measure that is both valid and reliable. Unreliable measures are either not valid (**a**) or inaccurate (**b**). However, reliable measures are not always valid (C)

that perhaps correlates with the target. *Accuracy* refers to how well the measure reflects the phenomenon being measured; how fine-grained the measure is and what the measurement error is. In the literature on introspection 'reliability' is used to mean some combination of all of these, such that 'reliable introspective reports' are those that veridically describe properties of a subject's experience. From now on, all uses of 'reliability' will follow measurement theory, so mean the repeatability of a measure.

As discussed in Chang (2004), testing whether a measure is reliable (repeatable) is a key step in establishing a measurement system. Whether a measure is reliable across the same conditions is the only property of a measurement that can be (more or less) directly observed; whether a measure is valid or accurate can only be inferred. As shown in Fig. 1, if a measure lacks reliability, this shows that the measure is either invalid or inaccurate. That is, if a measure of property P is not repeatable under conditions where it should be repeatable, then either this measure is measuring some other property Q, or it is an inaccurate measure of P. Unreliability, as in the Argument from Variability above, therefore highlights a problem with the measure. It is important to note that the mere presence of reliability in a measure does not however imply that it is accurate (it could routinely

make the same pattern of errors) or that it is valid (it could measure some other variable that is correlated with the variable of interest). Nevertheless, analyzing the reliability of a measure is an important step in analyzing a measurement procedure.[2]

Using this framework and terminology of measurement, the Argument from Variability above becomes:

> AV1′: There is evidence of unreliability in introspective data across conditions and across subjects (there is variability where we don't expect it).
> AV2′: Given the framework of measurement, introspective unreliability across conditions and subjects is best explained by introspection either being inaccurate (introspection is error-prone), or possibly invalid (introspective processes do not track properties of conscious experiences).
> So, by inference to the best explanation, introspection is either inaccurate or invalid.

This problem is found in various guises in the literature on introspection, and is a fairly recognizable problem given the framework of measurement. There are various ways of dealing with it in particular cases (e.g. Hohwy 2011 offers one), discussed later.

There are also other more explicit discussions of measurement and instrumentation in relation to introspection. Piccinini (2003, 2009) introduced the idea of introspective agents as 'self-measuring instruments'. This was primarily done to rebut the claim that introspective data is private data, and thus not scientific data. Instead, introspective data is firmly public, and consists of recorded behaviours and verbal reports.

Piccinini further argues that, like with any other instrument or measurement procedure, making use of the introspective reports generated by these self-measuring agents demands careful experimental design and precise task instructions, calibration (perhaps via training), and careful interpretation. Piccinini notes that this focus on the agent as an instrument, and the experimenter as the observer, is consistent with the idea that experimenters can use and interpret data from an instrument without knowing very much about how the instrument works, or about the phenomenon being measured (e.g. citing Hacking 1983). Instead, a skilled experimenter may be able to develop complex experimental protocols or interpretive frameworks with only the practical knowledge of the instrument gained in the lab. Given that there is currently very little scientific research on the process of introspection (see e.g. Overgaard et al. 2006a for a rare exception), this sounds promising.

However, Feest has argued that using introspection as a way of measuring properties of experience is far from straightforward (Feest 2012a, b). She argues

[2] Evaluating whether a measure is reliable requires identifying conditions under which the target phenomenon is in the same state, in order to repeatedly measure it. This is not always straightforward, and often relies on some basic theoretical assumptions about the target phenomenon, so is not an entirely independent or direct way of assessing the measure. This is potentially problematic for investigations of experience: we do not know which factors to control for when attempting to generate similar/same experiences.

that we in fact need to know rather a lot about introspection in order to use introspective reports, based in part on disanalogies between introspection and standard assumptions or techniques used in developing measurement methods. For example, compared to the use of standard instrumentation, there are still ongoing debates about what exactly counts as introspection (e.g. compared to perceptual reports), and whether the process of introspecting affects phenomenal states and whether this affects the validity of introspective reports (see discussion in Sect. 2). Feest also argues that we need to have a firmer grasp of exactly what the research questions are that are to be addressed by introspective data, as this will naturally affect experimental procedures and interpretive frameworks. Given this, she argues that forward progress must be made on a case by case basis, where "…our understanding of the status/meaning of introspective data co-evolves with our understanding of the ways in which they are generated…" (Feest 2012a, p. 13).

This co-evolution of theory and measurement fits well with Chang's claim of the necessity of using a coherentist approach in the development of measurement techniques (see esp. Chang 2004, pp. 220–234). Measurement devices and protocols, calibration techniques, and interpretive frameworks ultimately rely on theory about the phenomenon being measured, and theories about phenomena are developed from the very same measurements (as well as other sources of empirical evidence). Progress proceeds by indirectly testing and building on assumptions, broadening the scope of and adding detail to both measurement procedures and theory, and identifying and correcting errors, by relying on coherence as a (defeasible) marker of progress.

The rest of the paper unpacks the further potential of this approach, in particular where the framework of measurement can be used to identify and characterize methodological problems that might otherwise be missed.

4 A 'Bottom-Up' Approach

One way to analyse the potential evidentiary role of introspective reports is to see whether general processes of developing measurement procedures can be applied to them, and if so, what the evidential value of the ensuing reports are. The process analysed in this section is from Chang's (2004) work on measuring temperature, which provides a kind of 'bottom-up' approach to developing a scale of measurement. It is 'bottom-up' in the sense that it uses local coherentist strategies of fixing and testing constraints, making it possible to slowly expand the scope of the measurement procedure. This is in contrast to the explicitly comparative, cross-validation approaches considered later. While Chang notes that this pattern of development of a measurement procedure may not generalise, it provides a place to start, and as below, usefully identifies two core methodological problems related to the use of introspective reports about consciousness.

4.1 Developing Measurement Systems

As a very rough summary, there are four stages in the development of measurements of temperature (Chang 2004). First, there were thermoscopes, which indicate when one thing is hotter than another, but do not for example indicate how much hotter. Second, to allow comparisons across instruments, stable fixed points were identified and precisely specified (e.g. 0 and 100 Celsius as the boiling and freezing points of water under specific conditions), and instruments were calibrated to these fixed points. Third, it was established how the scale in the measuring instruments is related to temperature between the fixed points. For example, the expansion of mercury in a thermometer may not be linear as temperature increases, so a mark halfway between 0 and 100 °C on a mercury thermometer may not indicate a temperature of 50 degrees. Fourth, the measurement scale was expanded outwards beyond the fixed points, including how to measure temperature when standard thermometers freeze or melt.

Two important ontological assumptions underpin this process. The first is realism: it is assumed that temperature is a feature of the world, and is not a theoretical or experimental construct. The second is the Principle of Single Value: it is assumed that the property being measured has exactly one value at a time.

Both of these assumptions seem warranted when trying to measure temperature; it seems entirely reasonable to think that temperature is not a construct and things only have one temperature at a time. However, the Principle of Single Value is crucial not just as a starting assumption, but also to the practice of using reliability as a guide to problems with the validity or accuracy of a measure. In Fig. 1, there is precisely one bullseye on each diagram, which is equivalent to the Principle of Single Value. This makes it possible to claim that if repeated measures do not cluster together then the measure is either invalid or inaccurate, and so revisions must be made to the measurement procedure. However, without the assumption of the Principle of Single Value, it cannot be assumed that there is exactly one bullseye. If this is the case, then a lack of clustering means nothing: it is consistent with either having a measure that is wildly invalid or inaccurate (if the property does in fact only take one value at a time), or with having a perfectly valid and accurate measure (if the property takes multiple values at the same time).

4.2 Application to Introspective Reports

This section uses a fairly well known example to test whether something like this process of measurement development can work for properties of consciousness. This example is of introspective ratings of the ‘clarity’ of an experience, similar to subjective ratings of the ‘visibility’ of a stimulus. For example, Overgaard et al. (2006b; see also Ramsøy and Overgaard 2004) instructed participants to generate their own phenomenal categories of visual clarity, with associated verbal descriptions, in a training session, which were later integrated to form a standard ‘Perceptual Awareness Scale’ used in experimental studies (for reviews see Sandberg et al.

2010; Timmermans and Cleeremans 2015). Other researchers (e.g. Del Cul et al. 2007; Sergent and Dehaene 2004) regularly use similar subjective 'visibility' ratings to identify the presence or absence of conscious perception for subjects, and through this identify neural mechanisms for consciousness.

The use of these kinds of introspective reports is fairly widespread, and they have the potential for identifying the boundaries and features of conscious perception in a way that goes beyond standard and more 'behaviouristic' psychophysical measures. They are therefore an example of the more exciting and potentially revolutionary kinds of introspective reports that might help resolve long-standing debates about the nature and boundaries of consciousness. In trying to replace behavioural measures of consciousness, and in trying to use much of the same methodological machinery as these existing measures, they also easily fit under the umbrella of measurement. The question is then whether these rating scales can in fact be developed and validated within standard measurement procedures.

First, it seems reasonable to assume that introspectors are capable of thermoscope type measures, at least under normal conditions. That is, introspective agents are capable of telling, with a reasonable degree of validity and accuracy, when one experience is more clear than another, or when one stimulus is more visible than another. There might be some problems with this claim, but grant it for now.

The second stage is to identify fixed points, which makes it possible to calibrate different measuring instruments (e.g. different introspecting agents) to these anchor points. Fixed points in introspective scales would be incredibly useful, as for example a fixed point of having no clarity of experience, or no stimulus visibility, could be used as an indication of a lack of a conscious visual experience of a stimulus. This could then be used to delineate the boundary between conscious and unconscious perception. Indeed, this is the primary use of these introspective rating scales: to challenge and provide alternatives to theories of conscious perception that are based on more 'behavioural' measures (i.e. as is used in Ramsøy and Overgaard 2004; Del Cul et al. 2007; Sergent and Dehaene 2004).

However, fixed points cause problems when it comes to introspective scales like clarity. This is because these introspective judgements are tied to the task or broader context they are made within. As it turns out, how 'clear' you rate an experience depends on what you need to do with it. An experience that is clear enough to ground a simple response can be rated as having high clarity. However, if the same experience[3] is not clear enough to ground a more complex response, it will now be rated as having low clarity.

This idea is illustrated in more detail with the toy example in Fig. 2 (adapted from psychophysical research). In the first task (on the left), you the participant are shown a series of images on a screen, for 50 ms each, with a short break in between each image. Sometimes the screen shows a triangle as a stimulus, but sometimes it is a blank screen (no stimulus, marked by square brackets). Immediately after the presentation of each image you are asked two questions. The first question is a detection

[3] It is somewhat problematic to talk about 'the same experience'; it should be taken to mean an experience with the same (or very similar) phenomenal properties as another.

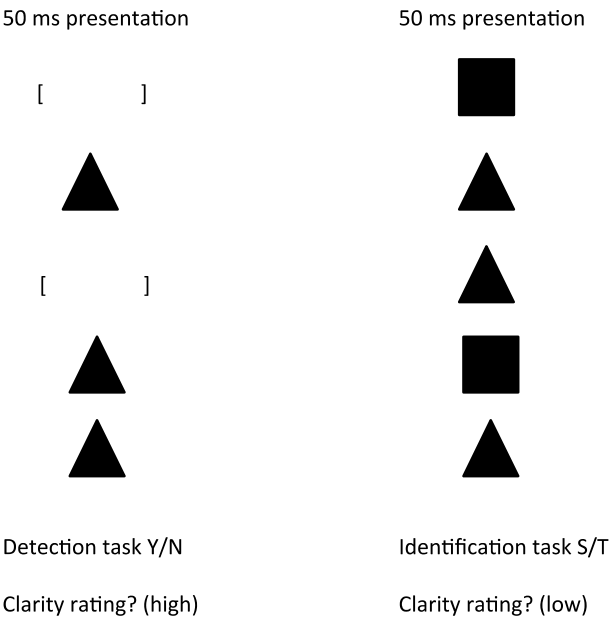

Fig. 2 Two different tasks using clarity ratings. See text for full explanation

question: was there a stimulus present (yes/no). The second is an introspective question: how clear was your experience of the stimulus (e.g. a clarity rating on a scale of 1–5).

With practice at this task, you can get fairly good at correctly answering the detection question. In this case it seems likely that you will give high clarity ratings about the experiences that allow you to successfully detect the stimuli, and you will be fairly confident in your answers.

In the second task (on the right), the set-up is basically the same. However, this time you are shown images of either a triangle or a square, and are now asked an identification question before giving a clarity rating. Identification questions are harder to answer correctly than detection questions, so even with practice your performance may not be good, and you will likely not be very confident in your answers. In this case you are likely to give lower clarity ratings for the experiences that support this lower level of performance.

Here is the problem: across these two simple tasks, you will give two different clarity ratings for the same stimulus (50 ms presentation of a triangle). On the assumption that the same stimulus will generate the same experience, you will therefore generate two different clarity ratings for the same experience, depending on what task you are currently performing.

Importantly, the results in this toy example are grounded in solid experimental work. In particular, confidence ratings tend to track task performance and task difficulty (Haase and Fisk 2001; Hertzman 1937; Nickerson and McGoldrick 1963; Sandberg et al. 2010) and clarity ratings track confidence. For example, when subjects in (Ramsøy and Overgaard 2004, see p. 12) generated verbal descriptions for

categories of visual clarity, different discrimination abilities and confidence levels featured heavily. The category of 'no experience' was defined as 'no impression of the stimulus is experienced. All answers are experienced as mere guessing', suggesting a lack of discrimination ability and no confidence in the participants' responses. The category of 'almost clear experience' was defined as the 'feeling of having seen the stimulus, but being only somewhat sure about it', here explicitly about having mid-level confidence in detection tasks. Even more specifically, Wierzchon et al. (2014) found that participants' responses about their awareness of stimuli using a variety of subjective rating scales were affected by responses they gave in identification tasks (see also Sandberg et al. 2010).

Importantly, this is true even for extreme points on a scale of clarity ratings. It is well known that for subjective reports, which include introspective reports, a range of factors affect when subjects report having 'no experience' of a stimulus, including task type, structure, difficulty, and motivation. As Timmermans and Cleeremans (2015) note: "One cannot emphasize enough how apparently small differences in procedures may lead one to strikingly different conclusions when it comes to distinguishing between conscious and unconscious cognition" (p. 41). This means that for two participants shown exactly the same set of stimuli, but doing different tasks, or paid differently, the instances under which they report having 'no experience' can be radically different (again see King and Dehaene 2014; Swets and Green 1966; Irvine 2009, 2012a, b).

Here then clarity ratings track task difficulty (detection vs. identification), task performance and confidence. This means that even if it is possible to set up fairly stable and shared fixed points within a specific task, these fixed points are not transferable across tasks. An experience of a triangle in an easy task can get a high clarity rating, but change the task difficulty and the (plausibly) same experience is now given a low clarity rating. The phenomenal properties of the same experience can be rated differently depending on the task at hand.

In sum, phenomenal properties like clarity and visibility are partly defined by, or are at any rate deeply sensitive to, features of task and context. This makes it impossible to generate task- and context-independent fixed points for phenomenological properties like clarity. But this is precisely what is needed to move beyond using simple thermoscope-type judgements, and to enable comparisons of introspective reports of phenomenal properties across instruments, measurements, and tasks.

One possible response is to accept that what subjects take phenomenal properties like clarity or visibility to mean vary task to task, and claim that this at least doesn't prevent us from investigating these various interpretations of these phenomenal properties. This, I think, is to just accept and ignore the problem though. As evidenced in the quotation from Timmermans and Cleeremans above, "apparently small differences in procedures" can generate very different interpretations of what phenomenal properties like clarity or visibility mean. If we cannot meaningfully compare introspective reports about phenomenal clarity, or subjective visibility, across even *small* differences in experimental procedures, then it is not clear that these concepts are doing any work in picking out interesting and stable properties of experience in the first place. In particular, if introspective reports of visibility genuinely pick out different properties of experience across very similar tasks, then it is

far from clear what use they can be in identifying the boundaries of conscious perception. Yet this is exactly what introspective reports about this phenomenal property is supposed to help us do.

Importantly, this point generalises. Any introspective reports that require some degree of comparison to other stimuli (either present or from memory), oor that are related to confidence ratings, or that enable different kinds of actions with greater or lesser likelihood of success, will be strongly modulated by the conditions under which they are made. It is hard to think of an introspective response that does *not* have one or more of these features: responses concerning the properties of a stimulus (colour, shape, duration, presence/absence), and responses concerning perhaps more elusive properties of experiences themselves (blurriness, clarity, richness, intensity), will all be affected. This strongly suggests that there it is not possible to find fixed points across many (if not all) kinds of introspective response. This further means that many (if not all) kinds of introspective reports cannot therefore move beyond stage 1 of the process of generating systems of measurement outlined above. At least one way to develop scientifically rigorous introspective measurements is out of bounds.[4]

4.3 Ontological Assumptions

There is a further problem though which runs even deeper, and applies more broadly, related to the ontological assumptions mentioned earlier. For temperature and many other phenomena, making the assumptions of realism and the Principle of Single Value are fairly straightforward. However, for properties of experience things get a bit more complicated.

In particular, it is not clear whether the Principle of Single Value (PSV) is really a safe assumption to make when it comes to properties of experience.[5] It is not outlandish to think that a property of experience might sometimes have multiple values, be indeterminate, or have no value at all. As explained in Sect. 4.1, an inability to apply the Principle of Single Value would mean that reliability or unreliability would be useless as a way of assessing a measurement procedure.

Indeed, whether or not the PSV can be made in certain cases is up for debate. For example, Schwitzgebel (2002a) notes that participants are often highly uncertain about features of their experiences when engaging in mental imagery. They might claim, for example, that they just don't know whether a mental image is in colour or in black and white, or how many windows an imagined house has. Schwitzgebel uses this as evidence towards a broad skepticism about epistemic claims based on introspection, but it might also show that some phenomenal properties are not determinate.

[4] Some interesting progress has been made on combining objective (bias-free) and subjective measures (Maniscalco and Lau 2012) but the resulting measures are not strictly introspective, and are usually associated more with measuring metacognitive abilities than (first order) experience.

[5] The assumption of realism is also potentially problematic, if the new 'illusionists' are to be believed (e.g. Frankish 2016).

Further, Hohwy (2011) claims that introspective uncertainty is to be expected with some cases of introspection on mental imagery. Hohwy argues that mental imagery is usually tightly constrained and goal-oriented, which ensures that salient features of the mental images are fixed. However, when used for introspective tasks, mental imagery is usually open-ended. In this case it is not surprising that various features of the mental image are not fixed, and so do not take exactly one value. This provides both a good reason to drop the PSV in this case, and explains the source of introspective uncertainty.

Note though the strategy used to decide whether or not the PSV can be taken to hold in the case of mental imagery, and so how much weight to place on the variability of subjects' reports about it. Here, Hohwy draws on evidence that is independent of introspective reports to suggest that introspective experiences of mental imagery really are uncertain, which explains why people are often uncertain in reporting about them. That is, behavioural evidence or cognitive frameworks are used to describe what experience is like (i.e. uncertain), which then explains away the unreliability and uncertainty found across a certain set of introspective reports. This then is an example where the status of introspective reports is directly evaluated by checking their contents (uncertainty) against more trusted sources of evidence (theoretical descriptions of the cognitive process of mental imagery). This is an initial clue about the evidential dependence of introspective evidence, discussed in more depth below.

For now though, note two methodological problems related to introspective reports that stem from considering this 'bottom-up' approach to developing measurement procedures. First, there are no task-independent fixed points for introspective rating scales related to properties of experience like clarity and visibility, and (plausibly) for other kinds of introspective responses too. This prevents introspective agents from moving beyond thermoscope type outputs, and prevents comparison of introspective reports across tasks and contexts. Second, it is far from clear when the Principle of Single Value actually holds with respect to properties of experience. If this is true, then researchers cannot rely on the reliability (repeatability) of a measure to tell them anything about the validity or accuracy of introspective reports. Further, in order to identify when the PSV holds, researchers are often forced to develop cognitive theories based on non-introspective evidence to suggest what experience is actually like in the case in question. If we can develop theories of phenomenal properties in this way though, in order to explain away introspective variability or uncertainty, then the justificatory role of these introspective reports is left obscure.

5 A Comparative Approach: Cross Validation of Measurement Procedures

It looks like the 'bottom-up' approach to developing introspective measures is unlikely to work, but there are of course other ways to develop and validate measurement procedures. One of the most common ways to do this in the cognitive sciences is to use cross-validation. Here, the outputs of a range of independent measurement procedures are compared with each other. If the results they

generate cohere with each other, then this is reason to think that each of these measurement procedures is valid (probes the target phenomenon/property) and provides accurate measurements (gets very close to the real value of the target property). This is because it is more likely that each of the procedures is valid and accurate rather than that each procedure is invalid and inaccurate, but generates the same results by chance. The more procedures that cohere with each other, the more likely it is that each is a valid (and accurate) one. The key reason for doing this is not just to test the validity and accuracy of a procedure in the particular cases in which cross-validation is attempted. It is instead used as a way of validating a procedure *in general*, such that it can be used across novel contexts, where researchers can be confident that the validity and accuracy of the procedure still holds.

However, while comparative approaches are used very broadly in consciousness science, it is not always clear if they are aimed at validating introspective procedures themselves, or aimed at validating specific sets of introspective reports. For example, comparative approaches are regularly used to assess different measures of consciousness (Sandberg et al. 2010; Seth et al. 2008), and introspective reports tend to be compared with 'external performance criteria' like other behaviours, neurophysiological data, or more 'objective' kinds of reports, in order to gauge the accuracy of each. However, these exercises are not usually explicitly aimed at validating introspective procedures per se. Similarly, Jack and Roepstorff (2002) suggest a method of 'triangulation' comparing introspective reports, other behavioural data, and neurophysiological evidence to better understand features of consciousness. Within the neurophenomenological approach this process is referred to as 'mutual circulation' (for early proposal see Varela 1996). Again though it is not explicitly aimed at generating valid procedures for eliciting introspective reports across a range of tasks, but rather at validating a particular set of introspective reports.

Nevertheless, some comparative approaches can be seen in this light, particularly when different procedures for introspective reports are compared with each other (e.g. as in Sandberg et al. 2010). Here, the aim was to identify the differences between the procedures and to come to some conclusion about which was better (more valid and accurate). In general, one might also argue that using cross-validation is key in introducing any new procedure for generating introspective reports, combined with any theoretical justification for it that can be provided.

Despite the potential power of the cross-validation approach towards introspective procedures, I will argue that there are significant problems that affect both stages of its application: the first stage requiring the development of an introspective procedure that generates evidence that coheres with non-introspective evidence, and second inferring that the procedure is valid and so can be confidently applied in new contexts. Wide (if implicit) recognition of these problems probably also explains why such an approach is not more widely used. The slightly different method of using cross-validation for validating sets of introspective reports (rather than procedures) is discussed in Sect. 6.

5.1 Stage One

The first stage in a cross-validating an introspective procedure is to develop one such that it generates evidence that coheres with non-introspective evidence, for a small set of calibration cases or experimental paradigms. One might think that the problem is that it is hard to develop introspective procedure that do this. However, given the sensitivity of introspective reports to a wide range of factors, this part is relatively easy: via introspective training it is possible to manipulate subjects' reports in a variety of ways, such that they cohere with sets of non-introspective evidence.

The problem is instead what the relevant non-introspective evidence is, such that introspective reports should cohere with it. There are a wealth of theories and related behavioural and neurophysiological ways of probing consciousness, all of which typically diverge, and as discussed above, sometimes quite dramatically (see also Timmermans and Cleeremans 2015; Irvine 2012a, b; Seth et al. 2008). For example, according to some measures and theories, consciousness is definitely present under some specific conditions, is always marked by the presence of feedback activity in the brain, and/or that all experiences are accompanied by phenomenal self-awareness. Other measures and theories say the exact opposite. It is possible to get subjective visibility ratings to cohere with 'objective' measures of consciousness which measure core discrimination capacity (e.g. Dehaene and Changeux 2011, p. 201), using substantial training on highly artificial tasks. However, this commits one to the controversial claim that objective measures of consciousness are genuinely only picking out instances of consciousness, such that introspective reports that cohere with them do the same.[6]

Picking a set of non-introspective evidence or a set of theoretical predictions that introspective procedures should cohere with is therefore to make a significant theoretical commitment about the nature of experience and how to (non-introspectively) probe it. If the aim of using introspective reports is to help resolve these debates by providing (relatively independent) justificatory evidence, this shows that they cannot do so when using this methodology. One must already pick a side in order to validate the introspective reports themselves, in which case they lend no justificatory support.

5.2 Stage 2

Second, there are problems in making the inference that an introspective procedure that (apparently) generates valid and accurate reports in one context will continue to do so in another. As illustrated in Sect. 4.2, many of the factors that affect introspective reports are task specific. The history of psychophysical research on the factors that affect subjective reports about even simple aspects of visual experience suggests that these factors are high in number, that they interact, and that introspective reports can be very sensitive to small changes in them. Given the complexity of the

[6] See Irvine (2012a, b) for further discussion of the incompatibility between subjective and objective measures of consciousness.

introspective process, it is likely that an introspective procedure that 'works' across one set of tasks in one group of individuals is unlikely to work elsewhere.

In evaluating whether there is likely to be a single process of introspection (and if so, what it is), Schwitzgebel (2012) makes a similar point. He writes that:

> What we have…is a cognitive confluence of crazy spaghetti, with aspects of self-detection, self-shaping, self-fulfillment, spontaneous expression, priming and association, categorical assumptions, outward perception, memory, inference, hypothesis testing, bodily activity, and who only knows what else, all feeding into our judgments about current states of mind (p. 41).

So, if an introspective procedure generates results that cohere with other evidence across a particular set of tasks, this then suggests that appropriate context-specific controls have been deployed in those tasks (possibly to encourage/facilitate coherence). But in order to use an introspective procedure more broadly, such that we would expect introspective reports to continue to cohere with non-introspective evidence, we would have to identify and use a different set of context-specific experimental controls, or develop some kind of interpretive framework to evaluate the responses generated.

The problem is that identifying and implementing controls or interpretive frameworks across contexts is far from straightforward. This is particularly true given that the use of an introspective procedure often relies on (extensive) context-specific training that is tied to the measure or theoretical framework that it is being validated against. Other than going through the whole validation procedure anew each time an introspective procedure is used, it is really not clear how to proceed. This means that confidently deploying introspective procedures across new tasks and individuals is just not easy as might be claimed.

These two problems provide good reason to think that cross-validating introspective procedures is, at the very least, extremely difficult. And perhaps not surprisingly, cross-validation of introspective procedures is not common. This seems to be largely based on an (implicit) recognition of the problems outlined here: that validating introspective procedures means making a commitment about the nature of consciousness, and that successful introspective procedures are likely to be specific to certain kinds of tasks and unlikely to generalise. As a result, researchers instead tend to focus on much more local cross-validation of the *outputs* of introspective procedures, rather than the procedures themselves. This is discussed below, in particular focussing on the evidentiary status of introspective reports within this methodology.

6 A Comparative Approach: Cross Validation of Evidence

Comparing and cross-validating introspective results with other behavioural data and neuropsychological evidence is common in consciousness science (see some of the comparative approaches outlined in Sect. 5), but here the focus is on validating sets of introspective evidence, rather than introspective procedures. The idea is that if introspective evidence coheres with non-introspective evidence then this is reason to think that each of the forms of evidence are valid and accurate.

This is because it is more likely that each of the sets of evidence are valid (relate to the target phenomenon/property) and accurate (get very close to the real value of the target property), rather than that the evidence is invalid and inaccurate, but matches through chance.

Similar problems arise with this approach, but with a different twist. One recurrent problem is what kind of non-introspective evidence should be used to compare introspective reports against. The problem is the same for the cross-validation of evidence as for the cross-validation of procedures as discussed in Sect. 5 above. But there is a further twist on it that is more visible here: what to do when sets of introspective and non-introspective evidence diverge, as they do fairly often. If introspective evidence coheres with or matches non-introspective evidence, then all is well, and all the evidence is validated. But if there are mis-matches there need to be pre-specified ways of moving forward; it is no good just reporting on differences and leaving it there. For cross-validation and comparative practices to be productive, they need detailed rules of engagement.

Crucially, when it comes to comparative approaches, these rules of engagement are often either missing or take a negative stance towards introspective reports. For example, in applying Jack and Roepstorff's (2002) method of 'triangulation' to introspective reports, there is no guidance about what to do in cases of mis-match between introspective reports and other data. In both Seth et al.'s (2008) and Timmermans and Cleeremans (2015) reviews of measures of consciousness, the strategy for moving forward is one of comparison and triangulation, but again with no suggestions of exactly how this should work.

In contrast, Bayne and Spener (2010) explicitly claim that in cases of mis-match, introspective reports can be ignored; only successfully scaffolded introspective reports should be trusted. 'Scaffolded' introspective reports are those whose content matches the content of a similar (perceptual) report. For example, the introspective report 'it looks like a red ball' would only be trusted (i.e. taken to be valid and accurate) if it could be scaffolded by a correct perceptual report like 'there is a red ball'. For introspective reports that cannot be scaffolded in this way, like judgements about cognitive phenomenology, a degree of scepticism is warranted towards introspective reports (though see Spener 2015 for alternative).

This is echoed more formally in comparisons of different subjective and introspective rating scales, where an objective behavioural baseline is used to assess how valid and accurate the subjective measures are. Roughly, objective measures of consciousness measure a subject's core discrimination capacity, which usually diverges from what is captured by subjective and introspective measures. All of these measures come with their own advantages and disadvantages (for review see Irvine 2013), but one of the distinct advantages of objective measures is that they are stable across tasks and free from response bias.

As a result, giving objective measures evidential priority is fairly standard methodology. Illustrative of this, Sandberg et al. (2010) state that they are "[g]oing with the tacit assumption that objective measures should be preferred over subjective (i.e. introspective) ones when studying consciousness" (p. 1077) in order to assess the merits of different subjective measures. That is, objective measures of consciousness are 'better' than subjective and introspective measures, which are simply not

as sensitive and accurate. Objective measures form the benchmark for introspective reports to be judged against.

The problem here is (I think) fairly obvious: when there are instructions about what to do in cases of mis-match between introspective and non-introspective evidence, cross-validation is done on the premise that introspective reports are evidentially inferior to other measures. To be clear, this negative stance towards introspective reports is a fairly sensible position to take. The arguments in Sect. 3 show that there are a wealth of unpredictable factors that can affect the content of introspective reports, and the Argument from Variability provides a further reason to be sceptical of their value. In this case, if introspective reports are the odd ones out in a comparative study, then it is reasonable to doubt the introspective evidence.

However, if non-introspective evidence is typically seen as having evidential superiority over introspective evidence, this entails that introspective reports have little or no independent justificatory status. First, in cases where introspective responses fails to fit with other evidence, introspective evidence is either ignored, assumed to be false, and/or form explananda in themselves. Obviously, in these cases, introspective evidence cannot play a strong justificatory role. Second, and based on this, in cases where the introspective evidence does fit with other evidence, it is not clear that it is particularly informative. That is, if introspective evidence is evidentially inferior, to the extent that it would be ignored if it did not fit with non-introspective evidence, then it is not clear what it (evidentially) adds when it does fit. This is compounded in cases where it is not known in any amount of detail why the introspective evidence fits or not, and in cases where there is dubious introspective training which artificially forces a 'fit'.

Together, this supports a strong and sceptical conclusion. The evidential value of introspective reports is deeply dependent on the evidential value of the non-introspective data that they are validated against, where the choice of relevant non-introspective data typically commits one to a significant theoretical stance regarding the nature of consciousness. Further, introspective evidence is only 'trusted' when it matches with non-introspective evidence, and ignored when it does not. Across all cases then, where introspective evidence either fits or does not fit with other evidence, introspective reports are not in a position to play a strong justificatory role in consciousness science, and are distinctly incapable of resolving long-standing theoretical and empirical debates.

7 An Objection

There is an obvious objection to the arguments above, which is to just learn more about the process of introspection.[7] If we knew more about how introspective reports are generated, we could gain more confidence in developing and applying introspective procedures, and so improve the evidentiary value of introspective reports to be at least as high, if not higher, than that of non-introspective evidence. In this case the

[7] Many thanks to Wayne Wu for pressing me on this.

arguments provided above would apply to current uses of introspective evidence, but possibly not future uses. However, the points raised above can be used to defend a stronger claim: that introspective evidence is unlikely to ever play a strong justificatory role in consciousness science. This can be illustrated by considering how such knowledge of introspection would be generated.

One approach would be to systematically study the generation of introspective reports in detail across a small range of tasks, identifying and manipulating factors to see how they change subjects' introspective reports. In essence, this is already done within classical psychophysics. One could imagine taking a broader approach though and try to isolate and manipulate factors that are traditionally out of bounds of psychophysics, including the sorts of factors identified in Schwitzgebel (2012): these might include self-shaping or self-fulfilment, memory and association. Once these factors are mapped out, one might then be able to control for them experimentally, or otherwise try to interpret introspective reports in light of them. This would not be easy: the number of factors that might be relevant, and how they affect introspective reports in isolation and in combination, is likely to be complex, to vary significantly across tasks, and also likely to vary across individuals (certainly for the more cognitive factors). At the very least, this will take a while.

However, this variability and complexity generates problems of its own. If different introspective agents generate different reports given the same stimuli, and are sensitive to small changes in experimental set ups, and sometimes even vary their own introspective reports about the same stimuli over time, then it is difficult to get started on mapping all this out. What is key is having something stable and reliable to work from. However, related to the 'bottom-up' approach to validation outlined in Sect. 4.2, this is usually absent when working with introspective reports alone. There are no phenomenal fixed points that introspective agents can be calibrated to such that they will give comparable introspective reports or ratings across tasks. And it is not even clear when one can assume the Principle of Single Value, that is, assume that some phenomenal property does indeed take a single value at a time. The massive variability in introspective reports, as seen in the Argument from Variability, makes it impossible to establish a stable baseline from which to construct a theory of introspection.

One very natural alternative strategy to take, and one is that is very common in consciousness science (and key in psychophysics), is to search for a stable baseline elsewhere, either from non-introspective measures of consciousness, or theoretical predictions on what consciousness is like in some particular case. On the assumption that this stable baseline accurately identifies when conscious perception is present or not, (and maybe some of its features in particular cases), researchers can then identify how to get subjects to generate valid and accurate introspective reports about these instances of consciousness. This may be done via precise manipulations of subjects' motivation, confidence, and by precisely wording task instructions. From here, one might be able to identify how, in detail, introspection works, by considering how it is affected by a barrage of different factors over different subjects.

This is to essentially to take a comparative strategy, but as seen from the above, comparative strategies do not work in introspection's favour. The biggest challenge is to identify what the best baseline is. For research on the difference

between conscious and unconscious perception, this is often taken to be an objective measure of consciousness, and similar behavioural tests are used as baselines for other aspects of consciousness. But again, this is to make a significant (and controversial) theoretical commitment about the nature of consciousness (e.g. that objective measures are good measures of the presence/absence of consciousness) before one has even started using introspective reports in a serious way.

Alternatively, one might take the predictions of a particular theoretical approach to provide a baseline. In the example of mental imagery earlier, there are introspective reports of phenomenal uncertainty, and a theoretical framework for understanding mental imagery. The framework predicts that mental imagery done under introspective conditions (i.e. not goal-driven) will generate uncertainty about phenomenal content. If this theoretical framework is right, then it explains away the uncertainty in introspective reports, and identifies the factors that generated it.

Yet both of these proposed 'solutions' illustrate the same fundamental problem: that making significant progress on how to collect and interpret introspective reports already demands that researchers commit to some fairly substantive claims about when conscious experience is like. In order to get better at evaluating introspective reports about (for example) the presence or absence of consciousness, or the clarity of conscious experience under certain conditions, we actually have to have some idea of when consciousness is present or not, and how certain conditions affect how 'clear' conscious experiences are. In order to get better at evaluating introspective reports about (for example) phenomenal uncertainty, blur, or other less obvious phenomenal features, we also have to have some idea of when and how they might occur, in order to rigorously evaluate introspective reports about them. That is, to engage with introspective reports will involve figuring out what experience is like based on more trustworthy kinds of evidence or (non-introspectively) well-tested theoretical frameworks, and from there, work backwards to identify the combinations of factors that explain why certain individuals make the kinds of introspective reports they do.

The key point then is that having done all this, introspective reports about a particular kind of experience might be scientifically usable, but at that point they are unlikely to tell us anything new. In particular, they are unlikely to play a role in providing strong justificatory evidence for claims about features of experience that have already been characterised by non-introspective means, and by which these new introspective responses were validated. Echoing this, and rather unsurprisingly, Froese et al. (2011) state that "[w]e are not aware of…any 'killer experiment' which would conclusively demonstrate that [an introspective method] has led to a substantial breakthrough in consciousness science" (Froese et al. 2011). The sections above explain why this is so. Introspective reports are so variable and their production is so complex that a non-introspective stable baseline of some kind is needed in order to evaluate them. But identifying a relevant baseline comes with problems and significant theoretical commitments of its own, and having chosen one in sufficient detail to engage in a validation process already answers the questions that introspective reports were supposed to help solve. In this case, due to basic methodological constraints, introspective reports themselves currently cannot and arguably will not play

a strong justificatory role in consciousness science, particularly in resolving the kind of long-standing debates where new forms of evidence would be most valuable.

8 Conclusion

In this paper I have argued that it is very unlikely, for basic methodological reasons, that introspective reports can play a strong justificatory role in consciousness science now, or in the future.

First, a 'bottom-up' non-comparative approach to validation does not take one very far. Constructing a measurement scale for introspective ratings runs into the problem of no fixed points: ratings about phenomenal properties of experience are massively context sensitive, so fixed points cannot be sustained across contexts. Using reliability as a defeasible marker of accuracy and validity requires the Principle of Single Value, which cannot always be assumed for phenomenal properties. This means that it is not possible to validate, in a bottom-up way, these kinds of introspective procedures such that they are guaranteed to work outside the narrow range of tasks and subjects where they have initially been (apparently) successful.

Second, cross-validation of introspective procedures runs into several recurring methodological dead ends. One is how to identify the relevant non-introspective measurement procedures that introspective procedures should be expected to cohere with. Any decision here is liable to be controversial. Another is that, as above, given the sensitivity of introspective reports to a range of factors, one is not warranted in making the inference that an introspective procedure that (apparently) generates valid and accurate reports in one context will continue to do so in another. If this inference cannot be sustained, then there is no value to trying to cross-validate introspective procedures (which may explain why it is rarely attempted).

Third, cross-validation of sets of introspective evidence is more common, but also runs into some of the same dead-ends. Again, it is not straightforward to identify the relevant non-introspective evidence to compare introspective reports to. It is also usually left unclear how to proceed when introspective evidence fails to cohere with the relevant non-introspective evidence. When it is made clear (though often only implicitly), the standard approach is to treat introspective evidence as being less trustworthy than non-introspective evidence. If introspective reports fit with what we already know from other methods then we accept them (but learn nothing new), and if they don't fit, then we ignore them, assume them to be incorrect and/or treat them as explananda in their own right. This means that whether or not introspective evidence coheres with non-introspective evidence, its evidential value is dependent on the value of the non-introspective evidence that it is validated against. In this case cross-validation offers no easy way out, and introspective reports end up with little justificatory power of their own.

Finally, the complexity of introspective processes ensures that this state of affairs will continue. Identifying how different groups of factors affect introspective reports across individuals, tasks and contexts well enough to interpret introspective reports in a scientifically rigorous way will likely be incredibly complicated. To do this will involve working backwards from 'simpler' and more trusted forms of evidence and

theoretical frameworks about what experience is like. At this point though, introspective data will add little of justificatory value; we will already know the answers that introspective data was supposed to provide.

However, introspective reports clearly do have an important set of roles to play in consciousness science and cognitive science more broadly. The very fact of introspective unreliability, uncertainty, or inaccuracy in particular cases is highly informative. If for example participants are highly uncertain in their reports (as in the case of mental imagery), this can perhaps be combined with independent evidence to suggest that the Principle of Single Value does not in fact hold for various properties of experience in certain cases. Introspective inaccuracy can be used to shed light on the nature of introspection and on participants' background beliefs about experience, and instances of introspective unreliability demand their own (often very local) explanations. Even if the nature of experience cannot be directly read off from the contents of introspective reports, it is still possible to learn about experience and the process of introspection by paying attention to higher level features (unreliability, uncertainty) of sets of introspective reports. Introspection will also obviously continue to play important and core roles in discovery, driving research questions and contributing to hypothesis generation. However, for the reasons discussed above, it is unlikely that introspective reports can ever play a strong justificatory role in consciousness science, or help resolve long-standing debates about the nature of experience.

Acknowledgements Many thanks to the anonymous referees for their comments which have made this a much better paper. Thanks to Wayne Wu and Matthias Michel for ongoing discussions about introspection and consciousness science, and the numerous conference and seminar audiences for their feedback on earlier versions of this paper.

References

Bayne, T., & Spener, M. (2010). Introspective humility. *Philosophical Issues, 20*(1), 1.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences, 30,* 481–548.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17). Thousand Oaks: Sage Publications.

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

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*(2), 200–227.

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology, 5*(10), e260.

Dennett, D. (2002). How could I be wrong? How wrong could I be? *Journal of Consciousness Studies, 9*(5–6), 13–16.

Dennett, D. (2003). Who's on first? Heterophenomenology explained. *Journal of Consciousness Studies, 10*(9–10), 19–30.

Dennett, D. C. (1993). *Consciousness explained*. Westminster: Penguin UK.

Dennett, D. C. (2007). Heterophenomenology reconsidered. *Phenomenology and the Cognitive Sciences, 6*(1), 247–270.

Ericsson, A. (2003). Valid and non-reactive verbalization of thoughts during performance of tasks towards a solution to the central problems of introspection as a source of scientific data. *Journal of Consciousness Studies, 10*(9–10), 1–18.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis; Verbal reports as data (Revised edition)*. Cambridge, MA: MIT Press.

Feest, U. (2012a). Introspection as a method and introspection as a feature of consciousness. *Inquiry, 55*(1), 1–16.

Feest, U. (2012b). Phenomenal experiences, first-person methods, and the artificiality of experimental data. In *Philosophy of science association 23rd biennial meeting*. San Diego, CA.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies, 23*(11–12), 11–39.

Froese, T., Gould, C., & Seth, A. K. (2011). Validating and calibrating first- and second-person methods in the science of consciousness. *Journal of Consciousness Studies*, *18*(2), 38–64.

Gallagher, S., & Sørensen, J. B. (2006). Experimenting with phenomenology. *Consciousness and Cognition, 15*(1), 119–134.

Haase, S. J., & Fisk, G. (2001). Confidence in word detection predicts word identification: Implications for an unconscious perception paradigm. *The American Journal of Psychology, 114*(3), 439.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science* (Vol. 5). Cambridge: Cambridge Univerity Press.

Hertzman, M. (1937). Confidence ratings as an index of difficulty. *Journal of Experimental Psychology, 21*(1), 113.

Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind and Language, 26*(3), 261–286.

Hurlburt, R. (2011). Descriptive experience sampling, the explicitation interview, and pristine experience in response to froese, gould and seth. *Journal of Consciousness Studies, 18*(2), 65–78.

Hurlburt, R. T., & Schwitzgebel, E. (2007). *Describing inner experience?: Proponent meets skeptic*. Cambridge: MIT Press.

Irvine, E. (2009). Signal detection theory, the exclusion failure paradigm and weak consciousness—Evidence for the access/phenomenal distinction? *Consciousness and Cognition, 18,* 551–560.

Irvine, E. (2012a). Old problems with new measures in the science of consciousness. *British Journal for Philosophy of Science, 63,* 627–648.

Irvine, E. (2012b). *Consciousness as a scientific concept: A philosophy of science perspective*. Berlin: Springer.

Irvine, E. (2013). Measures of consciousness. *Philosophy Compass, 8,* 285–297.

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: From stimulus–response to script–report. *Trends in Cognitive Sciences, 6*(8), 333–339.

King, J.-R., & Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 369*(1641), 20130204.

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences, 14,* 301–307.

Kriegel, U. (2013). A hesitant defense of introspection. *Philosophical Studies, 165*(3), 1165–1176.

Lutz, A., & Thompson, E. (2003). Neurophenomenology integrating subjective experience and brain dynamics in the neuroscience of consciousness. *Journal of Consciousness Studies, 10*(9–10), 31–52.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422–430.

Nickerson, R. S., & McGoldrick, C. C. (1963). Confidence, correctness, and difficulty with non-psychophysical comparative judgments. *Perceptual and Motor Skills, 17*(1), 159–167.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231.

Olivares, F. A., Vargas, E., Fuentes, C., Martínez-Pernía, D., & Canales-Johnson, A. (2015). Neurophenomenology revisited: Second-person methods for the study of human consciousness. *Frontiers in Psychology, 6,* 673.

Overgaard, M., Koivisto, M., Sørensen, T. A., Vangkilde, S., & Revonsuo, A. (2006a). The electrophysiology of introspection. *Consciousness and Cognition, 15*(4), 662–672.

Overgaard, M., Rote, J., Mouridsen, K., & Ramsøy, T. Z. (2006b). Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Consciousness and Cognition, 15*(4), 700–708.

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive Sciences, 5*(3–4), 229–269.

Piccinini, G. (2003). Data from introspective reports: Upgrading from common sense to science. *Journal of Consciousness Studies, 10*(9–10), 141–156.

Piccinini, G. (2009). First person data, publicity and self-measurement. *Philosopher's Imprint, 9*(9), 14–16.

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences, 3*(1), 1–23.

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition, 19*(4), 1069–1078.

Schooler, J., & Schreiber, C. A. (2004). Experience, meta-consciousness, and the paradox of introspection. *Journal of Consciousness Studies, 11*(7–8), 17–39.

Schwitzgebel, E. (2002a). How well do we know our own conscious experience? The case of visual imagery. *Journal of Consciousness Studies, 9*(5–6), 35–53.

Schwitzgebel, E. (2002b). Why did we think we dreamed in black and white? *Studies in History and Philosophy of Science Part A, 33*(4), 649–660.

Schwitzgebel, E. (2007). Do you have constant tactile experience of your feet in your shoes?: Or is experience limited to what's in attention? *Journal of Consciousness Studies, 14*(3), 5–35.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review, 117*(2), 245–273.

Schwitzgebel, E. (2011). *Perplexities of consciousness*. Cambridge: MIT Press.

Schwitzgebel, E. (2012). Introspection, what? In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness* (pp. 29–48). Oxford: Oxford University Press.

Schwitzgebel, E. (2013). Reply to Kriegel, Smithies, and Spener. *Philosophical Studies, 165*(3), 1195.

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science, 15*(11), 720–728.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences, 12*(8), 314–321.

Spener, M. (2013). Moderate scepticism about introspection. *Philosophical Studies, 165*(3), 1187.

Spener, M. (2015). Calibrating introspection. *Philosophical Issues, 25*(1), 300–321.

Swets, J. A., & Green, D. M. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Timmermans, B., & Cleeremans, A. (2015). How can we measure awareness? An overview of current methods. In M. Overgaard (Ed.), *Behavioural methods in consciousness research* (pp. 21–46). Oxford: Oxford University Press.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies, 3*(4), 330–349.

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition, 27,* 109–120.