Atten Percept Psychophys (2013) 75:1406–1426
DOI 10.3758/s13414-013-0505-2

Being confident without seeing: What subjective measures of visual consciousness are about

Michael Zehetleitner · Manuel Rausch

Published online: 19 July 2013

Abstract Can observers be confident about the accuracy of a discrimination response without a visual experience of the stimulus? In a series of five experiments, observers performed a masked orientation discrimination task, a masked shape discrimination task, or a random-dot motion discrimination task, followed by two subjective ratings after each trial, in which participants reported either their visual experience of the stimulus or their confidence in being correct. We observed that the threshold for ratings of the perception of the stimulus was above the threshold for ratings of a decision, that decision ratings outperformed stimulus ratings in predicting trial accuracy, and that different decision-related scales were more strongly associated with other decision-related scales than with ratings of stimulus clarity. We propose a taxonomy of subjective measures of consciousness that differentiates between subjective measures relating to the percept of the stimulus and measures relating to a discrimination decision and discuss the relation to type II blindsight.

Keywords Consciousness · Visual awareness · Confidence rating · Subjective experience · Masking · Heterophenomenology · Decision making · Signal detection theory

Introduction

The quest for neural correlates of consciousness relies typically on a comparison between two different types of measurements: those of neuronal processes and those of consciousness (Block, 2005; Crick & Koch, 1990; Rees, Kreiman, & Koch, 2002). This approach critically relies on defining measures of consciousness, which presents a huge obstacle in empirical science (Chalmers, 1998). With respect to measures of consciousness, several operationalizations are currently proposed in the literature.

Objective versus subjective measures

One prominent view distinguishes between objective measures and subjective measures (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). Measures of consciousness are considered objective if the participant's state of awareness is determined on the basis of his or her performance on a task. For example, it is often assumed that if observers are able to discriminate a stimulus or respond differentially to it, they are conscious of that stimulus (Erikson, 1960; Schmidt & Vorberg, 2006). When a participant performs at chance level on a discrimination task, this is typically considered a reliable indicator of the absence of conscious awareness of the presented stimuli (Hannula, Simons, & Cohen, 2005). Proponents of this view often make use of signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002), assuming that observers are conscious if their sensitivity in discriminating between signal and noise is above a predefined level (e.g., above zero).

A second approach to operationalizing consciousness is based on subjective measures. It has been questioned whether subjective measures are an acceptable method for empirical science at all (Hannula et al., 2005)—for example, because they might be corrupted by uncontrolled changes of the response criterion (Schmidt & Vorberg, 2006). By contrast, according to Daniel Dennett's heterophenomenology (Dennett, 2003, 2007), the participant's utterances about his or her experience should be considered as empirical raw data, which requires a scientific explanation. This means that the modulation of verbal reports in an experiment can be an object

Michael Zehetleitner and Manuel Rausch contributed equally to this work.

M. Zehetleitner (✉) · M. Rausch
Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstrasse 13, 80802 Munich, Germany
e-mail: mzehetleitner@psy.lmu.de

M. Rausch
Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Munich, Germany

of scientific study in the same way as other kinds of behavior, such as buttonpresses.

Several types of subjective measures are currently proposed in the literature. The most frequent measurements are confidence ratings: The participants indicate how confident they feel about the correctness of their response (Peirce & Jastrow, 1884). Another possibility is to ask participants about the reason why they chose a particular response alternative; for example, after a response is given, participants might attribute their response to guessing, intuition, memory, or knowledge (Dienes & Scott, 2005; Scott & Dienes, 2008). Also, recently, observers have been asked to place a wager on the accuracy of their decision, either with the possibility that the reward is lost if the wager is incorrect (Persaud, McLeod, & Cowey, 2007) or without the risk of losing the wager (Dienes & Seth, 2010). Since wagering is independent of speech, it has been used successfully to explore awareness in animals—specifically, in monkeys (Kornell, Son, & Terrace, 2007) and pigeons (Nakamura, Watanabe, Betsuyaku, & Fujita, 2011). A third approach asks the observers to make judgments directly about their visual experiences. For example, observers can be asked to rate the degree of visual experience evoked by a stimulus on a visual analogue rating scale (Del Cul, Baillet, & Dehaene, 2007; Sergent & Dehaene, 2004). Assessing the degree of visual experience as well, but avoiding the use of continuous scales, the Perceptual Awareness Scale (PAS) provides the participants with a discrete scale with verbal labels for each scale point to rate their visual experiences (Ramsøy & Overgaard, 2004).

Blindsight type 2 phenomena

A classical example held to support a dissociation between objective and subjective measures of consciousness is blindsight: After a unilateral lesion to V1, patients suffer from apparent blindness in the visual field contralateral to the lesion. Blindsight is defined as the ability of patients to discriminate visual stimuli presented in their seemingly blind visual field in forced choice tasks with remarkable accuracy, despite the fact that they report no visual experiences of these stimuli (Weiskrantz, 1986). The subjective reports of blindsight patients fall into two categories (Sahraie, Weiskrantz, Trevethan, Cruce, & Murray, 2002): blindsight type I and type II. In blindsight type I, patients report no awareness of the stimulus and very low confidence in discrimination choice, even though their choice is reliably above chance. However, the subjective reports of patients are apparently inconsistent in blindsight type II: These patients occasionally report a feeling or knowing that something happened in their blind visual field, although they insist that their experience was qualitatively different from normal seeing (Riddoch, 1917; Weiskrantz, Barbur, & Sahraie, 1995; Zeki & ffytche, 1998). Critically, these patients may report a considerable amount of confidence in two-alternative forced choice (2AFC) judgments (Sahraie, Weiskrantz, & Barbur, 1998) and even be willing to wager the same amount of money in the blind and in the intact hemifield when discrimination difficulty is matched (Persaud et al., 2011), although in these studies, no visual experience of the stimulus was reported at all.

A similar dissociation between subjective reports of confidence and visual experience has been reported when brain activity in the posterior cortex was only transiently disrupted via transcranial magnetic stimulation (TMS): Occipital TMS between 86 and 114 ms after the presentation of the stimulus suppressed reports of visual experience of the stimulus, although discrimination performance was still quite good (Boyer, Harrison, & Ro, 2005). Interestingly, confidence ratings were strongly correlated with the accuracy of the discrimination judgment, indicating that TMS affected the reports of subjective experience more than the reports of subjective confidence.

Stimulus ratings versus decision ratings

The discrepancy between subjective measures in type 2 blindsight and posterior TMS raises questions as to whether subjective measures of consciousness form one single category. In the present study, we propose a taxonomy of subjective measures of consciousness that differentiate between subjective measures relating to the percept of stimulus (*stimulus rating*) and measures relating to a discrimination decision (*decision rating*). In detail, we discuss whether stimulus and decision ratings (1) might relate to different events in terms of SDT, (2) can be interpreted as measures of different processes within the cognitive architecture, and (3) might be associated with different experiences from the first-person perspective.

First, it can be argued that the stimulus and decision ratings mirror a distinction in SDT between type 1 tasks and type 2 tasks (Galvin, Podd, Drga, & Whitmore, 2003). In SDT, the distinction between type 1 and type 2 tasks is based on the events about which an observer makes a discrimination decision. In type 1 tasks, the observer discriminates whether an event (a stimulus) is either signal or noise. The discrimination response of the observer can be considered as a new event, which can be either correct or incorrect. In SDT, type 2 tasks require the participant to make a judgment about whether the previous type 1 response was correct or incorrect. Subjective ratings can refer to the events of the type 1 task (e.g., when participants are asked to rate the clarity of their percept), but they can also refer to the events of the type 2 task (e.g., when participants give confidence

ratings). The mere wording of existing subjective measures suggests such a correspondence, since they semantically reference either to the stimulus or to the decision: "How clearly did you experience the *stimulus*?" or "how confident are you that your *decision* was correct?" Thus, it seems reasonable to assume that the events in the world that stimulus and decision ratings refer to are different.

Concerning the second point, it is possible to connect stimulus and decision ratings to different functions within the cognitive architecture. Nelson and Narens's model of metacognition distinguishes between two different levels of cognitive processing: On the one hand, there are processes concerned with performing the task, which they call the *object level*, and on the other hand, there are processes forming a dynamic model of the object level and giving rise to verbal reports, which they call the *meta-level* (Nelson & Narens, 1990). According to standard assumptions about processes on the object level, when an observer performs a visual discrimination task and a stimulus is presented, this stimulus first creates sensory data within the brain, which is integrated over time into a decision variable. A decision is selected by applying a decision rule to the decision variable, and the respective response is triggered (Gold & Shadlen, 2007; Ratcliff, 1978). When processes on the meta-level give rise to verbal reports about the stimulus or the decision, it is possible that both kinds of subjective reports are created by subsampling out of the same underlying dimension of sensory data. Another hypothesis might be that, when participants rate the clarity of their visual experience, they might estimate the strength or the quality of the internal signals that form part of the sensory data. In contrast, in confidence ratings or wagering, participants might evaluate those internal signals that are involved in the decision to make a response.

Third, stimulus and decision ratings are qualitatively different from the first-person perspective. When observers rate how clearly they perceived the stimulus, it seems to them that they judge their visual experience elicited by the presentation of the stimulus. This is different from the experience observers refer to when they give a decision rating: In this case, the first-person experience in question is, above all, a feeling of confidence in being correct or incorrect or, alternatively, a rational belief concerning the likelihood of being correct. For individuals, visual experience is not the primary referent of decision ratings, and likewise, a feeling of confidence is not the primary referent of stimulus ratings.

It should be noted that the distinction between stimulus and decision ratings proposed here overlaps with, but is not identical to, the distinction between introspective reports and metacognitive reports proposed by Overgaard and Sandberg (2012). They argued that introspective reports and metacognitive reports reveal different kinds of metacognitive access: Whereas introspective reports require participants to report their conscious experience directly, metacognitive reports are based on metacognitive judgments about a mental process (such as the selection of the task response), which is assumed to be dependent on introspection of one's conscious experience. In the view outlined in the present study, the relationship between stimulus ratings and decision ratings is symmetrical, in the sense that they are both based on a metacognitive judgment: When participants rate their percept of the stimulus, they evaluate cognitive processes involved in the representation of the stimulus. When participants rate their confidence in the discrimination judgments, they assess those processes involved in selecting one out of several task alternatives. However, both stimulus and decision ratings are associated with a certain subjective experience that is qualitatively different in both cases: in the first case, a visual experience of the stimulus; in the second case, a subjective feeling of being correct or incorrect.

In any case, since the cognitive functions of stimulus perception and decision making are closely connected, it is to be expected that the behavioral patterns of rating the stimulus and the decision are quite similar. The three lines of argumentation outlined above thus do not imply the prediction that both kinds of subjective reports contradict each other in a fundamental way but indicate the possibility of subtle differences.

To summarize, it is conceptually possible that ratings of visual experience can be sorted into one class of subjective measures, while confidence ratings, as well as wagering, belong to another class of subjective measures of consciousness. The two classes are probably not associated with fundamentally different behavioral patterns. At least in the case of disturbance of the occipital cortex, though, it has been demonstrated that the results obtained by the two classes are not identical. The present study aims to investigate whether there is empirical support for any dissociation between the two classes of decision-related and stimulus-related subjective reports in healthy human participants.

Evaluation criteria for subjective measures of consciousness

The selection of criteria to evaluate measurements of consciousness is nontrivial given the fact we cannot observe another person's consciousness from the third-person perspective (Nagel, 1974). As the extent to which a measurement "really" captures consciousness is impossible to determine, we will consider only three objective characteristics. Assuming that stimulus and decision ratings refer to different external events, the first relevant relationship is between the measures and properties of the stimulation. Specifically, measures might differ with respect to the relative sensitivity to changes of stimulus quality, as well as the thresholds they impose upon observers (analogous to SDT type 1 sensitivity

and criterion). The second relevant characteristic is their relation to the accuracy of the discrimination response. Again, measures might vary in their predictability for trial accuracy, as well as the response criterion (analogous to SDT type 2 sensitivity and criterion). According to the zero correlation criterion, an observer is assumed to be conscious if there is a positive correlation between his or her confidence ratings and task performance (Dienes, Altmann, Kwan, & Goode, 1995). This correlation can be assessed separately for each level of stimulation to determine the weakest level of stimulation with a positive correlation between the measure and trial. The third relevant property of subjective measures is their relation to other rating scales. Measures can vary in the degree to which their variance is specific to them or is shared by the other measures.

Empirical differences between subjective measures

Different subjective measures of consciousness have been previously compared with each other in two experiments with artificial grammar tasks and only one experiment with a visual discrimination task. Concerning artificial grammar tasks, one study compared confidence ratings and wagering, reporting that wagering is confounded by risk aversion, but no substantial differences between confidence and wagering occurred after the possibility of loss had been eliminated from wagering (Dienes & Seth, 2010). The second study reported that confidence ratings outperformed wagering and ratings of rule awareness in predicting trial accuracy and that confidence ratings imposed a more liberal criterion for ratings in terms of accuracy than did the other scales (Wierzchoń, Asanowicz, Paulewicz, & Cleeremans, 2012). Concerning the experiment with a visual paradigm, a masked object identification task, the PAS outperformed confidence ratings and wagering in predicting trial accuracy (Sandberg, Timmermans, Overgaard, & Cleeremans, 2010). By means of fitting psychometric functions to the data, the authors observed that the threshold in terms of stimulus duration for confidence ratings was below the threshold for the PAS. Furthermore, both the threshold for confidence and the PAS were below the threshold for wagering (Sandberg, Bibby, Timmermans, Cleeremans, & Overgaard, 2011).

Rationale of the present study

To summarize, the present study addressed two main research questions: First, we investigated whether the pattern of decisional confidence in absence of visual experience, as occasionally reported in blindsight patients, can also be found in healthy human observers. Second, we explored the hypothesis that subjective measures of consciousness fall into two categories, depending on whether these measures refer to the experience of the stimulus or to the correctness of a discrimination response.

To address these issues, we conducted a series of five experiments. In each experiment, observers performed a 2AFC discrimination task with varying levels of difficulty. Within each trial, participants were asked to give two out of four possible subjective ratings after their discrimination response. When rating the stimulus, participants reported their clarity of experience of explicitly stated features of the stimulus. When rating the decision, participants were instructed either to wager imaginary money on their decision, to express their confidence in being correct, or to give an attribution of choice rating whether their orientation discrimination judgment was based on a guess or on knowledge. In Experiments 1 and 2, participants performed a masked orientation discrimination task, followed by one stimulus rating and one decision rating (in Experiment 1) and two decision ratings (in Experiment 2). In Experiments 3 and 4, observers performed a masked shape discrimination task with a stimulus and a decision rating (in Experiment 3) and three different decision scales (in Experiment 4). Experiment 5 was conducted to compare stimulus and decision ratings in a motion discrimination task with random-dot kinematograms (RDK). We collected ratings with visual analogue rating scales (VARSs), because continuous scales might encourage participants to rely more on their intuition and less on verbal categorization, as discrete scales with verbal labels do. In addition, it has been suggested that VARSs are sensitive to gradual manipulations of target durations in masked discrimination tasks (Sergent & Dehaene, 2004). We manipulated the quality of stimulation by varying the stimulus onset asynchrony (SOA) between stimulus and mask in Experiments 1–4 and the proportion of dots moving coherently in one direction in Experiment 5, which allowed us to estimate psychometric functions relating the quality of stimulation with mean ratings. The slope of the psychometric functions quantifies the relative sensitivity of the scale to changes of stimulus quality, and the center of the function determines its threshold (Gescheider, 1997). In addition, we could test whether the zero correlation criterion was violated at each level of task difficulty by testing whether ratings in correct trials were higher than ratings in incorrect trials. After each single trial of the experiment, two ratings were presented; this procedure enabled us to assess the association of two different scale types on a single-trial basis. By using a hierarchical regression with random intercepts, we could, in addition, account for the clustered nature of the data across participants. In order to quantify the SDT type 2 characteristics of the different scales, we estimated receiver operating characteristics (ROCs) and determined sensitivity and response criterion on the basis of the area under the curve.

If subjective measures showed a similar pattern to type 2 blindsight, we hypothesized that stimulus ratings and decision ratings would exhibit different psychometric thresholds and different levels of difficulty where the zero correlation criterion was met: Decision ratings should have lower thresholds and should predict trial accuracy at a weaker level of stimulation. Second, concerning the classification of subjective measures into stimulus ratings and decision ratings, we predicted that the association of stimulus ratings and decision ratings would not be as close as the association of two different decision ratings. Third, since decision ratings, unlike stimulus ratings, refer primarily to trial accuracy, we predicted that decision ratings would exhibit a more pronounced SDT type 2 sensitivity than stimulus rating would. Decision ratings should only be more efficient in predicting trial accuracy, not stimulus quality; consequently, we expected that the psychometric slope of stimulus ratings would be at least the same as the psychometric slopes of decision ratings.

Experiment 1

Experiment 1 addressed the issue of comparing stimulus ratings against the three different decision-related scales. Observers performed a masked orientation discrimination task with varying SOAs between 10 and 140 ms. After each trial, observers submitted three responses: a 2AFC judgment about the orientation of the stimulus, a stimulus rating, and a decision rating. There were three different decision scales: Observers were asked to wager imaginary money on the correctness of their discrimination decision, to attribute the discrimination choice to a guess or to knowledge, or to give a confidence rating.

Method

Participants

Twenty participants (2 male, 2 left-handed) participated in the experiment. The age of the participants ranged from 19 to 29 years, with a median of 23. All reported having normal or corrected-to-normal vision, confirmed that that they did not suffer from epilepsy or seizures, and gave written informed consent. The experiment was conducted according to the principles expressed in the Declaration of Helsinki, 6th revision (World Medical Association, 2008), and the experimental procedure was approved by the ethics committee of the Department of Psychology of the Ludwig-Maximilians-Universität. Participants received either €8 per hour or course credits in return for participation.

Apparatus and stimuli

The experiment was performed in a sound-attenuated cabin with dim illumination to prevent reflections on the monitor. The stimuli were presented on a Diamond Pro 2070 SB (Mitsubishi) monitor with a 24-in. screen size and at a refresh rate of 100 Hz, driven by a PC with Windows XP as operating system. The viewing distance was approximately 80 cm. The experiment was programmed using MATLAB (MathWorks, U.S.) and Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). The target stimulus was a square filled with either a horizontal- or a vertical-oriented sinusoidal grating (frequency, 1 cycle per degree of visual angle; maximal luminance, 85.0 cd/m^2; minimal luminance, 9.5 cd/m^2), presented in front of a gray (12.5 cd/m^2) background. Squares subtended 3° × 3° degrees of visual angle. The mask consisted of a rectangular box (4° side length) with a black (1.3 cd/m^2) and white (85.0 cd/m^2) checkered pattern consisting of 6 × 6 equally sized squares. Both stimulus and mask were always presented at fixation. Concerning responses, participants performed the orientation discrimination judgment task by pressing "A" or "S" on the keyboard. When participants were presented with a rating, the corresponding question was displayed on the screen, with a continuous scale and labeled boundaries underneath, all colored black (1.3 cd/m^2). An index box was always initially located at the scale center. Participants used a Cyborg V1 joystick (Cyborg Gaming, U.K.) to move the index along the scale and to select a location on the scale. The question of the stimulus rating was always "how clearly did you see the grating?" with the anchors "unclear" and "clear." The three different decision scales were "how confident are you that your response was correct?" with the anchors "unsure" and "sure," "did you guess or did you know the response" with the anchors "guess" and "know," and finally, "how much money would you place as wager that you answer was correct?" with the anchors "€0" and "€20."

Trial structure

Each trial began with the presentation of a fixation cross at screen center for 1,000 ms. Then the target stimulus was presented for a brief period of time, until it was replaced by the mask. There were 10 possible SOAs between target and mask: 10, 20, 30, 40, 50, 60, 70, 90, 110, and 140 ms. In order to prevent participants from giving premature responses, there was a period of 600 ms after the onset of the mask when participants could not yet respond to the stimulus. After this delay period, participants gave a 2AFC judgment about the orientation of the sinusoidal grating of the target, while the mask remained on the screen. Immediately afterward, the first question appeared on the screen. Participants were always asked to deliver both a stimulus

rating and a decision rating after each single trial. The scale type of the decision ratings changed after three blocks in both sessions, with every scale being presented in each session and the sequence of scales being random. The sequence of whether the stimulus rating or a decision rating was asked first changed between sessions. When participants had given the first rating, they had to move the index back to the scale center, before the second rating was displayed on the screen. If the 2AFC orientation judgment had been erroneous, the trial ended with the display of "error" for 1,000 ms, before the next trial started (see Fig. 1).

Design and procedure

The experiment consisted of two sessions performed on two consecutive days at the same time of the day. For the orientation discrimination task, participants were instructed to perform the task as accurately as possible, to follow their intuition about the orientation if they had not seen the orientation, and to guess if they had no idea about the orientation. For the stimulus ratings, participants were told that the question "how clearly did you see the grating?" referred to the clarity of experience of the grating on the stimulus. For decision ratings, participants were told that the ratings referred to their previous orientation discrimination judgment. Furthermore, participants were instructed to give the two ratings as independently from each other as possible and to give their ratings as carefully and as accurately as possible. At the beginning of session one, participants performed 20 training trials to familiarize the participant with the task. Each session of the main experiment involved nine blocks with 40 trials each and took, on average, 45 min.

Analysis

All analyses were performed using R 2.12.2 (R Core Team, 2012). In order to assess the effect of asking a rating immediately after the trial or as a second rating after the trial, we did two separate ANOVAs with rating as dependent variable: one ANOVA with the factors sequence (whether the rating was first or second within the trial), scale type (stimulus rating vs. confidence vs. wagering vs. attribution of choice), and SOA (10–140); the other ANOVA with the factors timing, scale type, and trial accuracy (correct vs. false).

Psychometric functions To assess the relationship between stimulus and decision ratings and SOA, psychometric functions were fit on the data of each individual. Logistic functions were used because they produced slightly better fits than Weibull or error functions. Steepness, threshold, and upper and lower asymptotes were allowed to vary as free parameters, leading to the following formula:

$$f(x) = \delta + (1-\delta-\gamma)\frac{1}{1+e^{-\beta(x-\theta)}},$$

where β denotes the steepness of the function, γ indicates its upper asymptote, δ denotes its lower asymptote, x the logarithm of the SOA, and θ the threshold. The parameter sets of stimulus ratings and decision ratings were compared with two-tailed paired t-tests.

SDT type 2 analysis ROCs were constructed separately for each individual and for stimulus and decision ratings. For this reason, the rating data of each individual were divided into nine bins. ROC curves were obtained by plotting the cumulative frequencies for ratings in each interval for incorrect trials on the x-axis and for correct trials on the y-axis. Measures of SDT type 2 sensitivity (A_{roc}) and response bias (B_{roc}) were computed on the basis of formulae provided by Fleming, Weil, Nagy, Dolan, and Rees (2010) and Kornbrot (2006). One individual was excluded from SDT type 2 analysis because he or she was extremely reluctant in wagering, rating on average 2 standard deviations below the mean rating over all observers.

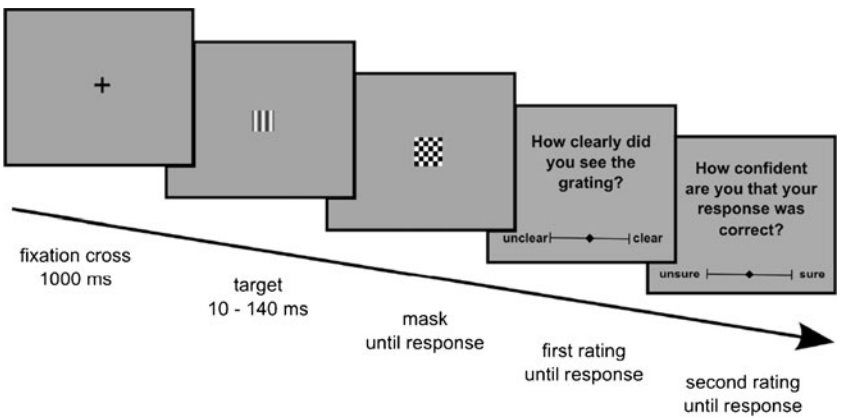

Fig. 1 Trial sequence in Experiments 1–4

In addition, to evaluate the zero correlation criterion, a series of one-tailed paired *t*-tests were computed separately for decision ratings and stimulus ratings and each SOA, assessing whether ratings were higher for correct trials than for incorrect trials. To avoid alpha error inflation, *p*-values were adjusted according to the Holm correction.

The relationship between stimulus ratings and each different type of decision rating was assessed by fitting a hierarchical linear model for each decision scale using nlme-package for R (Pinheiro, Bates, DebRoy, Sarkar, & the R Development Core Team, 2012), with decision rating as the dependent variable, SOA and stimulus rating as fixed factors, and a random intercept for each participant.

Results

Timing effects

The mixed ANOVA revealed significant effects of SOA, $F(9, 171) = 220.1$, $p < .001$, $\eta_G^2 = .81$, and scale type, $F(3, 57) = 6.8$, $p < .001$, $\eta_G^2 = .09$, as well as an interaction between these two, $F(27, 513) = 2.8$, $p < .05$, $\eta_G^2 = .02$. There was no effect of sequence and no interaction of sequence with SOA or scale type, all Fs < 1. The second ANOVA yielded significant effects of trial accuracy, $F(1, 19) = 180.4$, $p < .001$, $\eta_G^2 = .78$, scale type, $F(3, 57) = 5.4$, $p < .01$, $\eta_G^2 = .09$, as well as an interaction, $F(3, 57) = 7.0$, $p < .001$, $\eta_G^2 = .02$. Critically, there was again no effect of sequence, and no interaction of sequence with any of the other factors, all Fs < 1. Given these results, all subsequent analyses were performed without distinguishing between the first and second ratings.

Descriptive statistics

The mean error frequency in the discrimination task was .17 (*SD* = .08) and ranged from .41 for the shortest SOA to .01 for the longest SOA. Across the complete experiment, stimulus ratings averaged 46.8 % of the scale range (*SD* = 10.0). For the decision ratings, the mean rating was 55.0 % (*SD* = 12.2) for confidence, 50.4 % (*SD* = 16.8) for wagering, and 57.9 % (*SD* = 10.1) for attribution of choice ratings.

Psychometric functions

Within-subjects ANOVAs revealed that there were no significant differences between the three decision ratings in terms of threshold, $F(2, 38) = 1.2$, n. s., and slope, $F < 1$. Therefore, the rating data were pooled across different decision rating scales. An estimation of psychometric functions on stimulus ratings aggregated across participants revealed a threshold of 4.05 (*SE* = 0.09), a slope of 2.81 (*SE* = 0.64), a lower asymptote of .10 (*SE* = .3), and an upper asymptote of .10 (*SE* = .07). For decision ratings, the threshold was 3.93 (*SE* = 0.06), the slope 2.84 (*SE* = 0.54), the lower asymptote .10 (*SE* = .03), and the upper asymptote .03 (*SE* = .05; see Fig. 2). Paired *t*-tests on coefficients estimated on the level of each individual revealed that the threshold for decision ratings was lower than the threshold for stimulus ratings, $t(19) = 2.2$, $p < .05$, $d = 0.45$, and the upper asymptote was higher for decision ratings than for stimulus ratings, $t(19) = 2.6$, $p < .05$, $d = 0.61$. However, there were no significant differences in terms of slope, $t(19) = 1.6$, n. s., as well as lower asymptote, $t(19) = 0.8$, n.s.

SDT type 2 analysis

The data were again pooled across different decision scales, since a within-subjects ANOVA suggested there was no significant difference between the three decision ratings in terms of A_{roc}, $F(2, 38) < 1$, and B_{roc}, $F(2, 38) = 3.1$, n.s. Figure 3 displays the ROC curves for stimulus and decision ratings for the whole sample. The mean type 2 sensitivity as quantified by A_{roc} was .79 for decision ratings (*SD* = .07) and .78 for stimulus ratings (*SD* = .07). Paired *t*-tests revealed that the difference A_{roc} between stimulus ratings and decision ratings was significant, $t(18) = 2.4$, $p < .05$, $d = 0.20$. The mean type 2 criterion (B_{roc}) was −.15 (*SD* = .73) for decision ratings and −.74 (*SD* = .82) for stimulus ratings. The difference between stimulus and decision ratings in terms of B_{roc} was significant as well, $t(18) = 4.2$, $p < .001$, $d = 1.03$.

Zero correlation criterion

Multiple paired *t*-tests suggested that decision ratings on correct trials were always greater than decision ratings on

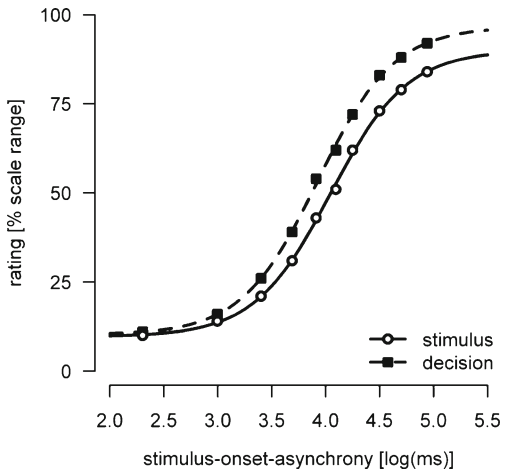

Fig. 2 Estimated logistic functions for stimulus ratings and decision ratings. Points indicate the averaged ratings for each SOA, the solid line indicates the estimated psychometric function for stimulus ratings, and the dashed line the estimated psychometric function for decision ratings

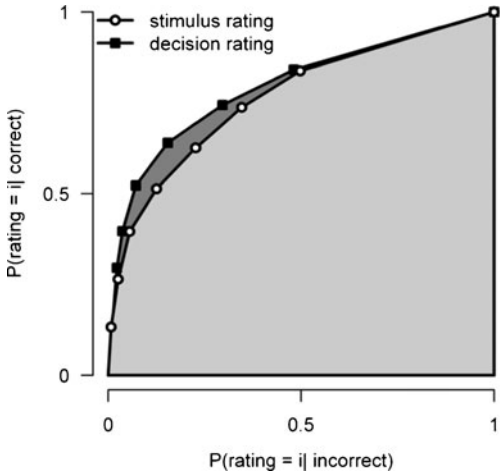

Fig. 3 Receiver operating characteristics. On the *x*-axis, there is the cumulative probability of each rating bin given that the trial was incorrect; on the *y*-axis, there is the cumulative probability for each rating given that the rating was correct. The area under the curve is used to determine the SDT type 2 sensitivity. White circles indicate binned stimulus ratings, black squares binned decision ratings

incorrect trials at each single SOA, all p_{cor}s < .05. By contrast, stimulus ratings were not significantly greater on correct trials than on incorrect trials at the shortest SOA, $t(19) = 0.89$, n.s. Significant results were obtained only for SOAs of 20 ms, $p_{cor} < .05$, and 40–90 ms, p_{cor}s < .05. In addition, for 9 out of 10 SOAs, the effect sizes of decision ratings as indexed by Cohen's *d* were greater for decision ratings than for stimulus ratings (see Table 1).

Within-trial regression

The hierarchical linear regressions revealed that for each scale type, decision ratings predicted stimulus ratings. The regression coefficients were .61, *SE* = .01, t (4770) = 51.8, $p < .001$, when stimulus ratings predicted confidence ratings, .64, *SE* = .01, $t(4770) = 58.6$, $p < .001$, when stimulus ratings predicted attribution of choice ratings, and .67, *SE* = .01, $t(4770) = 59.7$, $p < .001$, when stimulus ratings predicted wagering.

Discussion

Experiment 1 addressed the issue of whether subjective measures of consciousness show different properties depending on whether they refer to the stimulus or whether they refer to the decision. In addition, it was investigated whether the pattern of high confidence in absence of visual experiences known from blindsight patients can also be observed in normal participants.

We compared stimulus and decision ratings with respect to their psychometric functions, the zero correlation criterion at different SOAs, and SDT type 2 characteristics. It was observed that decision ratings were associated with a lower psychometric threshold than were stimulus ratings. We did not observe a substantial difference in the psychometric slope of stimulus and decision ratings, indicating that both types of ratings had comparable relative sensitivities to changes in the quality of stimulation. Concerning the analysis of zero correlation criterion, decision ratings were greater on correct trials than on incorrect trials for each SOA, while for stimulus ratings, the difference was not significant at SOAs of 10 and 30 ms. In addition, the effect sizes of the zero correlation criterion analysis were greater for decision ratings than for stimulus ratings at 9 out of 10 SOAs. Regarding SDT type 2 measures, decision ratings significantly outperformed stimulus ratings in predicting trial accuracy and imposed a considerably less conservative response criterion.

These results resemble to some degree the data pattern of subjective measures obtained in type 2 blindsight. Under

Table 1 *t*-tests comparing ratings in correct versus incorrect trials, separately for stimulus and decision ratings and each stimulus onset asynchrony (SOA)

SOA	*df*	Stimulus Ratings			Decision Ratings		
		t	p_{cor}	*d*	*t*	p_{cor}	*d*
10	19	0.9	n.s.	0.1	2.0	<.05	0.2
20	19	2.8	<.05	0.3	3.2	<.05	0.4
30	19	1.6	n.s.	0.3	2.4	<.05	0.4
40	19	2.7	<.05	0.5	3.2	<.05	0.8
50	17	4.3	<.01	1.2	4.3	<.01	1.4
60	18	3.9	<.01	1.1	4.8	<.001	1.3
70	14	4.4	<.01	1.4	4.2	<.01	1.3
90	13	4.4	<.01	1.5	7.0	<.001	3.4
110	9	2.4	n.s.	0.9	3.2	<.05	1.3
140	6	2.1	n.s.	1.1	3.1	<.05	1.6

certain stimulus conditions, these patients express a high degree of confidence in their decisions, although they report no visual experience (Persaud et al., 2011; Sahraie et al., 1998). In line with this, observers in the present experiment also exhibited higher thresholds for reporting visual experience than for reporting confidence. These data seem to suggest that a weaker level of stimulation is needed to elicit confidence in the decision than to elicit a visual experience of the stimulus in both blindsight patients and healthy participants.

A potential concern with the data presented here is the fact that our procedure of presenting two ratings after each trial might have biased the ratings. In particular, models that assume that ratings are formed by a stochastic diffusion process might predict the second rating to be higher or more accurate because there is more time for the sensory evidence to accumulate (Pleskac & Busemeyer, 2010). In the present study, we found no evidence that the sequence of ratings influenced the ratings directly or interacted with scale type, SOA, or trial accuracy. We cannot rule out the possibility that the procedure of asking two ratings after each trial might have influenced both of the two ratings—for example, if two contradicting ratings caused cognitive dissonance or if participants understood the instruction to give two ratings after each trial in such a way that they felt that the two ratings had to be different. However, if this was the case, the bias would affect both stimulus and decision ratings to the same extent and cannot account for the threshold offset between stimulus and decision ratings or for the difference in SDT type 2 sensitivity.

Experiment 2

Experiment 2 was designed to investigate the relationship between different subjective measures referring to the discrimination judgment. Observers performed the same discrimination task as in Experiment 1, except that each trial was followed by two out of three possible decision-related scales. Observers were asked how much money they would wager that the orientation discrimination was correct, were asked to report whether their orientation choice was based on a guess or on knowledge, or were asked to give a confidence rating.

Method

Participants

Twenty participants (6 male, 1 left-handed) participated in Experiment 2. The age of the participants ranged from 20 to 40 years, with a median age of 27. All participants reported having normal or corrected-to-normal vision, confirmed that that they did not suffer from epilepsy or seizures, and gave written informed consent. Participants received either €8 per hour or course credits in return for participation.

Apparatus, stimuli, design, and procedure

The apparatus, stimuli, design and procedure were identical to those in Experiment 1.

Trial sequence

The trials were identical to those in Experiment 1, except that instead of asking one stimulus rating and one decision rating after each trial, there were always two out of the three possible decision ratings. Each combination of ratings was presented for three blocks. The sequence of ratings was randomized and was opposite for the consecutive session.

Analysis

To ensure comparability between Experiments 1 and 2, the same analysis was performed for Experiment 2 as for Experiment 1, except that instead of comparing stimulus ratings against decision ratings, the three different decision-related scales—confidence, attribution of choice, and wagering—were compared against each other by an ANOVA with scale type (confidence vs. wagering vs. attribution) as a within-subjects factor. Significant main effects of scale type were further examined by two-sided *t*-tests with *p*-values adjusted according to the Holm correction. One participant was removed from the analysis of psychometric functions because her ratings were insensitive to varying SOA and the corresponding psychometric functions would have been parallel to the horizontal. Another participant was removed from the SDT type 2 analysis because his or her response criterion for all three scales was extremely conservative, so the B_{roc} value could not be computed.

Results

Descriptive statistics

On average, the proportion of erroneous trials in Experiment 2 was .16 (*SD* = .08). On average, observers gave confidence ratings of 63.1 % of the scale range (*SD* = 11.2), attribution of choice ratings of 65.1 % (SD = 10.6), and mean wagers of 59.1 % (*SD* = 12.9).

Psychometric functions

Figure 4 displays observed data and estimated psychometric functions for each scale type for the aggregated data. Comparing the parameters derived from the different scales, a within-subjects ANOVA revealed that there was a main effect of scale type on thresholds, $F(2, 36) = 5.6$, $p < .01$, $\eta_G^2 = .02$, and lower asymptote, $F(2, 36) = 6.8$, $p < .01$, $\eta_G^2 = .03$, but there were no effects on slope, $F(2, 36) = 1.1$, n.s., and upper

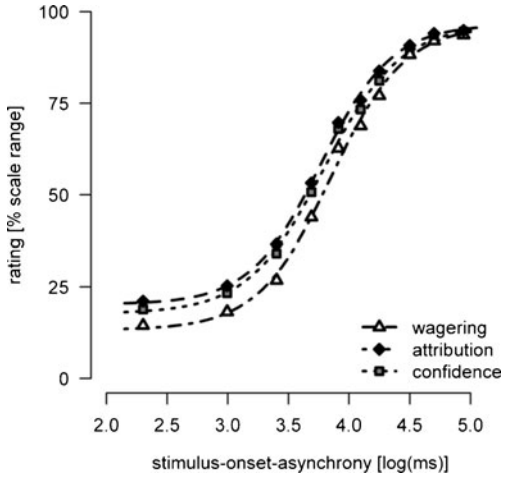

Fig. 4 Estimated functions for confidence ratings, attribution-of-choice ratings, and wagering. Squares indicate mean confidence ratings for each SOA, diamonds indicate attribution-of-choice ratings, and triangles indicate wagering. Separate lines indicate the estimated psychometric curves

asymptote, $F < 1$. Post hoc *t*-tests revealed that the threshold for wagering was above the threshold for attribution of choice, $t(18) = 2.7$, $p < .05$, $d = 0.35$, and for confidence as well, $t(18) = 3.6$, $p < .01$, $d = 0.22$, but there was no difference between thresholds for confidence and attribution of choice, $t(18) = 1.1$, n.s. For the lower asymptotes, post hoc comparisons suggested a significant difference between attribution-of-choice ratings and wagering, $t(18) = 3.0$, $p < .05$, $d = 0.41$, but there were no significant differences between attribution of choice and confidence, $t(18) = 1.5$, n.s., and between and between wagering and confidence, $t(18) = 2.5$, n.s.

SDT type 2 analysis

Figure 5 displays the ROC curves for the three different scales averaged across participants. The mean type 2 sensitivity as quantified by A_{roc} was .79 for confidence ($SD = .07$) and .80 for wagering ($SD = .06$) and attribution of choice ($SD = .05$). The main effect of scale type on A_{roc} was not significant, $F < 1$. The mean type 2 criterion (B_{roc}) was −0.94 ($SD = 0.62$) for confidence ratings, −1.05 ($SD = 0.39$) for attribution-of-choice ratings, and −0.92 ($SD = 0.55$) for wagering. There was no significant effect of scale type on B_{roc}, $F(2, 34) = 1.3$, n.s.

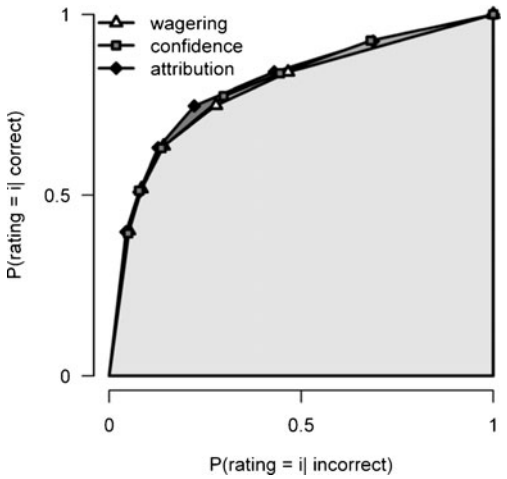

Fig. 5 Receiver operating characteristics in Experiment 2. The area under each curve indicates SDT type 2 sensitivity

Zero correlation criterion

As Table 2 shows, trial correctness predicted ratings in all three scale types starting with an SOA of 20 ms, all *p*s < .05. At the shortest SOA of 10 ms, only wagering differentiated between correct and incorrect trials, $t(19) = 2.6$, $p < .05$, but attribution-of-choice ratings did not, $t(19) = 0.6$, n.s., as well as confidence ratings, $t(19) = 0.7$, n.s. Effect sizes varied inconsistently between the different scales at different SOAs (see Table 2).

Within-trial regression

The hierarchical linear regressions revealed that ratings of each scale type could be predicted by ratings of each other scale type. The regression coefficients for wagering predicting attribution-of-choice ratings were .76, $SE = .01$, $t(4770) = 82.3$, $p < .001$, for wagering predicting confidence ratings .85, $SE = .01$, $t(4770) = 97.6$, $p < .001$, and for attribution of choice predicting confidence ratings .79, $SE = .01$, $t(4770) = 89.4$, $p < .001$.

Discussion

Experiment 2 was conducted in order to investigate the relationship between three decision-related subjective measures: confidence ratings, attribution-of-choice ratings, and wagering in terms of psychometric functions, SDT type 2 properties, zero correlation criterion, and within-trial regressions. Regarding psychometric functions, we observed no difference between the three scales in terms of slope, but the threshold for wagering was significantly above the threshold for confidence ratings and attribution-of-choice ratings. With respect to the ROC analysis, we found no significant differences regarding either SDT type 2 sensitivity or response criterion. Concerning the zero correlation criterion, the effects seemed to vary unsystematically between scales, with each scale being predicted by trial accuracy more efficiently at several SOAs. Concerning the association between the different types of ratings, we observed that all three scales were effective in predicting the other scale. Critically, the association of two different decision ratings in Experiment 2 seemed to be stronger than the association of decision ratings with stimulus ratings as observed in Experiment 1.

To summarize, Experiment 2 revealed a considerable amount of similar empirical properties of confidence ratings,

Table 2 Results of multiple *t*-tests comparing ratings on correct and incorrect trials in Experiment 2, separately for each different scale

SOA	*df*	Attribution of Choice			Wagering			Confidence		
		t	p_{cor}	*d*	*t*	p_{cor}	*d*	*t*	p_{cor}	*d*
10	19	0.6	n.s.	0.0	2.6	<.05	0.2	0.7	n.s.	0.1
20	19	4.1	<.01	0.6	2.7	<.05	0.4	3.0	<.05	0.3
30	19	5.3	<.001	0.7	4.8	<.001	0.7	5.8	<.001	1.0
40	19	3.6	<.01	0.9	5.3	<.001	1.3	4.5	<.001	1.3
50	15	3.6	<.01	1.1	3.5	<.01	1.0	2.7	<.01	0.8
60	17	5.1	<.001	1.4	4.3	<.01	1.4	4.6	<.001	1.4
70	13	6.5	<.001	2.7	4.1	<.01	1.4	5.5	<.001	1.9
90	11	5.8	<.001	2.2	4.0	<.01	1.5	4.7	<.001	1.9
110	9	4.0	<.01	1.8	4.7	<.01	2.7	6.0	<.001	3.0
140	7	3.6	<.01	1.7	4.2	<.01	1.7	3.5	<.01	1.8

attribution-of-choice ratings, and wagering, which is consistent with the view that all three scales belong to the same class of subjective measures of consciousness. Contradicting this view, the threshold for wagering was more conservative than that for the other two ratings. A potential explanation for this finding is that wagering not only is a measure of the cognitive processes involved in the discrimination task, but also might be biased by loss aversion (Fleming & Dolan, 2010) or risk aversion (Dienes & Seth, 2010). Presumably, risk aversion might influence wagering with imaginary money, although there was no objective risk of losing reward in the present experiment. We will resume the discussion of a distinct group of decision ratings after Experiment 4.

Experiment 3

Experiment 3 investigated whether the differences between stimulus and decision ratings as observed in Experiment 1 generalize to a masked object discrimination task. After each trial, observers indicated how clearly they had experienced the shape of the stimulus (instead of the orientation of its grating, as in Experiment 1) and how confident they felt about the accuracy of their discrimination choice.

Method

Participants

Sixteen participants (2 male, 1 left-handed) participated in Experiment 3. The age of the participants ranged from 19 to 26 years, with a median of 22. All participants reported having normal or corrected-to-normal vision, confirmed that that they did not suffer from epilepsy or seizures, and gave written informed consent. Participants received either €8 per hour or course credits in return for participation.

Apparatus and stimuli

The apparatus was the same as in Experiments 1 and 2, expect that the refresh rate was increased to 120 Hz. The target stimulus was either a square or a circle filled with either a horizontal- or a vertical-oriented sinusoidal grating (frequency, 1 cycle per degree of visual angle; maximal luminance, 85.0 cd/m^2; minimal luminance, 9.5 cd/m^2), presented in front of a gray (12.5 cd/m^2) background. Squares and circles subtended 3° × 3° of visual angle. Mask and rating scales were identical to those in Experiment 1.

Trial structure

The trial structure was the same as that in the previous experiments, except that SOAs of 8.3, 16.7, 25.0, 33.3, 50.0, 66.7, 83.3, and 116.7 ms were used. After onset of the mask and an additional delay period of 600 ms, participants gave a 2AFC judgment about the global shape of the stimulus by pressing "A" or "S" on the keyboard. After the discrimination response was given, two subjective ratings were presented on the screen, which were "How clearly did you perceive the shape?" with the anchors "unclear" and "clear" and "how confident are you that your response was correct?" with the anchors being "unsure" and "sure." Answers were collected via VARS. If the shape judgment was wrong, the trial ended with "error" displayed on the screen for 1,000 ms.

Design and procedure

Experiment 3 involved one session of approximately 1 h. Participants were instructed to prioritize accuracy over speed during the shape discrimination task. For verbal reports, it was ensured that participants understood that the stimulus

rating referred to their experience of the shape and the decision rating referred to their confidence in having discriminated the stimulus shape correctly. Again, participants were instructed to give the two ratings as independently from each other as possible and to give their ratings as carefully and as accurately as possible. At the beginning of the experiment, participants performed a training of 16 trials. Overall, the experiment comprised 12 blocks with 40 trials each.

Analysis

The analysis was the same as in Experiments 1 and 2. One participant was excluded from the analysis of psychometric functions because she gave the same subjective reports across all levels of difficulty, so no function fits could be obtained.

Results

Descriptive statistics

The mean error frequency in Experiment 3 was .23 (*SD* = .05). On average, observers gave a stimulus rating of 41.1 % (*SD* = 12.9) and a decision rating of 52.2 % (*SD* = 14.0).

Psychometric functions

Paired *t*-tests performed on individual parameters suggested that the decision ratings were associated with lower thresholds than were stimulus ratings, $t(14) = 2.0$, p(one-tailed) $< .05$, $d = 0.42$ (see Fig. 6a). In addition, we observed a marginal difference of lower asymptotes, $t(14) = 2.1$, $p = .06$, $d = 0.52$, but no difference between slopes, $t(14) = 1.5$, n.s., or upper asymptotes, $t(14) = 0.8$, n.s.

SDT type 2 analysis

Analysis of the SDT type 2 sensitivity resulted in mean A_{roc} of .77 (*SD* = .08) for stimulus ratings and mean A_{roc} of .78 (*SD* = .08) for decision ratings. For the response criterion, B_{roc} was −0.93 (*SD* = 1.19) for stimulus ratings and −0.39 (*SD* = 0.78) for decision ratings. Paired *t*-tests suggested that there was no significant difference between A_{roc}, $t(15) = 0.9$, n.s. (see Fig. 6b), but the response criterion of decision ratings was more liberal, $t(15) = 2.6$, $p < .05$, $d = 0.61$.

Zero correlation criterion

Multiple *t*-tests suggested that both stimulus and decision ratings were greater on correct trials than on incorrect trials at SOAs of 25.0 ms or greater. At shorter SOAs, no significant effects were observed (see Table 3).

Within-trial regression

The hierarchical linear regressions revealed that decision ratings could efficiently be predicted by stimulus ratings. The regression coefficient was .79, *SE* = .01, $t(7400) = 104.7$, $p < .001$.

Discussion

Experiment 3 investigated whether a pattern of subjective reports similar to those for type II blindsight—that is, high ratings of confidence in combination with low ratings of visual experience—can be observed in a masked shape discrimination task. In addition, we predicted that stimulus ratings and decision ratings showed different characteristics in terms of psychometric functions, SDT type 2 measures, and shared variance within trials.

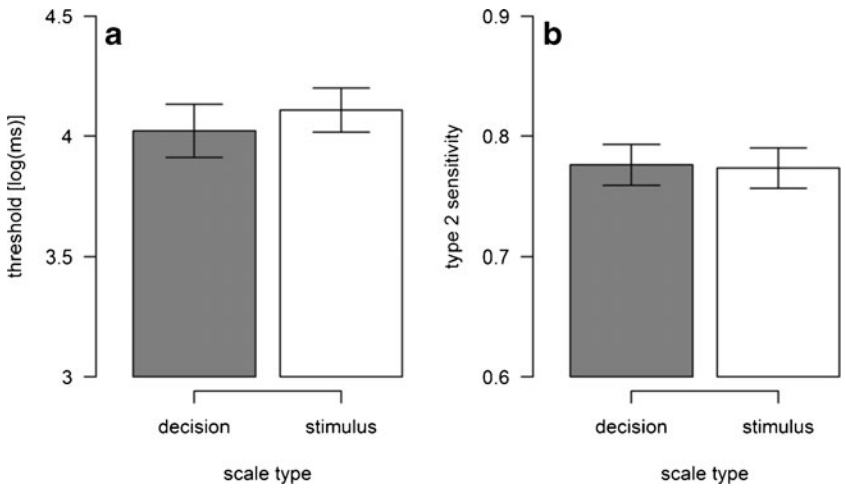

Fig. 6 Results of Experiment 3. **a** Mean thresholds derived from stimulus ratings and decision ratings. **b** Type 2 sensitivity of decision ratings and stimulus ratings

Table 3 Multiple *t*-tests comparing ratings on correct and incorrect trials in Experiment 3, separately for each different scale

SOA	Stimulus Ratings				Decision Ratings			
	t	*df*	p_{cor}	*d*	*t*	*df*	p_{cor}	*d*
8.3	0.4	15	n.s.	0.0	−0.4	15	n.s.	0.0
16.7	1.8	15	n.s.	0.1	1.2	15	n.s.	0.1
25.0	3.2	15	<.05	0.3	3.4	15	<.01	0.4
33.3	6.1	15	<.001	0.9	6.4	15	<.001	1.1
50.0	7.8	15	<.001	1.1	6.9	15	<.001	1.9
66.7	4.5	11	<.01	0.7	4.9	11	<.001	1.5
83.3	3.8	13	<.01	1.2	5.7	13	<.001	2.1
116.7	3.1	6	<.05	1.4	5.5	6	<.001	2.3

Regarding psychometric functions, we observed that the threshold of decision ratings was significantly higher than the threshold of stimulus ratings, albeit the relative sensitivity to changes of the stimulation was comparable. With respect to the SDT type 2 analysis, we observed that the response criterion induced by decision ratings was more liberal but there was no reliable difference in sensitivity. In contrast to our prediction, while decision ratings were associated with higher effect sizes than were stimulus ratings at longer SOAs, the patterns of the zero correlation criteria at short SOAs were the same.

In support of a type 2 blindsight-similar behavior of normal participants, observers in Experiment 3 had a higher threshold for decision ratings than for stimulus ratings, meaning that they would report confidence in being correct about the discrimination task already at a level of stimulation where their reports of visual experience were still low. The magnitude of this effect was nearly the same as in the orientation discrimination task, implying that the offset of psychometric curves derived by reports about the stimulus and reports about the decision is consistent across tasks.

Concerning the classification of subjective measures of consciousness into two classes, the results of Experiment 3 are more divergent than those of Experiment 1. We observed differences between stimulus and decision ratings in terms of thresholds and SDT type 2 criteria, indicating that observers are more conservative in reporting an experience of the stimulus than reporting confidence about a judgment. However, the difference between SDT sensitivity was not significant, and the patterns of the zero correlation criteria were the same. Consequently, at least for shape discrimination tasks, it seems to depend on the research question whether the distinction between stimulus and decision ratings is relevant. If the focus is on the correlation between subjective reports and objective performance (e.g., by zero correlation criteria), stimulus and decision ratings converge to the same results. In cases where criteria are more important (e.g., if it is determined whether a stimulus is above or below a subjective threshold), stimulus and decision ratings might lead to opposite conclusions.

Experiment 4

Experiment 4 was conducted to explore whether the lag of psychometric curves between wagering and the other decision-related scales generalizes to shape discrimination. Observers reported whether a masked target stimulus was either a square or a circle, followed by subjective reports about how confident they felt about their discrimination decision, whether they guessed or knew their discrimination response, or how much money they would place as a wager that their response was correct.

Method

Participants

Sixteen participants (6 male, 1 left-handed) participated in Experiment 4. The age of the participants ranged from 20 to 40 years, with a median age of 25. All participants reported having normal or corrected-to-normal vision, confirmed that that they did not suffer from epilepsy or seizures, and gave written informed consent. Participants received either €8 per hour or course credits in return for participation.

Apparatus and stimuli

The apparatus and stimuli were the same as those in Experiment 3, expect that the refresh rate was set to 160 Hz.

Trial structure

The trial structure was the same as in the previous experiments, except that SOAs of 6.25, 12.5, 18.75, 25.0, 31.25, 37.5, 50.0, 62.5, 75.0, 87.5, and 120.0 ms were used. After

onset of the mask and a delay period of 600 ms, participants gave a 2AFC judgment of whether the global shape of the stimulus was a square or a circle. After the discrimination response was given, two out of the three possible decision-related scales were presented on the screen.

Design, procedure, and analysis

Design, procedure, and analysis were the same as those in Experiment 2.

Results

Descriptive statistics

The mean error frequency in Experiment 4 was .26 (*SD* = .08). On average, observers gave a confidence rating of 51.1 % of the scale range (*SD* = 12.3), an attribution-of-choice rating of 51.9 % (*SD* = 10.6), and a 49.5 % (*SD* = 15.7).

Psychometric functions

Figure 7a displays mean psychometric thresholds of each scale in Experiment 4. A comparison of the estimated parameters via a within-subjects ANOVAs revealed no effects of scale type on thresholds, slopes, upper asymptotes, or lower asymptotes, all Fs < 1.

SDT type 2 analysis

The mean type 2 sensitivity as quantified by A_{roc} was .72 for confidence (*SD* = .09) and attribution of choice (*SD* = .08) and .71 for wagering (*SD* = .10). The main effect of scale type on A_{roc} was not significant, $F < 1$. The mean type 2 criterion (B_{roc}) was 0.22 (*SD* = 2.46) for confidence ratings, −0.17 (*SD* = 1.81) for attribution-of-choice ratings, and 0.05 (*SD* = 1.54) for wagering. There was no significant effect of scale type on B_{roc}, $F < 1$ (see Fig. 7b).

Zero correlation criterion

As is shown in Table 4, ratings were significantly larger on correct trials than on incorrect trials for all three scales at SOAs between 31.2 ms, all p_{cor}s < .05. At shorter SOAs, all t-tests were not significant.

Within-trial regression

The hierarchical linear regressions suggested that ratings of each scale type could be predicted by ratings of the other scale types. The regression coefficients were for wagering predicting attribution-of-choice ratings .91, *SE* = .01, $t(3813) = 119.6$, $p < .001$, for wagering predicting confidence ratings .92, *SE* = .01, $t(3813) = 131.5$, $p < .001$, and for attribution-of-choice predicting confidence .91, *SE* = .01, $t(3813) = 129.7$, $p < .001$.

Discussion

Experiment 4 investigated whether confidence ratings, attribution-of-choice ratings, and wagering form one coherent class of subjective measures of consciousness with respect to their psychometric functions, SDT type 2 characteristics, zero correlation criteria, and within-trial regressions. Specifically, it was examined whether a lag in thresholds between wagering and the other two scales, as observed in Experiment 2, also would emerge in the masked shape discrimination task.

An analysis of psychometric functions showed no difference between curves fitted on wagering, attribution-of-choice, and confidence data in terms of slopes and thresholds, just as there were no differences in terms of type 2 sensitivities and type 2 criteria. The zero correlation criterion was rejected starting at the same SOA on all scales, and within-trial regressions showed that the three scales shared their variance almost completely. In accordance with the classification of subjective measures as either decision ratings or stimulus ratings, the association between two different decision ratings in

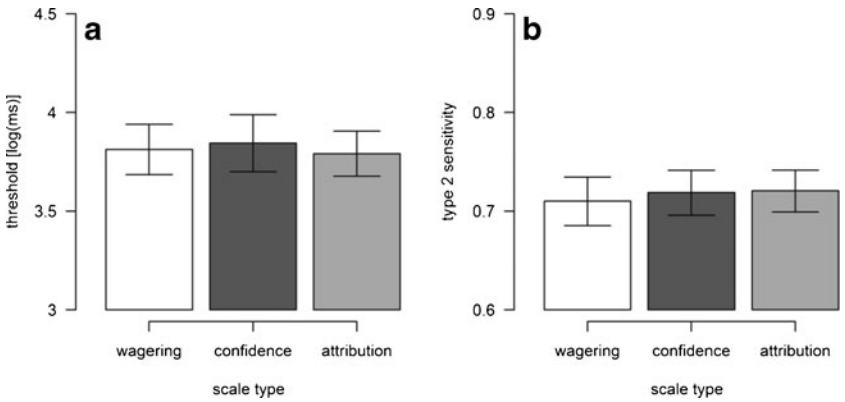

Fig. 7 Results of Experiment 4. **a** Thresholds for confidence ratings, attribution-of-choice ratings, and wagering. **b**: SDT type 2 sensitivities

Table 4 Multiple *t*-tests comparing ratings on correct and incorrect trials in Experiment 4, separately for each different scale

SOA	Attribution of Choice				Wagering				Confidence			
	df	*t*	p_{cor}	*d*	*df*	*t*	p_{cor}	*d*	*df*	*t*	p_{cor}	*d*
6.25	15	1.1	n.s.	0.0	15	0.7	n.s.	0.1	15	−0.7	n.s.	0.0
12.5	15	0.0	n.s.	0.1	15	−0.2	n.s.	0.0	15	0.2	n.s.	0.0
18.75	15	0.7	n.s.	0.0	15	0.6	n.s.	0.1	15	0.6	n.s.	0.1
25.0	15	2.2	n.s.	0.4	15	2.0	n.s.	0.5	15	1.9	n.s.	0.4
31.25	14	6.6	<.001	1.7	14	4.4	<.01	1.3	14	5.9	<.001	1.3
37.5	14	5.9	<.001	1.3	15	3.5	<.05	1.0	15	5.5	<.001	1.4
50.0	14	7.2	<.001	2.1	15	5.5	<.001	1.6	14	5.8	<.001	1.6
62.5	14	3.9	<.01	1.3	12	6.4	<.001	2.3	12	5.2	<.001	1.7
75.0	10	3.6	<.05	1.6	11	2.8	n.s.	1.3	10	2.1	n.s.	1.0
87.5	4	2.2	n.s.	1.7	4	1.4	<.01	0.9	4	3.5	n.s.	1.9
120.0	4	3.8	n.s.	1.7	4	10.5	n.s.	3.3	2	5.1	n.s.	2.8

Experiment 4 seemed to be stronger than the association between a stimulus rating and a decision rating in Experiment 3.

Overall, Experiments 1, 2, and 4 concurrently indicate that verbal reports that refer to the discrimination decision are very similar in their patterns in terms of within-trial regressions, psychometric slopes, and SDT type 2 characteristics. The only indication of a difference between measures, a lag of the psychometric threshold of wagering with respect to the other two scales, was observed only in Experiment 2 but did not replicate in Experiment 4. Thus, our experiments provide converging evidence that attribution-of-choice ratings, confidence ratings, and wagering form one coherent category of subjective measures of consciousness.

Experiment 5

In Experiments 1–4, sensory evidence was always manipulated by short presentation of the stimulus in conjunction with backward masking. In Experiment 5, we investigated whether the discrepancy between subjective reports of the stimulus and of the discrimination decision can be replicated when sensory evidence is varied by another manipulation—that is, the proportion of coherently moving dots of RDKs. After indicating the direction of motion of the coherently moving dots, observers delivered ratings both of the subjective clarity of motion and of confidence in the motion discrimination decision.

Method

Participants

Twenty-one participants (4 male, 2 left-handed) participated in the experiment. The age of the participants ranged from 19 to 40 years, with a median age of 22. All participants reported having normal or corrected-to-normal vision, confirmed that that they did not suffer from epilepsy or seizures, and gave written informed consent.

Apparatus and stimuli

The experiment was conducted in a sound-attenuated cabin, controlled by MATLAB and Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). Stimuli were presented on a Diamond Pro 2070SB at a refresh rate of 120 Hz driven by a Mac with OS X 10.7 as the operating system at a viewing distance of approximately 60 cm. The stimulus was a random dot kinematogram, consisting of small white squares (16.7 dots per square degree of visual angle, sized 2 × 2 pixels, luminance 78.5 cd/m^2) in front of a black background (0.0 cd/m^2), which appeared in a circular aperture (diameter: 5º) centered at the fixation. A set of dots was shown for one video frame and then replotted three video frames later. When replotted, a subset of dots was offset from their original location to create apparent motion, while the remaining dots were relocated randomly. The proportion of coherently moving dots was randomly chosen from among 0.7 %, 1.3 %, 2.7 %, 5.3 %, 10.7 %, 21.3 %, or 42.7 %. Dots moved horizontally to the left or to the right at a velocity of 4° per second. Participants responded to leftward and rightward motion by pressing the left and right arrow buttons on the keyboard. Subjective reports were collected in the same way as in the previous experiments. The stimulus rating was "How clearly did you see the coherent motion?" with the anchors "unclear" and "clear"; the decision rating was "how confident are you that your response was correct?" with the anchors "unsure" and "sure."

Trial structure

Each trial began with the presentation of a fixation cross at screen center for 1,000 ms. Then an RDK was presented until participants gave a 2AFC judgment about the direction of the random-dot motion. Immediately afterward, the first question appeared on the screen. Participants were always asked to deliver both a stimulus rating and a decision rating after each single trial, with the sequence of the two ratings counterbalanced across participants. If the 2AFC orientation judgment was erroneous, the trial ended with the display of "error" for 1,000 ms.

Design and procedure

Experiment 5 involved one session of 45 min on average. For the motion discrimination task, participants were instructed to prioritize accuracy over speed and to guess if they did not know the direction of motion. For verbal reports, it was ensured that participants understood that the stimulus rating referred to motion experience created by the coherently moving dots, and the decision rating referred to their confidence in having discriminated the motion direction correctly. Again, participants were instructed to give the two ratings as independently from each other as possible and to give their ratings as carefully and as accurately as possible. At the beginning of the experiment, participants performed a training block with 49 trials. The main experiment involved seven blocks with 49 trials each.

Analysis

The analysis was the same as that in previous experiments, except that it was performed with respect to levels of coherence rather than SOAs.

Results

Descriptive statistics

The mean error rate in Experiment 5 was .22 (SD = .53). On average, observers gave a confidence rating of 59.7 % of the scale range (SD = 11.0) and a stimulus rating of 52.0 % (SD = 12.6).

Psychometric functions

Two-tailed paired t-tests of the estimated parameters revealed that the offset of thresholds between stimulus ratings and decision ratings was significant, $t(20) = 4.0$, $p < .001$, $d = 0.73$ (see Fig. 8a); however, there was no difference between slopes, $t(20) = 1.3$, n.s., lower asymptotes, $t(20) = 2.0$, n.s., and upper asymptotes, $t(20) = 0.8$, n.s.

SDT type 2 analysis

For SDT type 2 sensitivity, the mean A_{roc} was .73 (SD = .05) for stimulus ratings, as compared with .74 (SD = .03) for decision ratings. Two-tailed paired t-tests suggested that the difference was significant, $t(20) = 2.2$, $p < .05$, $d = 0.41$ (see Fig. 8b). For the response criterion, B_{roc} was −0.63 (SD= 0.74) for stimulus ratings and 0.10 (SD = 1.0) for ratings of the decision. As well, t-tests suggested that B_{roc} was different between stimulus and decision ratings, $t(20) = 5.0$, $p < .001$, $d = .82$.

Zero correlation criterion

Table 5 shows overviews of t-tests performed between correct and erroneous trials at each level of coherence. Both stimulus and decision ratings were significantly different between correct and incorrect trials at a coherence of 2.7 %. At a coherence of 1.3 %, the effect of trial correctness on decision ratings was marginally significant, $t(20) = 1.3$, $p = .06$, $d = 0.2$, but could not be observed for stimulus ratings, $t(20) = 0.4$, n.s.

Within-trial regression

The hierarchical linear regressions suggested that decision ratings predicted stimulus ratings on a single-trial basis. The regression coefficient was .59, $SE = .01$, $t(7175) = 71.2$, $p < .001$.

Discussion

Experiment 5 was conducted to test whether the observed discrepancy between stimulus and decision ratings is specific to masking experiments or whether it generalizes to motion discrimination with random-dot motion kinematograms as well. We observed that the threshold for stimulus ratings required a higher proportion of coherently moving dots than did decision ratings, although the relative sensitivities of both kinds of ratings were not substantially different. In addition, we found that decision ratings outperformed stimulus ratings in predicting trial accuracy and were associated with a more liberal type 2 response criterion. Concerning the zero correlation criterion, decision ratings were marginally greater on correct trials than on incorrect trials at a coherence level of 1.3 %, while stimulus ratings were associated with trial accuracy at a coherence of at least 2.7 %. The magnitude of this effect was greater for decision ratings than for stimulus ratings for six out of seven levels of coherence. The association between stimulus and decision ratings was comparable to that in Experiment 1 and was considerably smaller than the association between confidence, wagering, and attribution-of-choice ratings in Experiments 2 and 4.

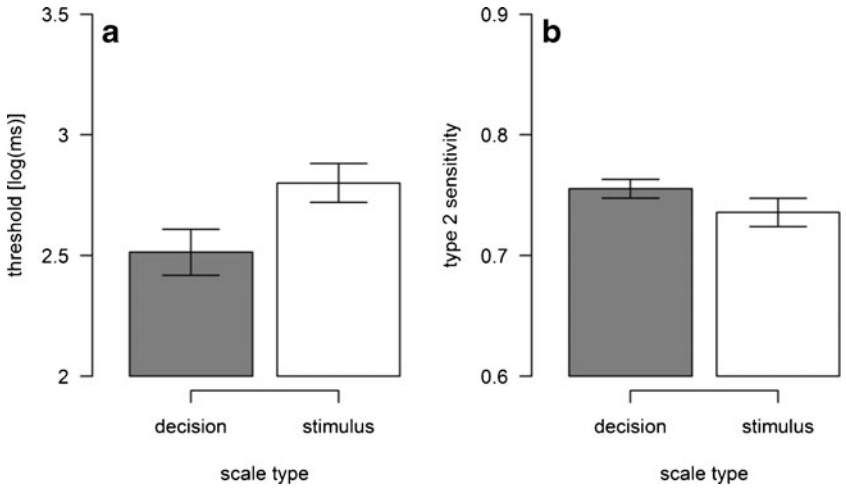

Fig. 8 Results of Experiment 5. **a** Thresholds derived from decision ratings and stimulus ratings. **b** Mean SDT type 2 sensitivities of stimulus ratings and decision ratings

Overall, the results of Experiment 5 support nicely the distinction between stimulus and decision ratings, which has thus been shown for masked orientation discrimination, shape discrimination, and random-dot motion discrimination.

General discussion

The five experiments presented here addressed two research questions. First, we investigated whether reports of high confidence and low visual experience, as is reported for type II blindsight, can be observed when healthy observers perform a masked orientation discrimination task. Second, we explored the hypothesis that subjective measures of consciousness can be sorted into two categories, depending on whether they refer to the stimulus or to the discrimination decision of the participant.

We compared ratings of the stimulus with ratings of the decision in a masked orientation discrimination task (Experiment 1), a masked shape discrimination task (Experiment 3), and a motion discrimination task (Experiment 5). Concerning psychometric functions, the thresholds of decision ratings were substantially lower than the thresholds of stimulus ratings in all three experiments, although the relative sensitivity to the quality of stimulation as indexed by psychometric slopes was comparable. With respect to SDT type 2 characteristics, decision ratings were associated with a more liberal response criterion in all experiments and a greater sensitivity in two out of three experiments. Concerning the zero correlation criterion, the results were more diverse: In Experiments 1 and 5, decision ratings were associated with correct trials at a lower level of stimulation despite the fact that the psychometric functions of both types of ratings had the same lower asymptote in both experiments. By contrast, in Experiment 3, we observed no differences in the zero correlation criterion at short SOAs.

Confidence ratings, attribution-of-choice ratings, and wagering were compared during a masked discrimination task with respect to orientation (Experiment 2) and shape (Experiment 4). Regarding psychometric functions, wagering was associated with a lower threshold than the other two scales in Experiment 2, but no differences appeared in

Table 5 Multiple *t*-tests comparing ratings on correct and incorrect trials in Experiment 5, separately for each different scale

Coherence	Stimulus Ratings				Decision Ratings			
	t	*df*	p_{cor}	*d*	*t*	*df*	p_{cor}	*d*
0.7	0.5	20	n.s.	0.0	1.7	20	n.s.	0.1
1.3	0.4	20	n.s.	0.0	2.2	20	n.s.	0.2
2.7	3.8	20	<.01	0.2	5.5	20	<.001	0.5
5.3	4.1	20	<.01	0.4	5.1	20	<.001	0.7
10.7	3.3	17	<.05	1.2	4.7	17	<.01	1.4
21.3	3.1	10	<.05	1.5	5.6	10	<.01	1.9
42.7	0.9	4	n.s.	0.0	−0.4	4	n.s.	−0.2

Experiment 4. All three scales had the same psychometric slopes, the same SDT type 2 sensitivity, and response criterion. In addition, the zero correlation criterion analysis revealed no systematic differences between the three scale types across different levels of stimulation.

In all five experiments, there was a considerable association between the two ratings that were required after each trial, indicating that the patterns of the ratings are quite similar. However, beyond that similarity, decision ratings were more efficient in predicting one of the other decision ratings in Experiments 2 and 4 than in predicting the stimulus ratings in Experiments 1, 3, and 5, suggesting there is a proportion of variance not shared between the two types of measures.

Type 2 blindsight in normal observers?

The present experiments might contribute to the theoretical interpretation of type 2 blindsight. In type 2 blindsight, patients report a feeling or some knowledge that something has happened in the visual field corresponding to the damaged V1 region (Sahraie et al., 2002). It has been reported that these patients can be very confident about discrimination decisions on stimuli presented in their blind visual field (Persaud et al., 2011; Sahraie et al., 1998). It has been proposed that blindsight in these patients is best understood as degraded conscious vision rather than preserved unconscious vision (Zeki & ffytche, 1998). In our data, the threshold for decision ratings was lower than that for stimulus ratings, meaning that participants reported confidence in the accuracy of their discrimination judgements at a lower level of stimulus quality than they reported experience of the stimulus. In addition, in Experiments 1 and 5, but not Experiment 3, decision ratings predicted trial accuracy at a weaker level of stimulation than stimulus ratings did. Although the discrepancy between report confidence and experience seems to be considerably stronger for blindsight patients, it seems as if our data show at least qualitatively the same pattern, indicating that confidence at a low degree of visual experience is not special to blindsight type 2 but can occur in healthy observers as well.

Stimulus versus decision ratings

The traditional view of subjective measures of consciousness assumes that all subjective measures of consciousness form one coherent category (Seth et al., 2008). In the present study, we observed a series of systematic differences between ratings of the stimulus and ratings of the decision: The psychometric threshold for decision ratings was lower than for stimulus ratings in all three experiments. With regard to SDT type 2 characteristics, decision ratings always imposed a more liberal response criterion and were associated with a higher sensitivity in two out of three experiments. We expected an advantage of decision ratings in type 2 sensitivity over stimulus ratings because decision ratings refer logically to the accuracy of the trial. Moreover, wagering, confidence, and attribution-of-choice ratings were more strongly associated with other decision-related scales within single trials than with stimulus ratings for both orientation discrimination in Experiments 1 and 2 and for shape discrimination in Experiments 3 and 4. Thus, consistent with our classification of subjective measures as stimulus ratings or decision ratings, both kinds of measures differed according to a variety of characteristics; these differences were replicable and generalized across several tasks. It is tempting to interpret stimulus ratings and decision ratings as measurements of the strength of overlapping but not identical neural signals, although our data support a distinction only at the level of measurements, but not at the level of mechanisms. We have speculated that stimulus ratings might constitute a measurement of neural signals during sensory processing; while decision ratings might be a measurement of neural signals during decision making. An alternative interpretation might explain the present findings by referring to only one kind of neural signal. According to this view, when participants rate the stimulus or decision, they are in fact rating the strength of the same underlying signals in both cases. Subjective measures are different in how accurately participants are able to translate these neural signals into a point on the scale. If the translation of neural signals into stimulus ratings was more prone to noise than was the translation into decision ratings, it could be explained why decision ratings are associated with a higher SDT type 2 sensitivity and why trial accuracy could be predicted at lower levels of stimulus quality than stimulus ratings. However, since noise is unsystematic, this account would predict that the correlation of stimulus ratings with all events in the world would be corrupted by noise, not only the correlation with trial accuracy. Contrary to this prediction, we observed no substantial differences between stimulus and decision ratings with respect to the steepness of psychometric functions, which indexes the relative sensitivity of the subjective measures to changes of stimulus quality. This means that decision ratings are more closely related only to correct and incorrect discrimination decisions than are stimulus ratings but there is no difference between stimulus and decision ratings in their relation to stimulus quality. Overall, this pattern of results is not consistent with the view that subjective measures are different only in their susceptibility to noise, but it supports the view that the characteristics of subjective measures influence the events subjective measures refer to.

A continuum of multiple thresholds?

The discrepancy between stimulus and decision ratings reported in the present study implies that the ascription of how

conscious a stimulus is depends on the type of subjective measure researchers adopt. In this respect, the present study relates to the classical distinction between subjective and objective thresholds of awareness (Cheesman & Merikle, 1984; Merikle, Smilek, & Eastwood, 2001). They assumed that while a stimulus of a certain strength is sufficient to reach the objective threshold and elicit a correct response, the strength of stimulation needs to be even stronger to reach the subjective threshold and elicit a verbal report; that is, the objective threshold is lower than the subjective one. Our study suggests that there might be more than one subjective threshold; specifically, the threshold for confidence and attribution of choice ratings is below the threshold for reports of visual experience. Weak stimuli might result in a weak form of representation enabling participants to perform above chance, although at the same time they deny any experience of the stimulus and claim that their performance was due to guessing (low decision and low stimulus ratings). If the stimulation is stronger, a more stable or a different kind of representation emerges, and participants report some confidence in being correct (decision ratings increase), but they still claim to have little experience of the stimulus (stimulus ratings lower than decision ratings). Only with even greater stimulation performance, decision ratings and stimulus ratings indicate concurrently that the participant is conscious of the stimulus. In other words, our data suggest that the set of events when observers perform above chance is larger than the set of events when they report to be confident, which, in turn, is larger than the set of events when observers report having visual experiences. Consequently, if a participant reports a visual experience, it is very likely that he or she will also be able to discriminate the stimulus and report confidence in the discrimination decision. The reverse is not the case: If a participant reports confidence in the discrimination decision, there is still uncertainty whether he or she reports a clear visual experience as well. However, this hierarchical relationship between experience and confidence does not necessarily hold for other paradigms. For example, in iconic memory tasks, participants typically report having seen all the items on display, although memory performance is restricted to three to five items (Sperling, 1960). To investigate the relationship between thresholds derived from stimulus ratings and decision ratings, more studies employing different paradigms and different stimulus modalities are required. Therefore, we recommend always considering stimulus and decision ratings in consciousness research.

Relation to previous studies

The results reported here are in line with a previous artificial grammar study that reported SDT type 2 sensitivity of confidence ratings to be greater than the sensitivity of awareness ratings (Wierzchoń et al., 2012). However, our results only partially replicate the results of prior visual studies (Sandberg et al., 2011; Sandberg et al., 2010). In a masked object discrimination task, Sandberg and colleagues reported, in line with our results, that the psychometric threshold for a stimulus-based rating scale, the PAS, was more conservative than for confidence. However, unlike in our results, PAS outperformed both confidence ratings and wagering in predicting discrimination performance. One methodological difference between their study and our studies is the employed stimulus rating. In the study by Sandberg and colleagues, participants rated their experience on the PAS, a 4-point scale that distinguished between *no experience*, *brief glimpse*, *almost clear experiences*, and *clear experiences*. Critically, the choice *brief glimpses* is defined as "a feeling that something has been shown, but is not characterised by any content, and cannot be specified any further" (Ramsøy & Overgaard, 2004). In the present study, participants rated their clarity of visual experience of the task-relevant stimulus feature—for example, coherent motion. Supposing that an observer had an experience that matches the definition of a brief glimpse in the PAS—an experience without any content—in the present study, the observer would nevertheless veridically indicate a maximally unclear experience, because he or she would not have any experience of the task-relevant stimulus feature. However, using the PAS, the participant would veridically report a brief glimpse. In other words, the PAS might measure a larger set of experiences than our stimulus ratings because it requires participants to report experiences without content as well, which could also be nonvisual intuitions. However, this is entirely post hoc reasoning; a valid comparison between the PAS and our scales would require a comparison of all scales based on the same paradigm and balanced briefing of participants.

Conclusion

In summary, the present experiments indicate that participants' verbal reports when being asked to rate their perception of the stimulus versus their discrimination response—although being similar in many ways—show reliable and important differences. Similar to type II blindsight patients, subjective ratings that referred to a discrimination decision had lower thresholds than did subjective measures that referred to the percept of the stimulus; that is, observers reported confidence or knowledge about the correctness of their responses at a greater level of stimulus ambiguity than when they reported experience of the stimulus. Moreover, decision ratings exhibited different SDT type 2 characteristics, and different decision-related scales were more strongly correlated with other decision-related scales than with reports of experience. We suggest that consciousness research has to consider the use of a subjective measure that refers to the experience of

the stimulus in addition to a measurement that assesses confidence in the discrimination decision.

Author Note This research was supported by the following grants: DFG (Deutsche Forschungsgesellschaft; i.e., German Research Council) grant ZE 887/3-1 and German-Israeli Foundation for Scientific Research and Development (GIF) grant 1130–158 (both to M.Z.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Correspondence concerning this article should be addressed to Michael Zehetleitner.

References

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences, 9,* 46–52. doi:10.1016/j.tics.2004.12.006

Boyer, J. L., Harrison, S., & Ro, T. (2005). Unconscious processing of orientation and color without primary visual cortex. *Proceedings of the National Academy of Sciences, 102,* 16875–16879.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Chalmers, D. (1998). On the search of neural correlates of consciousness. In S. Hameroff, A. Kaszniak, & A. Scott (Eds.), *Toward a science of consciousness II: The second Tucson discussions and debates*. Cambridge, MA: MIT Press.

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics, 36,* 387–395.

Crick, F., & Koch, C. (1990). Towards a neurobiologcial theory of consciousness. *Seminars in the Neurosciences, 2,* 263–275.

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology, 5,* e260. doi:10.1371/journal.pbio.0050260

Dennett, D. C. (2003). Who's on first - heterophenomenology explained. *Journal of Consciousness Studies, 10,* 19–30.

Dennett, D. C. (2007). Heterophenomenology reconsidered. *Phenomenology and Cognitive Science, 6,* 247–270. doi:10.1007/s11097-006-9044-9

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1322–1338.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological Research, 69,* 338–351. doi:10.1007/s00426-004-0208-3

Dienes, Z., & Seth, A. K. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition, 19,* 674–681. doi:10.1016/j.concog.2009.09.009

Erikson, C. W. (1960). Discrimination and learning without awareness: A methodological survey and evaluation. *Psychological Review, 67,* 279–300.

Fleming, S. M., & Dolan, R. J. (2010). Effects of loss-aversion on post-decision wagering: Implications for measures of awareness. *Consciousness and Cognition, 19,* 352–363. doi:10.1016/j.concog.2009.11.002

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329,* 1541–1543. doi:10.1126/science.1191883

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10,* 843–876.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Mahwah, New Jersey: Lawrence Erlenbaum Publishers.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30,* 535–574. doi:10.1146/annurev.neuro.29.051605.113038

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

Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: Promise and pitfalls. *Nature Reviews Neuroscience, 6,* 247–255.

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: Model-based and distribution-free measures and evaluation. *Perception & Psychophysics, 68,* 393–414.

Kornell, N., Son, L. K., & Terrace, H. S. (2007). Tranfer of metacognitive skills and hint seeking in monkeys. *Psychological Science, 18,* 64–71. doi:10.1111/j.1467-9280.2007.01850.x

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory. A user's guide*. Mahwah, NY: Lawrence Erlbaum Associates.

Merikle, P. M., Smilek, D., & Eastwood, J. D. (2001). Perception without awareness: Perspectives from cognitive psychology. *Cognition, 79,* 115–134.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83,* 435–450.

Nakamura, N., Watanabe, S., Betsuyaku, T., & Fujita, K. (2011). Do birds (pigeons and bantams) know how confident they are of their perceptual decisions? *Animal Cognition, 14,* 83–93. doi:10.1007/s10071-010-0345-6

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26,* 125–174.

Overgaard, M., & Sandberg, K. (2012). Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B, 367,* 1287–1296. doi:10.1098/rstb.2011.0425

Peirce, C. S., & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences, 3,* 73–83.

Pelli, D. G. (1997). The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Persaud, N., Davidson, M., Mansicalco, B., Mobbs, D., Passingham, R. E., Cowey, A., & Lau, H. (2011). Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *NeuroReport, 58,* 605–611. doi:10.1016/j.neuroimage.2011.06.081

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decisional wagering objectively measures awareness. *Nature Neuroscience, 10,* 257–261. doi:10.1038/nn1840

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & the R Development Core Team. (2012). nlme: Linear and nonlinear mixed effects models. R package version 3.1-105.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117,* 864–901. doi:10.1037/a0019737

R Core Team. (2012). R: A language and environment for statistical computing. (Version 2.14.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Rees, G., Kreiman, G., & Koch, C. (2002). Neural correlates of consciousness in humans. *Nature Reviews Neuroscience, 3,* 261–270.

Riddoch, G. (1917). Dissociation of visual perceptions due to occipital injuries, with especial reference to appreciation of movement. *Brain, 40,* 15–57.

Sahraie, A., Weiskrantz, L., & Barbur, J. L. (1998). Awareness and confidence ratings in motion perception without geniculo-striate projection. *Behavioural Brain Research, 96,* 71–77.

Sahraie, A., Weiskrantz, L., Trevethan, C. T., Cruce, R., & Murray, R. D. (2002). Psychophysical and pupillometric study of spatial channels of visual processing in blindsight. *Experimental Brain Research, 143,* 249–256. doi:10.1007/s00221-001-0989-1

Sandberg, K., Bibby, M. B., Timmermans, B., Cleeremans, A., & Overgaard, M. (2011). Measuring consciousness: Task accuracy and awareness as sigmoid functions of stimulus duration. *Consciousness and Cognition, 20,* 1659–1675. doi:10.1016/j.concog.2011.09.002

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness- Is one measure better than the other? *Consciousness and Cognition, 19,* 1069–1078. doi:10.1016/j.concog.2009.12.013

Schmidt, T., & Vorberg, D. (2006). Criteria for unconscious cognition: Three types of dissociation. *Perception & Psychophysics, 68,* 489–504.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1264–1288. doi:10.1037/a0012943

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science, 15,* 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,* 314–321. doi:10.1016/j.tics.2008.04.008

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74,* 1–29.

Weiskrantz, L. (1986). *Blindsight: A case study and implications*. New York: Oxford University Press.

Weiskrantz, L., Barbur, J. L., & Sahraie, A. (1995). Conscious versus unconscious visual discrimination with damage to the visual cortex (VI). *Proceedings of the National Academy of Sciences, 92,* 6122–6126.

Wickens, T. D. (2002). *Elementary signal detection theory.* New York: Oxford University Press.

Wierzchoń, M., Asanowicz, D., Paulewicz, B., & Cleeremans, A. (2012). Subjective measures of consciousness in artificial grammar learning task. *Consciousness and Cognition, 21,* 1141–1153. doi:10.1016/j.concog.2012.05.012

World Medical Association. (2008). WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects. from http://www.wma.net/en/30publications/10policies/b3/index.html

Zeki, S., & ffytche, D. H. (1998). The Riddoch syndrome: Insights into the neurobiology of conscious vision. *Brain, 121,* 25–45.