Metacognitive sensitivity of subjective reports of decisional confidence and visual experience

Manuel Rausch, Hermann J. Müller, and Michael Zehetleitner
Ludwig-Maximilians-Universität München

This is the final draft of the manuscript published at Consciousness and Cognition. Please cite this work as follows:
Rausch, M., Müller, H. J., & Zehetleitner, M. (2015). Metacognitive sensitivity of subjective reports of decisional confidence and visual experience. *Consciousness and Cognition, 35*, 192-205. doi:10.1016/j.concog.2015.02.011

Author Note

Manuel Rausch, Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany, Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Munich, Germany; Hermann J. Müller, Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany, School of Psychological Sciences, Birkbeck College, University of London, London, UK; Michael Zehetleitner, Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

This research is supported by grant 1130-158 of the German-Israeli Foundation for Scientific Research and Development (GIF) and grant ZE 887/3-1 of the Deutsche Forschungsgesellschaft (DFG) (both to MZ). The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Correspondence should be addressed at Manuel Rausch, Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstraße 13, 80802 Munich, Germany. Phone: +4989 2180 5152; Email: manuel.rausch@psy.lmu.de

Abstract

Previous studies provided contradicting results regarding metacognitive sensitivity estimated from subjective reports of confidence in comparison to subjective reports of visual experience. We investigated whether this effect of content of subjective reports is influenced by the statistical method to quantify metacognitive sensitivity. Comparing logistic regression and meta-d in a masked orientation task, a masked shape task, and a random-dot motion task, we observed metacognitive sensitivity of reports regarding decisional confidence was greater than of reports about visual experience irrespective of mathematical procedures. However, the relationship between subjective reports and the logistic transform of accuracy was often not linear, implying that logistic regression is not a consistent measure of metacognitive sensitivity. We argue that a science of consciousness would benefit from the assessment of both visual experience and decisional confidence, and recommend meta-d as measure of metacognitive sensitivity for future studies.

Keywords: Consciousness, visual awareness, subjective report, confidence, meta-d, logistic regression, signal detection theory, type 2 signal detection theory

1. Introduction

Empirical approaches to human consciousness crucially rely on measures to determine whether or not an observer is conscious of a stimulus (Chalmers, 1998). Many researchers prefer objective measures, where conscious awareness is ascribed based on performance in a discrimination task (e.g. Erikson, 1960; Hannula, Simons, & Cohen, 2005; Schmidt & Vorberg, 2006). However, at least two popular theoretical perspectives imply that conscious awareness ought to be measured by subjective reports: First, according to higher-order theories, perception of a stimulus is conscious only if it is associated with a higher-order representation, i.e. a representation of oneself as perceiving the stimulus (Carruthers, 2011; Lau & Rosenthal, 2011). While discrimination performance is not necessarily accompanied by a corresponding higher-order representation, a subjective report does require some higher-order knowledge (participants need to know that they are aware of the stimulus in order to report that they are aware) and are thus considered more valid measures of conscious awareness than discrimination performance (Dienes, 2004, 2008; Lau, 2008). Second, according to the perspective of heterophenomenology, participants' verbal reports about their subjective experience are themselves objects of study in consciousness research (Dennett, 2003, 2007) and are thus the appropriate raw data that needs to be recorded and explained (Dehaene, 2010; Dehaene & Naccache, 2001).

1.1 Visual experience and confidence as content of subjective reports

A consequence of these theoretical reasons for using subjective measures of conscious awareness is the need of appropriate scales to record subjective reports. One characteristic of subjective reports that requires special consideration is *the content of subjective report*, i.e. what the subjective report is about. The contents queried in visual awareness experiments fall into two categories depending on whether participants are asked to make a report about their experience of the stimulus, or about the accuracy of a discrimination task response (Zehetleitner & Rausch, 2013). We will refer to the first kind of content as "visual experience", and the second kind as "confidence". Examples for scales with visual experience as content of subjective reports are ratings how visible the stimulus was (Sergent & Dehaene, 2004) or how clear a specific stimulus feature was experienced (Rausch & Zehetleitner, 2014). Examples for the discrimination response as content are reports of how confident participants were about the preceding task response (Peirce & Jastrow, 1884), or whether the last task response was made by guessing or based on knowledge (Zehetleitner & Rausch, 2013).

Aiming to identify the best scale to measure conscious awareness empirically, a series of previous studies has compared subjective reports collected with different scales (Dienes & Seth, 2010; Rausch & Zehetleitner, 2014; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010; Szczepanowski, Traczyk, Wierzchoń, & Cleeremans, 2013; Wierzchoń, Asanowicz, Paulewicz, & Cleeremans, 2012; Wierzchoń, Paulewicz, Asanowicz, Timmermans, & Cleeremans, 2014). As subjective scales are often used to determine whether

performance in a specific task is conscious or unconscious, the scales were compared by examining the correlation between subjective reports and task accuracy: On the assumption that the correlation between reports and accuracy is mediated by conscious processes, if one scale was found to predict accuracy better than the other scales, it was concluded that this scale is more sensitive in detecting conscious processes (that the other scales miss) and is thus closer to being an exhaustive measure of conscious awareness (Overgaard & Sandberg, 2012). This reasoning rests on the assumption that the scales under comparison are equally valid from a conceptual point of view, but some are more suitable research instruments than others.

In contrast to the assumption that all scales are a priori valid measurements of conscious experience, we have proposed that which content of subjective reports is appropriate depends on the set of conscious experiences relevant to a specific research question (Rausch & Zehetleitner, 2014). The reason is that participants might already experience some conscious intuition about being correct in a discrimination task while not yet consciously seeing the stimulus feature relevant for the task judgment (Zehetleitner & Rausch, 2013). A similar dissociation between knowledge about the accuracy of task decisions and the knowledge underlying those task decisions was shown for artificial grammar tasks (Dienes & Scott, 2005). These observations suggest that studies investigating the neural correlates of a specific visual content (such as the redness of an apple) may encounter false positives if they rely on confidence judgments because confidence may not necessarily require a conscious visual experience of the relevant stimulus feature. On the other hand, if the full set of experiences during visual perception is of theoretical interest to a specific study, the use of a scale that measures only visual experience of one specific feature leaves out subjective feelings of confidence (Zehetleitner & Rausch, 2013), and possibly other qualitatively different experiences along the unawareness/awareness continuum, such as awareness of an event without a phenomenology of seeing, as reported by some blindsight patients (Sahraie, Weiskrantz, Trevethan, Cruce, & Murray, 2002), or experiences without any content (Ramsøy & Overgaard, 2004). Finally, if a study investigates whether performance in a specific task is conscious, confidence ratings are a convenient choice since participants should consider all their conscious experiences relevant for their performance in this case (Dienes, 2008). Overall, should reliable differences between subjective scales with different contents exist, then researchers would have to decide which set of conscious experiences is relevant to their particular research questions, and choose a measure accordingly.

1.2 Type 2 signal detection theory

As subjective reports entail making a decision for one out of the several response alternatives offered by the scale, it is legitimate to apply theories of decision making to subjective reports. One of the most prominent theories of decision making under uncertainty is signal detection theory (SDT, Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). According to SDT, when observers decide which out of two possible event

types occurred, their perceptual systems create sensory evidence delineating the two response options. As there is noise in the system, the sensory evidence is not constant, but a random sample out of a distribution for each of the two event types. Participants select a response by comparing the sensory evidence with a response criterion, choosing one option if the sensory evidence is greater than the criterion and the other option otherwise. SDT allows distinguishing between two aspects of decision making: sensitivity and bias. The more sensitive an observers is, the smaller is the overlap between the two distributions of evidence created by the two events. Bias towards one response options however depends on the position of the response criterion (see Fig. 1a).

SDT tasks can be classified based on the events participants have to discriminate: In type 1 tasks, the standard application of SDT, participants differentiate between two different kinds of stimulation (e.g. two distinct stimuli, or the presence or absence of the stimulus). However, SDT can also be applied to type 2 tasks, where the task is to differentiate correct and incorrect responses to a type 1 task (Galvin, Podd, Drga, & Whitmore, 2003). Type 2 tasks allow the assessment of sensitivity and bias just as in type 1 tasks (see Fig. 1b): *Metacognitive sensitivity*, the sensitivity in type 2 tasks, is defined as the extent to which the observers' type 2 responses differentiate between correct and incorrect type 1 responses. *Metacognitive bias* indicates how liberal or conservative participants' type 2 responses are with respect to their task performance (Fleming & Lau, 2014; Galvin et al., 2003). Quantifying metacognitive sensitivity is challenging because metacognitive sensitivity depends on type 1 sensitivity and bias and standard models predict heavily skewed distributions of evidence for type 2 decisions (Barrett, Dienes, & Seth, 2013; Galvin et al., 2003). Nevertheless, type 2 SDT analysis is both conceptually and practically useful for the study of subjective reports because it allows a separation of observers' degree of insight into their own performance in the task from observers' response strategies.

INSERT FIGURE 1 ABOUT HERE

1.3 Empirical studies on confidence and visual experience

Is there an effect of experience and confidence as content of subjective reports on metacognitive sensitivity and bias? Concerning metacognitive bias, there is a considerable amount of evidence that participants apply different criteria when they make a report concerning their subjective confidence in being correct in a discrimination judgment, compared to when they report their visual experience of the task-relevant stimulus feature. Extreme examples for dissociations between visual experience and confidence stem from neuropsychological patients. For instance, Carota and Calabrese (2013) described a patient with achromatopsia after bilateral occipital damage, who claims to be entirely color-blind, but is still able to make accurate color discriminations and reports being confident about these color judgments. A similar pattern has been documented in blindsight type 2, which, unlike classical blindsight, is characterized by awareness of some event, but without the phenomenology of normal seeing (Sahraie et al., 2002). Patient G.Y. reported being confident

in discrimination judgments without experiencing the stimuli visually (Sahraie, Weiskrantz, & Barbur, 1998) and even wagered the same amount of money for the blind as for the intact hemifield when discrimination difficulty was matched (Persaud et al., 2011). In normal observers, decisional confidence is associated with more liberal criteria across a wide range of visual tasks, such as a stimulus localization task (Schlagbauer, Müller, Zehetleitner, & Geyer, 2012), a masked orientation discrimination task, a masked shape discrimination task, and a random-dot motion discrimination task (Zehetleitner & Rausch, 2013).

For metacognitive sensitivity, the evidence for a distinction between experience and confidence is less consistent. The only neuropsychological study informative of metacognitive sensitivity reported that blindsight patient G.Y.'s area under the receiver operating characteristic (ROC) is larger when it is estimated from confidence judgments as compared to visual awareness at low stimulus intensities (Sahraie et al., 1998). In normal observers, subjective reports of perceptual experience outperformed confidence ratings in predicting trial accuracy in a masked object discrimination task (Sandberg et al., 2010) as well as a masked face discrimination task (Wierzchoń et al., 2014); however, subjective reports of decisional confidence were more efficient in predicting trial accuracy in a masked orientation discrimination task and a random-dot motion discrimination task (Zehetleitner & Rausch, 2013); and no substantial differences were found in a masked discrimination task of affective face expressions (Sandberg, Bibby, & Overgaard, 2013; Szczepanowski et al., 2013) and a masked shape discrimination task (Zehetleitner & Rausch, 2013).

These discrepant results of previous studies raise the question what are the factors that determine when visual experience and when confidence is associated with greater metacognitive sensitivity. One candidate factor may be the *method used to quantify metacognitive sensitivity*: Those two studies that found metacognitive sensitivity of visual experience to be higher than that of decisional confidence were both based on logistic regression analysis (Sandberg et al., 2010; Wierzchoń et al., 2014). By contrast, Szczepanowski et al. (2013) and Zehetleitner and Rausch (2013), who used type 2 ROC analysis to quantify metacognitive sensitivity (Fleming, Weil, Nagy, Dolan, & Rees, 2010), observed that metacognitive sensitivity of confidence was substantially greater than metacognitive sensitivity of experience or at least confidence tended to be associated with a greater metacognitive sensitivity. Since the measure of metacognitive sensitivity is closely associated with the effects of the scale across previous studies, the question arises if the effect of confidence versus experience is entirely dependent on which measure is applied.

1.4. Meta-d as measure of metacognitive sensitivity

The development of meta-d, a relatively new approach to quantifying metacognitive sensitivity (Maniscalco & Lau, 2012), offers the possibility assess metacognitive sensitivity with improved control (Fleming & Lau, 2014). The conceptual idea of meta-d is to express metacognitive sensitivity in terms of sensitivity of a type 1 SDT model (see Fig. 2). In such a model, participants are assumed to make objective discrimination responses and subjective reports based on identical sensory evidence. Subjective reports and task decisions are

considered to form one continuum of responses such as “I’m sure it’s A”, “I guess A”, “I guess B”, “I’m sure it’s B”. Participants select one response out of the continuum based on comparisons of one value of sensory evidence, which is a random sample out of different distributions generated by A and B, with criteria that delineate the different response options. If participants had the same amount of evidence for subjective reports as they have for the task response, the distance between the two distributions should be same no matter whether it is estimated from A versus B decisions alone, or from A versus B decisions plus subjective reports. Thus, meta-d indicates the distance between the two distributions of evidence available for subjective responses. If meta-d is smaller than d’, the distance between distributions of evidence estimated from “objective“ decisions alone, this would means that there is less sensory evidence for subjective reports than for task responses and that, accordingly, metacognitive sensitivity is suboptimal. An introduction into the mathematics of meta-d is provided by Barrett et al. (2013).

INSERT FIGURE 2 ABOUT HERE

Meta-d and type 2 ROC analysis have both advantages and disadvantages: On the one hand, type 2 ROC analysis has the advantage of being free of assumptions about the underlying distributions of evidence, while meta-d requires making assumptions about the shape of these distributions, which may be incorrect. On the other hand, meta- d provides two advantages over type 2 ROC analysis: First, meta-d accounts for bias regarding the two task alternatives. Second, meta-d can be used to easily compare metacognitive sensitivity to objective task performance because meta-d is expressed in the same signal-to-noise units as the standard d’ from signal detection theory (Fleming & Lau, 2014; Maniscalco & Lau, 2012). However, no study to date has compared different subjective reports of visual experience and decisional confidence in terms of meta-d.

1.5. Logistic regression as measure of metacognitive sensitivity

Despite the merits of type 2 SDT analysis, the majority of previous studies comparing subjective reports have quantified the relation between trial accuracy and subjective reports by logistic regression (Rausch & Zehetleitner, 2014; Sandberg et al., 2013; Sandberg et al., 2010; Wierzchoń et al., 2012; Wierzchoń et al., 2014). Logistic regression, a special case of generalized linear regression models, is a method to quantify the relationship between a binary outcome variable and one or several predictors. Linear regression methods assume a linear relationship between outcome and predictor: To obtain such a linear relationship, the outcome variable is transformed into the logarithm of the odds of the two possible outcome events. In case of metacognitive sensitivity, the correctness of the trial serves as binary outcome variable, and subjective report as linear predictor. Thus, the subjective report is used to predict the logarithm of the odds of the trial being correct to being incorrect (see Fig. 3). The more efficient subjective reports differentiate between different levels of accuracy, the

steeper the slope of the resulting regression line is. Thus, the slopes of logistic regression are interpreted as measure of metacognitive sensitivity.

INSERT FIGURE 3 ABOUT HERE

On the one hand, logistic regression provides several advantages over other methods to analyze non-linear data: First, it is possible to include random effects to account for hierarchical clusters in the data, such as blocks nested within participants nested within experiments (Bolker et al., 2008; Pinheiro & Bates, 2000). Second, logistic mixed-model regression can be applied when the data is unbalanced (Bolker et al., 2008), that is, when the number of observations varies between conditions or even if there are empty cells in the design matrix. This is particularly useful for studies of metacognition because the number of errors may vary greatly among participants and conditions in the same experiment.

On the other hand, the assumption of a linear relationship between subjective reports and transformed accuracy logistic regression relies upon is unlikely to hold. First, the data provided by rating scales is inherently categorical, not continuous, and linear models are inappropriate in particular for rating scales with small numbers of categories (Christensen & Brockhof, 2013). In contrast, ratings on a visual analog scale (VAS) may be at least approximately equidistant (Reips & Funke, 2008). Second, even if scale steps were equidistant, a non-linear relationship between the transformed accuracy and subjective reports might be expected in all tasks where participants have to select one out of a finite number of options: If there is a chance p of guessing correctly, the transformed odds of being correct cannot vary between $-\infty$ and ∞; instead, it will asymptotically approach a lower bound at the logarithm of $p/(1-p)$. A non-linear relationship between subjective reports and transformed accuracy would have two implications: (i) the interpretation of logistic regression slopes as indices of metacognitive sensitivity would be ambiguous because the slope of the regression would vary across different parts of the scale, being close to zero for the lower part of the scale, and increasing only at the upper part; (ii) logistic regression might underestimate the metacognitive sensitivity of scales imposing liberal criteria for lower scale steps, because the more liberal criteria are, the larger will be the part of the scale where the transformed accuracy cannot decrease any further due to the lower bound imposed by the guessing probability.

1.6. Rationale of the present study

In present paper, we investigate two issues: First, we examined whether an analysis of meta-d and logistic regression would reveal the same effect of visual experience versus decisional confidence (as contents of subjective reports) on metacognitive sensitivity as suggested by previous type 2 ROC analyses. Second, we investigated whether the assumption of a linear relationship between subjective reports and transformed accuracy, which is required if logistic regression is used as an index of metacognitive sensitivity, is justified.

Specifically, we predicted that if the method of assessing metacognitive sensitivity is indeed the reason for the discrepancy of results observed in previous studies, logistic regression coefficients of reports of visual experience should be greater than those of decisional confidence. If the effect of confidence associated with a larger area under the type 2 ROC curve than visual experience as observed previously reflected a stable pattern of the data, then metacognitive sensitivity of confidence should be greater no matter if quantified by meta-d or logistic regression. In addition, if the assumption of a linear relationship between subjective reports and transformed accuracy is well-founded, then no non-linear trends should be observed. In contrast, if there was a bias to logistic regression due to a lower bound to the transformed accuracy, we would expect positive quadratic trends between subjective reports and transformed accuracy, and the quadratic trends should be more pronounced for decisional confidence as confidence is associated with more liberal criteria.

To address these issues, we performed a reanalysis of three previously published experiments, a masked orientation discrimination task, a masked shape discrimination task, and a random-dot motion discrimination task (Zehetleitner & Rausch, 2013). In each of these experiments, participants submitted three responses on each trial: A 2-AFC discrimination judgment was followed by a report of the visual experience of the task-relevant stimulus feature along with a report of subjective confidence in being correct on the just performed discrimination judgment. For each of experiment, we analyzed metacognitive sensitivity based on logistic regression analysis as well as meta-d.

2. Material and Methods

In the present paper, we reanalyzed Experiment 1, Experiment 3, and Experiment 5 conducted by Zehetleitner and Rausch (2013). A detailed description of the methodology can be found there. Experiments 2 and 4 were not considered for reanalysis because these experiments did not require participants to report their visual experience.

2.1. Experimental tasks

The experiments involved a masked orientation discrimination task (N = 20), a masked shape discrimination task (N = 16), and a motion discrimination task (N = 21). All three experiments had an identical trial structure (see Fig. 4). First, participants were presented with a stimulus always at fixation. For the masked orientation task, the stimulus was a sinusoidal grating oriented either horizontally or vertically, followed by a checkerboard mask after a stimulus onset asynchrony (SOA) of 10, 20, 30, 40, 50, 70, 90, or 140 ms. For the masked shape task, the stimulus was either a circle or a square filled with the same sinusoidal grating as in the orientation task, succeeded by the checkerboard mask after SOAs of 8.3, 16.7, 25.0, 33.3, 50.0, 66.7, 83.3, or 116.7 ms. For the motion discrimination task, the stimulus was a random dot kinematogram, with 0.7, 1.3, 2.7, 5.3, 10.7, 21.3, or 42.7 % of the dots coherently moving to either the left or the right, and the remaining dots relocated randomly. Participants had to make a non-speeded two-alternative forced-choice by key press about the stimulus they just had been presented with: For the masked orientation task, they

indicated whether the sinusoidal grating had been horizontal or vertical; for the masked shape task, they reported whether the stimulus had been a square or a circle; and for the motion discrimination task, they indicated whether the dots had moved towards the left or the right. After each discrimination response, participants made two subjective reports, one regarding their visual experience of the stimulus, and one regarding their confidence in being correct in the discrimination task. For that, each question was displayed on the screen, which was: "How clearly did you see the grating/shape/coherent motion?" or "How confident are you that your response was correct?" In the orientation task, participants were asked not only to report their confidence, but additionally, in one third of the blocks , to wager money on the outcome of the judgment, and, in another third, to indicate whether their response was more due to guessing or to knowledge. The sequence of questions was balanced within participants in the orientation task, and across participants in the other two tasks. Participants delivered subjective reports using a joystick and a VAS, which means that participants selected a position along a continuous line between two end points by moving a cursor. The end points were labeled as "unclear" and "clear" for the experience scale and "unconfident" and "confident" for the confidence scale, i.e. observers indicated their experience or confidence by the selected cursor position on the continuous scale (see Fig. 4). If the discrimination judgment was erroneous, the trial ended by displaying the word "error" for 1,000 ms on the monitor. There was no feedback with respect to the subjective report.

INSERT FIGURE 4 ABOUT HERE

2.2. Analysis

All analysis were conducted in the free software R 3.0.2 (R Core Team, 2013). Trials of the masked orientation task on which participants did not report their subjective confidence in being correct were excluded from the analysis.

2.2.1. Logistic regression

Logistic mixed regression analysis was performed using the R library lme4 (Bates, 2005; Bates, Maechler, Bolker, & Walker, 2013), with error as dependent variable and stimulus quality (logarithm of SOA for the orientation and the shape discrimination task, logarithm of coherence for the motion task), first report, second report, scale (confidence first vs. experience first), as well as all possible two-way and three-way interactions as fixed effects, and a random effect on the intercept. All numerical predictors were centered and scaled. Statistical significance was assessed via likelihood ratio tests conducted by dropping the effect to be tested out of a model containing all effects of the same order. Contrasts were coded in a way that the regression coefficients of scale can be directly interpreted as difference between experience and confidence. Confidence intervals were estimated around fixed effects from the local curvature of the likelihood surface. To resolve the interaction between scale, stimulus quality, and subjective reports, we performed likelihood ratio tests

comparing models that only included main effects of report and scale against models with an interaction between report and scale, separately for each level of stimulus quality, with p-values adjusted according to the Bonferroni method to account for multiple comparisons.

2.2.2. Meta-d

Meta-d was computed using an implementation of the maximum likelihood procedure described by Maniscalco and Lau (2012) in the free software R (code is found in the Supplementary Material), assuming normal distributions of evidence with non-equal variances. First, the continuous VAS rating data was divided into 13 equal bins. Then, meta-d was computed separately for each participant and each condition and then subjected to a mixed linear regression model with the fixed factors scale (experience vs. confidence), time (first vs. second report), and stimulus quality and a random effect on the intercept (again based on the R library lme4). We used mixed linear regression models instead of ANOVAs because the factors time and scale varied within participants, but were not crossed in the shape discrimination and the motion discrimination experiments. Contrasts were coded in a way that the regression coefficients of scale and time can directly be interpreted as difference in meta-d between conditions. Confidence intervals around fixed effects were estimated from 10,000 parametric bootstrap samples. Significance was assessed by Wald t-tests using degrees of freedom estimated by Satterthwaite's approximation implemented in the R library lmerTest (Kuznetsova, Brockhof, & Christensen, 2014). To resolve interactions between stimulus quality and scale, separate t-tests were computed for each level of stimulus quality, with p-values corrected using the Bonferroni method. We repeated this analysis assuming two other distributions of evidence, the logistic distribution and the distribution of the smallest extremes, which gave essentially the same pattern of results as we obtained with the normal distribution.

2.2.3. Association between reports and stimulus quality

To assess the relationship between reports and stimulus quality, we computed non-parametric Goodman and Kruskal's gamma correlation coefficients separately for each participant and for visual experience and confidence. Paired t-tests were conducted to test for a difference between scales.

3. Results

3.2. Logistic regression

The complete results of the mixed logistic regression models can be seen in Table 1. We found significant interactions between the first report and scale in the masked shape task and the motion task, as well as between the second report and scale in all three experiments. Only for the first report in the masked orientation task, no significant interaction was detected. The sign of the coefficients of each interaction term between scale and report indicated concurrently that subjective reports of decisional confidence were more efficient in

predicting trial accuracy than the reports of visual experience. While there were no three-way interactions of ratings, scale, and stimulus quality in the masked orientation task and in the motion task, we observed significant interactions between rating, stimulus quality and scale in the masked shape task. To resolve these three-way interactions, we tested the interaction between scale and rating with separate logistic regression models for each level of stimulus quality of the masked shape task, observing significant interactions at the SOAs of 50, 66, and 116.7 ms, $\chi^2(2) = 22.7$, $p_{cor} < .001$, $\chi^2(2) = 13.1$, $p_{cor} < .05$, and $\chi^2(2) = 12.1$, $p_{cor} < .05$, respectively.

INSERT TABLE 1 ABOUT HERE

The relationships between subjective report and transformed accuracy are depicted separately for scale, experiment, and time of the report in Fig. 5. For the masked orientation task, we detected no substantial quadratic trend at the first report, $\chi^2(1) = 0.6$, n.s., but we did at the second, $\chi^2(1) = 22.5$, $p < .001$. For the masked shape task, there was a significant quadratic trend at the first report, $\chi^2(1) = 18.0$, $p < .001$, but not at the second, $\chi^2(1) = 1.1$, n.s. For the motion discrimination task, we again detected no significant quadratic trend at the first report, $\chi^2(1) = 0.1$, n.s., while there was one at the second report, $\chi^2(1) = 18.2$, $p < .001$.

Significant interactions between quadratic trends and scale were only detected for the masked shape task, first report: $\chi^2(1) = 11.1$, $p < .001$, second report: $\chi^2(1) = 6.2$, $p < .05$. Separate models for only experience and confidence revealed a significant quadratic trend for confidence only, $\chi^2(1) = 45.5$, $p < .001$, but not for experience, $\chi^2(1) = 2.0$, n.s.

INSERT FIGURE 5 ABOUT HERE

3.3. Meta-d

As can be seen from Fig. 6, meta-d scores estimated from confidence ratings were greater than meta-d scores for experience in all three experiments. In the masked orientation experiment and the motion experiment, this effect emerged already at very low stimulus quality (i.e., short SOAs), where meta-d of experience was still at chance level; in the masked shape task, by contrast, the effect became evident only at longer SOAs.

INSERT FIGURE 6 ABOUT HERE

The results of the mixed linear regression models can be seen in Table 2. We found substantial negative effects of scale in all three experiments, indicating that meta-d scores computed from visual experience were indeed always smaller than meta-d scores of decisional confidence. Substantial effects of time or an interaction between time and any of the other variables were not detected. However, we observed significant interactions between stimulus quality and scale in the masked orientation and the masked shape experiment, but

not in the motion experiment. Post-hoc tests comparing meta-d between experience and confidence separately at each SOA revealed that for the orientation task, meta-d of confidence was greater than that of experience for each SOA longer than 50 ms, all $t(19)$'s > 2.2, all p_{cor}'s < .05. For the masked shape experiments, we found meta-d of confidence to be above meta-d of experience at the SOA of 50 ms, $t(15) = 3.7$, $p_{cor} < .05$, as well as the SOA of 116.7 ms, $t(15) = 5.0$, $p_{cor} < .01$.

INSERT TABLE 2 ABOUT HERE

3.4. Correlation between reports and stimulus quality

The mean gamma correlation coefficient between reports and stimulation strength were .68 for experience and .69 for confidence in the masked orientation task, both .62 in the masked form task, and .59 and .60, respectively, in the motion task. None of these differences were significant, all t's > .7, n. s.

4. Discussion

The analysis presented here was conducted to examine two issues: (i) does the effect of visual experience versus decisional confidence (as contents of subjective reports) on metacognitive sensitivity depend on the method used to quantify metacognitive sensitivity, and (ii) is logistic regression biased owing to a non-linear relationship between transformed accuracy and subjective reports?

Concerning the effect of content, meta-d indicated that metacognitive sensitivity of decisional confidence was greater than of visual experience in all three tasks. Consistent with the hypothesis that the effect of experience versus confidence is largely independent of the method to quantify metacognitive sensitivity, we detected the same effect in five out of six tests using logistic regression analysis. The correlation between subjective reports of visual experience and quality of stimulation was the same as the correlation between confidence and the quality of stimulation, indicating that none of the two scales was compromised by a large amount of noise.

Concerning the relationship between transformed accuracy and subjective reports, logistic regression revealed at least one quadratic trend out of the two subjective reports in each experiment, indicating that the interpretation of logistic regression slopes as metacognitive sensitivity is often ambiguous and may be confounded by response criteria settings. While the quadratic trend in the masked shape task was primarily driven by decisional confidence, we observed no differences between experience and confidence in terms of non-linear trends in the other two experiments.

4.1. Why confidence outperforms visual experience in predicting trial accuracy

There are three potential explanations why subjective reports of confidence are different from subjective reports of visual experience: (i) independent conscious access of different stimulus features, (ii) distinct metacognitive mechanisms (Overgaard & Sandberg, 2012), and, respectively, (iii) placement of different criteria (Wierzchoń et al., 2012; Wierzchoń et al., 2014).

The first account is closely linked to the theoretical proposal that a stimulus is represented by a hierarchy of features, and conscious access to the different features of a stimulus can vary independently (Kouider, de Gardelle, Sackur, & Dupoux, 2010). According to this theory, partial awareness is a state where some features are consciously accessible while other features cannot be accessed. Decisional confidence may depend to a large degree on conscious access of the relevant feature to the discrimination decision (Dienes, 2008). If additional task-irrelevant features of the stimulus contribute to the quality of visual experience to a greater extent than they do to confidence judgments, this would explain why confidence judgments are more strongly associated with task accuracy. At the same time, conscious access of both task-relevant and task-irrelevant features varies as a function of physical stimulus quality; consequently, a state of partial awareness would also explain why the correlations of confidence and visual experience with task difficulty are the same. Finally, if decisional confidence requires conscious access to only that feature which is task-relevant, but visual experience requires conscious access to other features in addition to the task-relevant one, the condition for reporting confidence may be met more frequently than the condition for reporting a visual experience, thus explaining why reports of visual experience are associated with more restrictive criteria (Carota & Calabrese, 2013; Sahraie et al., 1998; Schlagbauer et al., 2012; Zehetleitner & Rausch, 2013).

The second explanation for varying metacognitive sensitivity between different scales posits different metacognitive mechanisms underlying the making of subjective reports: Overgaard and Sandberg (2012) suggested that subjective reports of experience rely on introspection, an online inspection of ongoing mental states, whereas confidence judgments are mediated by additional more complex metacognitive processes requiring insight into the decision processes during the objective task. Based on the second assumption that insight into one's decision making is more error-prone than pure introspection, Overgaard and Sandberg (2012) predicted that metacognitive sensitivity of visual experience is greater than that of decisional confidence. However, the pattern we observed was just reversed, indicating that reporting one's confidence is not more difficult than reporting one's visual experience. If reporting one's visual experience was then a more difficult task than reporting one's confidence, it would be expected that experience is compromised by a higher level of unsystematic noise in general. However, unsystematic noise would also decrease the correlation with the quality of stimulation, but we observed no indication of such an effect. Overall, we did not find any evidence that either subjective reports of experience or confidence are more difficult to make. Nevertheless, our data do not rule out the possibility that subjective reports of experience and confidence are mediated by independent but similarly effective metacognitive processes.

According to the third account for differences between scales, each scale is composed of different criteria along the awareness spectrum; thus, each step of each scale estimates a slightly different level of awareness (Wierzchoń et al., 2012; Wierzchoń et al., 2014). If the differences between scales were only due to metacognitive bias, rather than metacognitive sensitivity, there should be no effect of different scales if subjective criteria are controlled for. However, we find meta-d of confidence to be greater than meta-d of experience across all three experiments, indicating that the difference between experience and confidence is not due to metacognitive bias alone.

4.2. What factors contribute to the variability across studies?

The starting point for our reanalysis was the observation that the patterns of results in previous studies were closely associated with the method employed to quantify metacognitive sensitivity: While metacognitive sensitivities of decisional confidence were greater than those of visual experience in several studies (Sahraie et al., 1998; Szczepanowski et al., 2013; Zehetleitner & Rausch, 2013), two other studies both using logistic regression analysis found the opposite pattern (Sandberg et al., 2010; Wierzchoń et al., 2014). Our comparison between logistic regression and meta-d as measures of metacognitive sensitivity revealed that the overall pattern of metacognitive sensitivity of confidence compared to experience was largely independent of the method used to assess metacognitive sensitivity. Consequently, the question what factors determine whether subjective reports of experience or confidence are associated with greater metacognitive sensitivity is still open: The first and most obvious possibility is that the variability across studies is due to the different stimuli. While those studies that reported greater metacognitive sensitivity of confidence employed tasks with fairly simple stimulus features such as motion and orientation, studies reporting the reversed pattern used either an object identification task or a masked face discrimination task. It is possible that confidence is associated with a greater metacognitive sensitivity than visual experience for very basic stimulus features only, while the effect is reversed with more complex stimuli. A second possibility relates to the different techniques of how subjective reports were recorded: While Sandberg et al. (2010) and Wierzchoń et al. (2014) provided participants with four labelled scale steps, participants in our own experiments operated a joystick to select a position on a VAS. It is possible that recording techniques interfere with the content of the subjective scales, for example, if participants are unable to report their visual experience in the same fine-grained manner as their decisional confidence. A previous study did not detect any effect of recoding technique on metacognitive sensitivity of motion experience (Rausch & Zehetleitner, 2014), but to our knowledge, no study so far has addressed this issue with respect to decisional confidence. A third possibility lies in the precise content of the scale assessing visual experience: While the scale in our study measured visual experience of the task-relevant feature, previous studies frequently used the perceptual awareness scale, which measures visual experiences of the task-relevant feature in conjunction with "brief glimpses", defined as "experiences without any content that cannot be defined any further" (Ramsøy & Overgaard, 2004). Thus, the surplus of metacognitive

sensitivity of visual experience could be driven entirely by experiences without content (see Rausch & Zehetleitner, 2014; Zehetleitner & Rausch, 2013, for more detailed discussions). Finally, although logistic regression and meta-d converged in our data, it is still possible that these methods would create conflicting results if applied to other data sets. Overall, further experiments would appear necessary to explore which of these options can explain the variability of previous studies concerning metacognitive sensitivity of experience and confidence.

4.3. How should we quantify metacognitive sensitivity?

Comparisons between previous studies on metacognitive sensitivity are limited due to the fact that there are several competing measures such as logistic regression, type 2 ROC analysis (Fleming et al., 2010), and meta-d (Maniscalco & Lau, 2012). Our reanalysis based on logistic regression and meta-d revealed a consistent effect of confidence versus experience as content of subjective reports across all three tasks, although a previous analysis based on type 2 ROC curves failed to detect an effect in the masked shape task (Zehetleitner & Rausch, 2013). Since the results of the present reanalysis are consistent across all three tasks and both methods, the most likely reason why we failed to find an effect in the previous 2 ROC analysis is lack of statistical power. Meta-d may be more powerful than type 2 ROC analysis due to the control of discrimination response biases or because it is possible to apply adjustments for extreme proportions (Hautus, 1995). Logistic regression analysis may benefit from the analysis being conducted on a single trial basis.

However, we observed two downsides to the use of logistic regression, owing to the fact that the relationship between subjective reports and the transformed accuracy was not linear, but often approached a lower bound instead. First, the slope of the regression curve changed over the range of the scale, tending towards zero at lower parts of the scale and increasing only at higher parts of the scale. As a consequence, there is no single logistic regression slope in each condition, and thus the interpretation of logistic regression slopes in terms of metacognitive sensitivity is ambiguous. Second, logistic regression may have a bias towards greater slopes with more conservative reports because the more liberal a scale is, the larger will be the part of the scale where the transformed accuracy is within the asymptotic range of performance; the more conservative a scale is, the larger will be the part of the scale where transformed accuracy increases. Indeed, in the masked shape task, we observed that the non-linear trend was confined to decisional confidence, the more liberal scale, and was absent in subjective reports of visual experience, which are known to be more conservative.

As control of subjective criteria is a critical feature of measures of metacognitive sensitivity (Barrett et al., 2013), and given that meta-d also controls discrimination bias and may provide increased statistical power, we recommend meta-d for all future studies where it can be applied.

5. Conclusion

We report that logistic regression and meta-d consistently indicated that subjective reports of confidence are more efficient in predicting trial accuracy than subjective reports of visual experience. Our data is consistent with the interpretation that participants consider stimulus features irrelevant to the current discrimination decision in addition to task-relevant ones for making subjective reports about their visual experience. We suggest that the choice of a scale to measure visual awareness should be based on theoretical considerations of exactly what are the conscious contents relevant for a particular research question. As we observed multiple non-linear relationships between subjective reports and the logit transform of accuracy, logistic regression is not a consistent and possibly biased measure of metacognitive sensitivity, which is why we recommend meta-d for future studies.

References

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods, 18*(4), 535-552. doi: 10.1037/a0033268

Bates, D. (2005). Fitting linear mixed models in R - Using the lme4 package. *R news, 5*(1), 27-30. doi: 10.1159/000351027

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5. Retrieved from http://CRAN.R-project.org/package=lme4

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2008). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution, 24*(3), 127-135. doi: 10.1016/j.tree.2008.10.008

Carota, A., & Calabrese, P. (2013). The achromatic 'philosophical zombie', a syndrome of cerebral achromatopsia with color anopsognosia. *Case reports in neurology, 5*(1), 98-103.

Carruthers, P. (2011). Higher-Order Theories of Consciousness. In E. N. Zalta (Ed.), *The Standford Encyclopedia of Philosophy* (Fall 2011 ed.).

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.

Christensen, R. H. B., & Brockhof, P. B. (2013). Analysis of sensory ratings data with cumulative link models. *Journal de la Société Française de Statistique, 154*(3), 58-79.

Dehaene, S. (2010). Conscious and nonconscious processes: Distinct forms of evidence accumulation? *Biological Physics, 60*, 141-168.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79*, 1-37.

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. (2004). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias. *Journal of Consciousness Studies, 11*, 25-45.

Dienes, Z. (2008). Subjective measures of unconscious knowledge. *Progress in Brain Research, 168*, 49-64.

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

Eriksen, C. W. (1960). Discrimination and learning without awareness: a methodological survey and evaluation. *Psychological Review, 67*, 279-300.

Fleming, S. M., & Lau, H. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 1-9. doi: 10.3389/fnhum.2014.00443

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.

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.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers, 27*, 46-51.

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences, 14*, 301-307. doi: 10.1016/j.tics.2010.04.006

Kuznetsova, A., Brockhof, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.0-11. Retrieved from http://CRAN.R-project.org/package=lmerTest

Lau, H. (2008). Are we studying consciousness yet? In L. Weißkrantz & M. Davies (Eds.), *Frontiers of Consciousness: Chichele Lectures*. Oxford: Oxford University Press.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*, 8-16. doi: 10.1016/j.tics.2011.05.009

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

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*, 420-430. doi: 10.1016/j.concog.2011.09.021

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.

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

Pinheiro, J., & Bates, D. M. (2000). *Mixed-effects models in S and S-plus*. New York: Springer.

R Core Team. (2013). R: A language and environment for statistical computing. (Version 2.15.2). 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.

Rausch, M., & Zehetleitner, M. (2014). A comparison between a visual analogue scale and a four-point scale as measures of conscious experience of motion. *Consciousness and Cognition, 28*, 126-140. doi: 10.1016/j.concog.2014.06.012

Reips, U.-D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in internet-based research: VAS generator. *Behavior Research Methods, 40*(3), 699-704. doi: 10.3758/BRM.40.3.699

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., & Overgaard, M. (2013). Measuring and testing awareness of emotional face expressions. *Consciousness and Cognition, 22*, 806-809. doi: 10.1016/j.concog.2012.12.003

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

Schlagbauer, B., Müller, H. J., Zehetleitner, M., & Geyer, T. (2012). Awareness in contextual cueing of visual search as measured with concurrent access- and phenomenal-consciousness tasks. *Journal of Vision, 12*(11), 1-12. doi: 10.1167/12.11.25

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

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.

Szczepanowski, R., Traczyk, J., Wierzchoń, M., & Cleeremans, A. (2013). The perception of visual emotion: Comparing different measures of awareness. *Consciousness and Cognition, 22*, 212-220. doi: 10.1016/j.concog.2012.12.003

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

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition, 27*, 109–120. doi: 10.1016/j.concog.2014.04.009

Zehetleitner, M., & Rausch, M. (2013). Being confident without seeing: What subjective measures of consciousness are about. *Attention, Perception, & Psychophysics, 75*, 1406-1426. doi: 10.3758/s13414-013-0505-2

Table 1

Results of a logistic mixed model regression for accuracy across experiments

Experiment	Effect	*B*	*95% CI*		χ^2	*p*
			Lower	Upper		
Masked orientation task	First report	0.71	0.52	0.90	60.1	<.001
	Second report	0.37	0.17	0.57	20.4	<.001
	SOA	0.71	0.54	0.89	60.3	<.001
	Scale	-0.42	-0.70	-0.14	7.1	<.01
	First report * second report	0.23	0.04	0.41	2.1	n.s.
	First report * SOA	0.38	0.17	0.58	3.6	n.s.
	First report * scale†	0.21	-0.20	0.62	3.4	n.s.
	Second report * SOA	0.35	0.15	0.55	6.7	<.01
	Second report * scale†	-0.93	-1.35	-0.51	13.3	<.001
	SOA * scale	-0.19	-0.53	0.15	0.1	n.s.
	First report * second report * SOA	0.33	0.16	0.51	13.3	<.001
	First report * second report * scale	-0.41	-0.77	-0.04	3.9	<.05
	First report * SOA * scale	0.06	-0.35	0.47	0.8	n.s.
	Second report * SOA * scale	-0.42	-0.83	-0.01	2.9	n.s.
Masked shape task	First report	0.71	0.50	0.92	44.3	<.001
	Second report	0.36	0.15	0.57	31.4	<.001
	SOA	0.85	0.72	0.98	310.5	<.001
	Scale	0.54	0.16	0.93	2.4	n.s.
	First report * second report	0.22	0.06	0.37	3.5	n.s.
	First report * SOA	0.42	0.21	0.63	5.9	< .05
	First report * scale†	1.05	0.63	1.46	12.8	<.001
	Second report * SOA	0.26	0.07	0.45	23.0	<.001
	Second report * scale†	-0.72	-1.14	-0.31	6.4	< .05
	First report * second report * SOA	0.21	0.06	0.36	8.7	<.01
	First report * second report * scale	0.14	-0.18	0.45	0.6	n.s.
	First report * SOA * scale	1.02	0.60	1.44	27.5	<.001
	Second report * SOA * scale	-0.63	-1.00	-0.25	10.8	<.001
Motion discrimination task	First report	0.44	0.25	0.63	20.3	<.001
	Second report	0.38	0.21	0.56	28.3	<.001
	Coherence	1.15	1.02	1.27	518.4	<.001
	Scale	-0.08	-0.49	0.34	1.0	n.s.
	First report * second report	0.03	-0.09	0.15	0.0	n.s.

First report * coherence	0.17	-0.01	0.35	3.0	n.s.
First report * scale†	0.29	-0.08	0.66	7.2	<.01
Second report * coherence	0.38	0.21	0.56	23.9	<.001
Second report * scale†	-0.59	-0.95	-0.24	12.6	<.001
First report * second report * coherence	0.04	-0.08	0.15	0.3	n.s.
First report * second report * scale	0.03	-0.21	0.27	0.1	n.s.
First report * coherence * scale	-0.15	-0.50	0.20	1.0	n.s.
Second report * coherence * scale	-0.19	-0.53	0.16	1.1	n.s.

†Note that the effect of scale codes if the report of confidence was collected before the report of experience or vice versa. Consequently, a positive coefficient of the first report * scale interaction effect indicates that confidence predicted accuracy more efficiently than experience, whereas a positive coefficient of the second report * scale indicates just the reverse pattern.

Table 2

Results of a linear mixed regression model for meta-d across experiments

Experiment	Effect	*B*	*95% CI*		|*t*|	*df*	*p*
			Lower	Upper			
Masked orientation task	Scale	-0.62	-0.75	-0.49	9.3	772.9	<.001
	SOA	0.82	0.75	0.88	24.7	772.9	<.001
	Time	0.01	-0.12	0.14	0.2	772.9	n.s.
	Scale * SOA	-0.30	-0.43	-0.17	4.5	772.9	<.001
	Scale * Time	0.24	-0.02	0.50	1.8	772.9	n.s.
	SOA * Time	-0.10	-0.23	0.03	1.6	772.9	n.s.
	Scale * SOA * Time	0.20	-0.06	0.46	1.5	772.9	n.s.
Masked shape task	Scale	-0.42	-0.60	-0.23	4.4	234.0	<.001
	SOA	0.99	0.90	1.08	20.7	234.0	<.001
	Time	0.00	-0.19	0.19	0.0	234.0	n.s.
	Scale * SOA	-0.47	-0.66	-0.28	4.9	234.0	<.001
	Scale * Time	0.16	-0.77	1.10	0.3	14.0	n.s.
	SOA * Time	-0.01	-0.20	0.18	0.1	234.0	n.s.
	Scale * SOA * Time	-0.26	-0.63	0.12	1.4	234.0	n.s.
Motion discrimination task	Scale	-0.28	-0.46	-0.11	3.1	267.0	<.01
	Coherence	1.13	1.04	1.21	25.1	267.0	<.001
	Time	-0.03	-0.20	0.15	0.3	267.0	n.s.
	Scale * Coherence	-0.10	-0.27	0.08	1.1	267.0	n.s.
	Scale * Time	0.21	-0.46	0.86	0.6	19.0	n.s.
	Coherence * Time	-0.05	-0.23	0.12	0.6	267.0	n.s.
	Scale * Coherence * Time	0.16	-0.20	0.51	0.9	267.0	n.s.

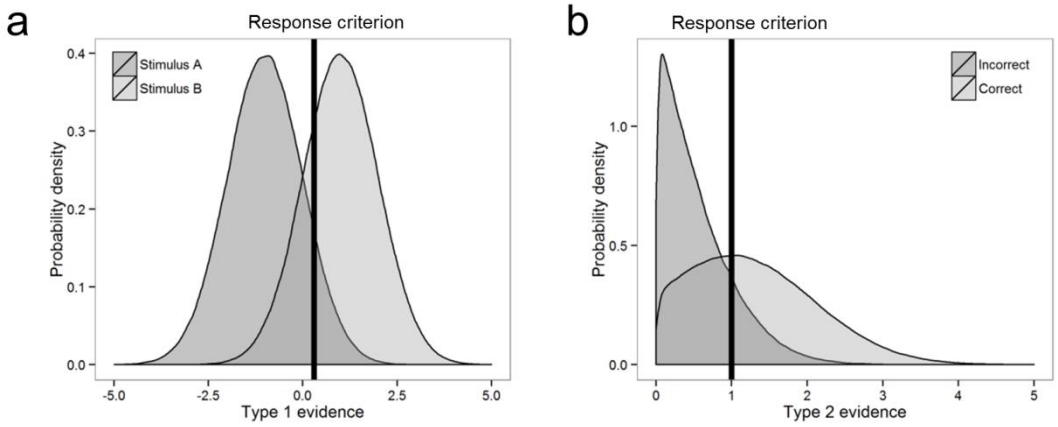

Figure 1. Signal detection theory. (a) Distributions of evidence created by the two stimuli A and B in a type 1 task, i.e. the observers' task is to decide which one of the two stimuli has been presented. When the type 1 evidence is greater than the response criterion, observers respond "B", and "A" otherwise. (b) Distributions of evidence created by correct and incorrect trials in a type 2 task, i.e. the observers' task is to decide if the preceding judgment was correct. Note that the decision process is analogous to a type 1 task except the distributions of evidence created by correct and incorrect trials are expected to deviate strongly from the normal distribution.

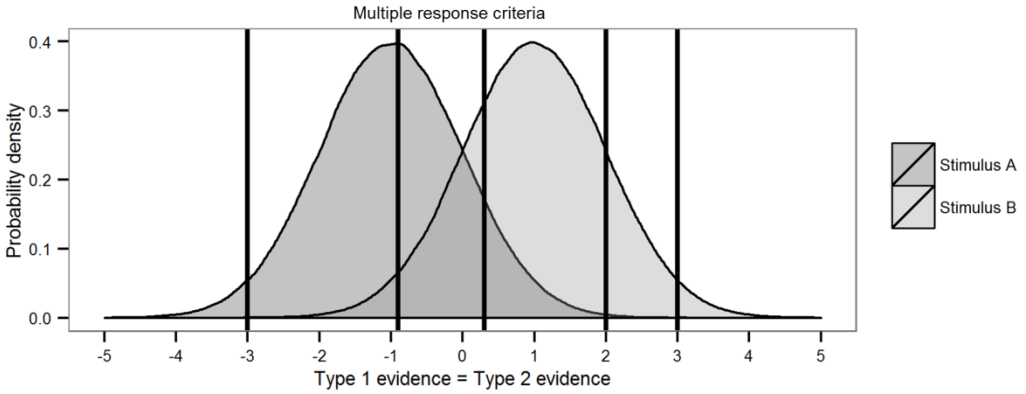

Figure 2. Signal detection model underlying meta-d. Meta-d is computed assuming metacognition is ideal, i.e. the same evidence is available for subjective reports than for discrimination judgments. The model is the same as a standard SDT model for a type 1 task, except that discrimination decisions and subjective reports are assumed to form one dimension of response options, i.e. “It is A for sure”, “I’m guessing A”, I’m guessing B”, “It is B for sure”, delineated by several response criteria.

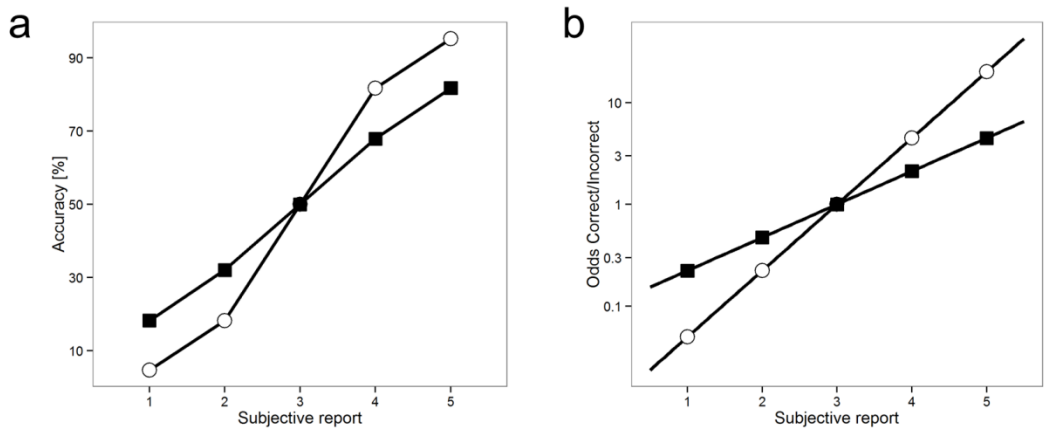

Figure 3. Quantifying the relationship between trial accuracy and subjective reports by logistic regression. (a) Data of a hypothetical experiment. Task accuracy in % correct is plotted as a function of subjective report. Lines indicate two separate conditions. (b) Same data with accuracy transformed into the odds of being correct to incorrect and plotted on log-scale. Logistic regression is based on fitting a linear function on such transformed data. The more subjective reports differentiate between different levels of accuracy, the steeper the slopes of the regression line will be. Note that such a linear relationship is unlikely to occur in real data.

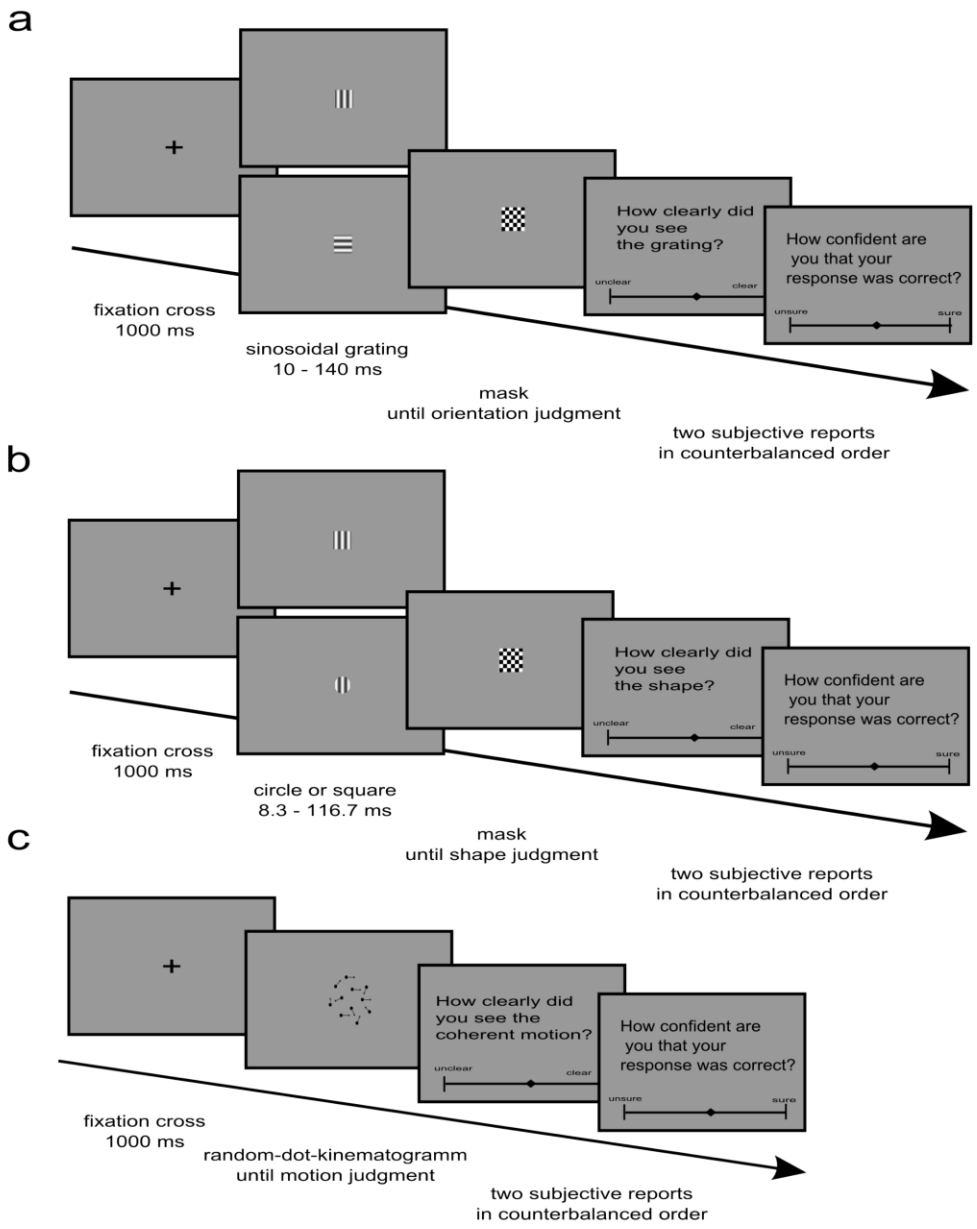

Figure 4. Trial sequence for (a) the masked orientation task, (b) the masked shape discrimination task, and (c) the random-dot motion discrimination task.

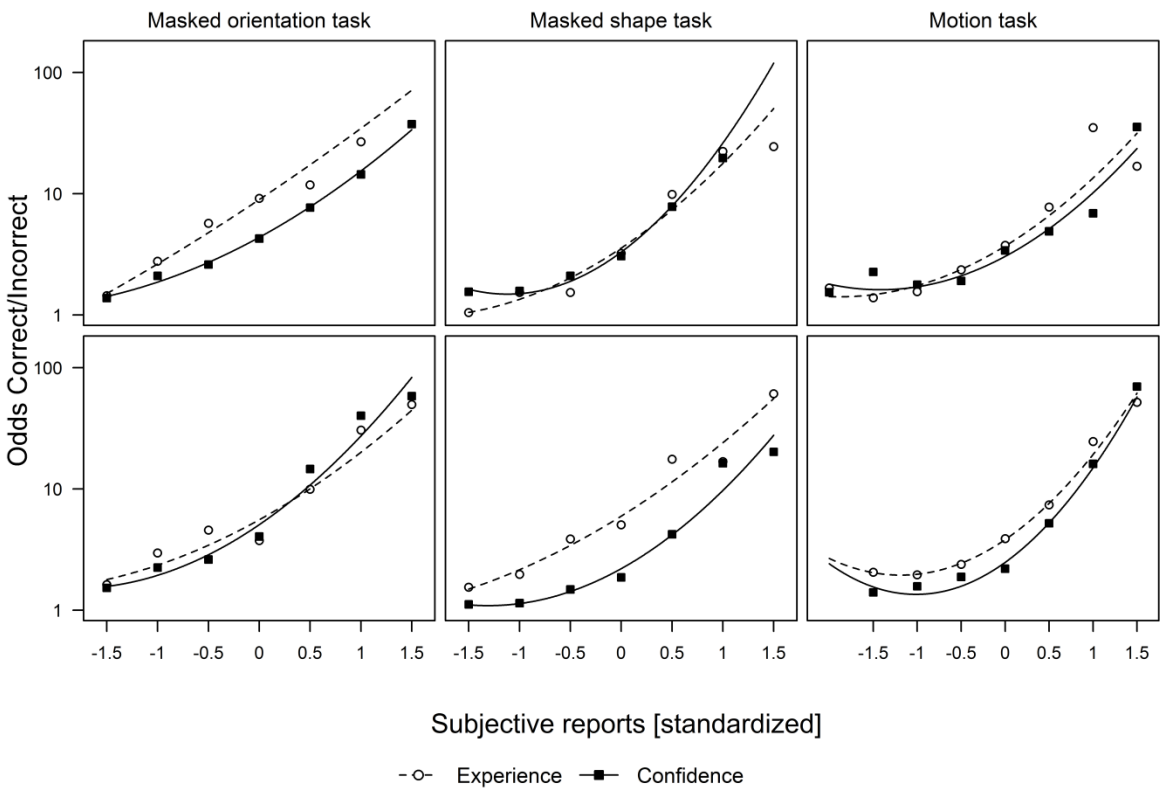

Figure 5. Relationship between subjective report and the odds of being correct, separately for scale, experiment, and time of the report. Upper row: First subjective report within one trial, Lower row: Second subjective report within one trial. Lines indicate the prediction from logistic regression models including quadratic effects.

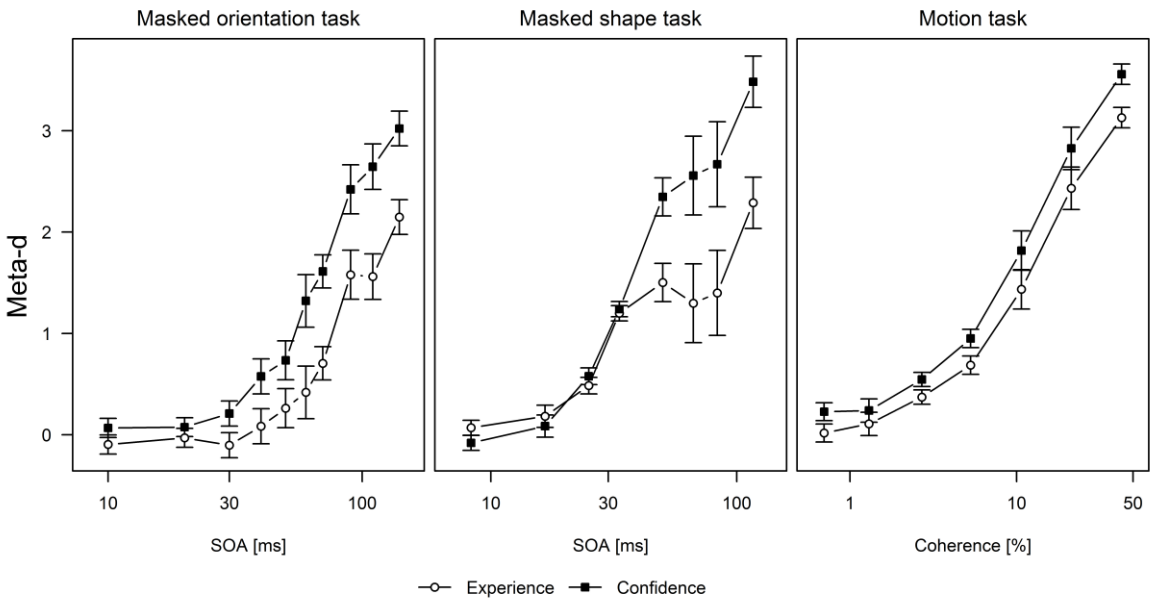

Figure 6. Meta-d as a function of stimulus quality, separately for each task in separate panels and scale as separate lines.