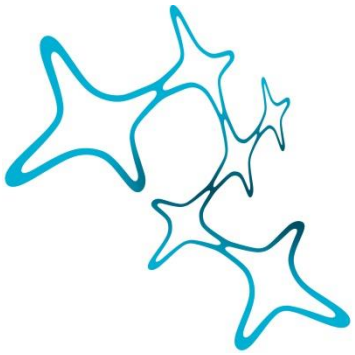

Graduate School of
Systemic Neurosciences
LMU Munich

Content, granularity, and type 2 sensitivity of subjective measures of visual consciousness

Manuel Rausch

Dissertation
at the Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

submitted by
Manuel Rausch
from Munich

Munich, 1 October, 2015

Day of disputation: 18 April 2016

Thesis advisory committee:

Prof. Michael Zehetleitner

Prof. Hermann J. Müller

Prof. Stephan Sellmaier

Examination committee:

Prof. Michael Zehetleitner

Prof. Stephan Glasauer

Prof. Hermann J. Müller

Prof. Marco Steinhauser

Keeper of the minutes

Prof. Marco Steinhauser

ACKNOWLEDGEMENTS

Mein erster Dank gilt Saskia, und zwar dafür, dass sie einerseits tolerierte, wenn meine Gedanken auch nach Feierabend um Forschung kreisten, und andererseits sichergestellt hat, dass ich mich nicht ausschließlich mit Forschung beschäftigte.

Desweiteren danke ich Michael Zehetleitner dafür, dass er mir diese Dissertation abseits der Standardthemen der Psychologie ermöglichte, und für das große Vertrauen, das er mir und meiner Arbeit stets entgegen brachte.

Außerdem danke ich Hermann Müller für seine Untersützung immer dann, wenn sie vonnöten war.

Bernhard Schlagbauer danke ich für hilfreiche Kommentare und für die Gesellschafft bei Kaffee, Mensa, und Whisky.

Emil Ratko-Dehnert danke ich dafür, dass er mir Progammieren beigebracht hat, und für die Erweiterung meines gustatorischen Horizontes.

Finally, I would like to express my gratitude towards the whole Graduate School of Systemic Neurosciences for providing such a stimulating multidisciplinary environment.

ABSTRACT

According to several major theories in the field of consciousness research, the valid assessment of conscious awareness requires subjective measures, i.e. participants' reports about their conscious experience. However, there is a considerable amount of uncertainty in the field if and how scientifically valuable data can be obtained from subjective measures.

The present work empirically examines how subjective measures of conscious awareness need to be designed and applied to provide maximally useful data for empirical studies of visual consciousness. Specifically, it is investigated what contents subjective measures should require participants to report, at which granularity subjective measures ought to be recorded, and what statistical procedures should be used to quantify the relation between subjective measures and discrimination task performance.

Concerning content, subjective measures that referred to the accuracy of a preceding discrimination response and subjective measures referring to participants' visual experience of the task-relevant stimulus feature were compared during a series of visual psychophysical experiments. Subjective measures about the accuracy of the responses were associated with more liberal psychophysical thresholds: At lower stimulus quality, participants reported that they feel confident that their discrimination response was correct without reporting a visual experience of the stimulus feature. Only at greater stimulus quality, they reported that they had a visual experience of the stimulus feature in addition to being confident. Moreover, subjective measures about confidence in discrimination responses predicted task accuracy more efficiently than measures about visual experience. Finally, subjective measures of experience and task accuracy as content were compared while event-related potentials (ERP) were recorded. The earliest electrophysiological correlates of subjective measures where predictive of the fact if participants reported that they selected the response to the discrimination task based on knowledge instead of guessing, but were not yet predictive whether participants reported a clear experience over and above making the task response based on knowledge. The strongest ERP correlate of visual experience occurred a short period in time before participants responded to the discrimination task. As a consequence, it is argued that conceptual considerations are required which conscious contents are relevant for a specific research question, and subjective measures should be about the relevant contents accordingly.

Concerning the granularity of subjective measures, a continuous scale and a scale with four discrete labelled categories were compared as subjective measure of conscious experience of motion. The subjective measures contained more information when participants used the continuous scale instead of the discrete scales. The greater amount of information provided by continuous scales rendered subjective measures more predictive of task accuracy and enhanced internal consistency.

Regarding the statistical procedure to quantify the relation between subjective measures and task performance, it was found that logistic regression is a suboptimal method because the relationship between subjective measures and the transformed accuracy was frequently not linear. In contrast, meta-d_a, a measure of the relationship between subjective reports and task accuracy derived from signal detection theory (SDT), provided the most consistent results across all studies.

Overall, it is concluded that subjective measures are suited to provide highly useful data to address non-trivial research questions for the scientific study of consciousness: As prerequisite, the content of a subjective measures should be tailored to the current research question. In addition, the problem of a lacking objective standard can be addressed by using the relation between subjective measures and task performance as a reference frame.

TABLE OF CONTENTS

1. INTRODUCTION

The neural correlates of consciousness (NCC) is probably the most prominent yet unresolved problem in modern neuroscience (Crick & Koch, 1990, 2003; Rees, Kreiman, & Koch, 2002). According to a standard definition, the NCC is the minimum set of neural events jointly sufficient to give rise to a specific conscious experience (Mormann & Koch, 2007). Empirical studies of the NCC are usually based on the same principle, a comparison of two different kinds of measurements: (i) a measurement of on-going neural events, and (ii) a measurement of the conscious experiences of the subject given a specific stimulus or situation. A major obstacle to the prosperity of the field of consciousness research is that there is great disagreement on how to measure the latter, i.e. conscious experience (Chalmers, 1998).

What is the appropriate measurement for consciousness research? First and foremost, choosing a measurement of consciousness poses a conceptual question: Several different concepts of consciousness exist, each providing a different definition of what it means to be conscious (for overviews see e.g. Block, 2002; Rosenthal, 2009; van Gulick, 2014). Some of these concepts imply specific measurement procedures by implying that specific behaviours are indicative of consciousness (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). Conceptual clarity is not only important for identifying appropriate behavioural measures; in fact, different concepts of consciousness may be realized by separate NCCs (Block, 2005; Rees et al., 2002). As a consequence, empirical studies on consciousness need to clarify first which concept of consciousness is addressed by their research.

However, while the key features of a measure of consciousness can be determined by conceptual considerations, many specific features cannot be deduced from the concepts. For example, higher order theories of consciousness typically imply that visual consciousness needs to be assessed by some kind of subjective measure (Dienes, 2004, 2008; Lau & Rosenthal, 2011; Lau, 2008b; c.f. section 1.2.3.), but they do not specify whether visual consciousness should be measured by a visibility rating, a confidence rating, whether reports should be made at a continuous or a multi-category scale, or whether joystick or keyboards should be used as measurement device. Whenever several measures are conceptually equally valid, empirical studies are required to investigate how a measurement needs to be designed

to provide the most useful data (Dienes & Seth, 2010; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010).

This work addresses the basic research question how subjective measures need to be designed to provide useful measures of visual consciousness. It begins with a discussion of the distinctive features of subjective measures in comparison to so-called objective measures, which concepts of consciousness can be assessed by subjective measures, and what is the rationale to prefer subjective measures over objective ones. Subsequently, three key characteristics of subjective measures are empirically investigated:

(i) the content of subjective measures
(ii) the granularity of subjective measures
(iii) the method to quantify the relation between subjective measures and objective task performance

Regarding content, it is investigated whether subjective measures that refer to the visual experience of the stimulus are interchangeable with subjective measures that refer to participants' confidence in having made a correct response at the discrimination task. The effect of content of subjective measures is examined in terms of behaviour in psychophysical experiments (cf. Chapter 2; Zehetleitner & Rausch, 2013) and event-related potential (ERP) correlates (Chapter 3). Regarding the granularity of subjective measures, it is investigated if subjective reports should preferably be recorded by scales with four discrete categories, or by a visual analogue scale (VAS; Chapter 4; Rausch & Zehetleitner, 2014). Finally, concerning the association between subjective measures and discrimination performance, it is examined if mixed-model logistic regression or the recently suggested meta-d_a is a more convenient method of analysis (Chapter 5; Rausch, Müller, & Zehetleitner, 2015).

1.1. What are subjective measures of conscious awareness?

Two general approaches to measuring conscious awareness can be distinguished: *objective measures* and *subjective measures* (Cheesman & Merikle, 1984; Lau, 2008; Seth et al., 2008). A measure is considered objective if conscious experiences are ascribed to the subject based on performance in a task (Eriksen, 1960). For example, assuming that participants are presented with one out of two possible stimuli "A" and "B", the participants are said to be conscious of the stimulus if they respond correctly more often than expected from chance when asked to classify the stimulus. In contrast, measures are considered as

subjective if participants are required to make a report about their conscious experiences (Cheesman & Merikle, 1984). In this case, the participants might be asked if they are seeing the stimulus, and it is assumed that they have a visual experience if they respond "yes".

According to a standard view in the consciousness studies literature, the fundamental difference between objective and subjective measures lies in the kind of processes indicated by each measure (Dienes, 2008). Fig. 1-1 provides an overview of the hypothesized causal pathways leading to a discrimination response (i.e. an objective measure) as well as subjective report (i.e. a subjective measure). Discrimination performance above chance shows that the visual system of the observer provided at least some sensory evidence, which was available to the processes engaged in decision making. The decision must also have successfully triggered a motor response. In contrast, subjective measures require participants to report their mental state directly (Dienes, 2004; Seth et al., 2008), for instance their visual experience of the stimulus. Such a report requires the participant to make a decision between the different response options offered by a subjective measure (called *rating decision* in Fig. 1-1). This rating decision depends on metacognitive processes in the sense that participants need to know they saw the stimulus in order to report that they saw the stimulus (Dienes, 2008). Some of the causal pathways underlying rating decisions are yet controversial: Standard signal detection theory assumes that rating decision are made based on the same sensory evidence as the discrimination decision (Kepecs, Uchida, Zariwala, & Mainen, 2008; Ko & Lau, 2012; Macmillan & Creelman, 2005). Other models have suggested a parallel causal pathway for sensory evidence to rating decisions bypassing the discrimination decision (Cleeremans, Timmermans, & Pasquali, 2007; Dehaene, 2010). Some models unify both accounts in assuming rating decisions are influenced by both sensory evidence used in the discrimination task as well as parallel sensory evidence (Jang, Wallsten, & Huber, 2012; Pleskac & Busemeyer, 2010). Finally, recent studies proposed that rating decisions are influenced by the motor action of the previous discrimination response (Fleming et al., 2015). Overall, the mechanisms underlying subjective measures may involve metacognitive components in addition to those processes engaged in objective measures.

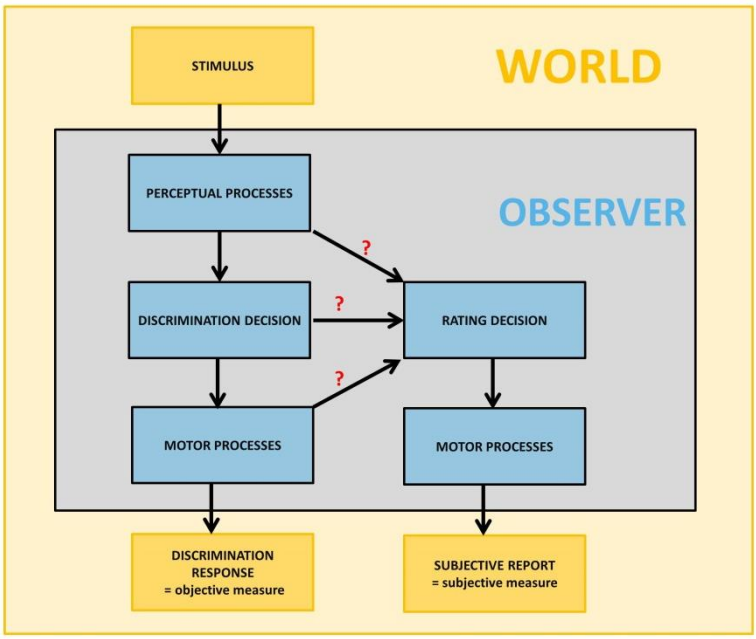

Figure 1-1. A schematic view of the causal pathways underlying objective and subjective measures of consciousness. A discrimination decision is a decision process where participants select one out of the possible responses to the task, based on sensory evidence provided by perceptual processes. A rating decision is a process where participants select one out of several options provided by the rating scale. The causal pathways leading to rating decisions are still controversial, as indicated by the question marks.

Besides these differences, Fig. 1-1 also illustrates an important feature objective and subjective measures have in common: Both kinds of measures ultimately depend on an overt behavioural response by the participant. For objective and subjective measures, the most common response in the context of an experiment is that the participant presses one out of several buttons. This is important to acknowledge because subjective measures seem to depend on introspection, i.e. the observation of one's own mind, which has been exiled from scientific psychology for decades (Boring, 1953; Danziger, 1980). However, participants' subjective report recorded and interpreted by the experimenter is a physical fact about the world just in the same way as a discrimination judgement, public and thus readily accessible for scientific investigations. Consequently, subjective reports about conscious experience are legitimate data for science as long as the experimenter and the participant who engages in introspection are not the same person (Dennett, 2003, 2007). Even more, it was argued that subjective reports about conscious experience are exactly those events that a science of consciousness should collect as data and strive to explain (Dehaene & Naccache, 2001; Dehaene, 2010; Dennett, 2003, 2007).

1.2. Concepts of consciousness

Before discussing whether subjective and objective measures are more appropriate for a specific research question, researchers should always clarify first which concept of consciousness is investigated in a specific study. Concepts of consciousness fall into the two superordinate categories of transitive consciousness and intransitive consciousness (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006). When used in the intransitive sense, consciousness refers to a person or to an agent as a whole: Intransitive consciousness differentiates awake humans from non-conscious entities (e.g. robots, philosophical zombies) or other humans in dreamless sleep or coma. In the transitive sense, consciousness always refers to events or mental states one is conscious or aware *of* (Rosenthal, 1986, 2009; Van Gulick, 2014): Transitive consciousness differentiates conscious stimuli or perception from unconscious stimuli or perception. Researchers often use the term *awareness* or *conscious awareness* to explicitly refer to the transitive meaning of consciousness of something. Both superordinate categories involve various different concepts (see Van Gulick, 2014, for an overview). As the present work is only concerned with subjective measures of conscious awareness of visual stimuli, only the three most influential concepts of transitive consciousness for contemporary research and the corresponding theories of consciousness are described briefly:

(i) phenomenal consciousness (Block, 2002; Jackson, 1982; Nagel, 1974)
(ii) conscious access (Baars, 2002, 2005; Block, 2002; Dehaene et al., 2006; Dehaene & Naccache, 2001)
(iii) higher-order consciousness (Carruthers, 2011; Lau & Rosenthal, 2011; Timmermans, Schilbach, Pasquali, & Cleeremans, 2012)

For each of these concepts, it is discussed whether it can be adequately assessed by means of subjective measures.

1.2.1. Phenomenal Consciousness

Conscious awareness in the sense of phenomenal consciousness is defined as what-it-is-like to have an experience of an external stimulus or an inner event. What-it-is-likeness, the qualitative aspect of experience, always depends on the first person perspective of the observer (Nagel, 1974). If conscious awareness is understood along the lines of phenomenal consciousness, observers are conscious of a stimulus if they have first person experiences of

the stimulus (Block, 2002; Chalmers, 1994; Nagel, 1974). Concerning the NCC, proponents of phenomenal consciousness often defend a sensory cortex view, assuming that conscious experience is instantiated by brain events during sensory processing of the stimulus, for instance recurrent processing along the visual ventral stream (Block, 2005; Lamme, 2006).

Is there a possibility of measuring phenomenal experience by subjective measures? According to one view, in order to really know what a specific experience is like, it is necessary to experience that particular sensation oneself (Jackson, 1982). As phenomenal consciousness is defined as experience from the first person perspective (Block, 2002; Chalmers, 1994; Nagel, 1974), but science is an endeavour that crucially relies on the third person perspective (Dennett, 2007), it appears unlikely that measurements from the third-person perspective are ever able to measure phenomenal consciousness in a scientifically appropriate way, irrespective of the measures being subjective or objective. However, although phenomenal experience cannot be accessed from third person perspective, it may be possible that phenomenal experience can be shared (Velmans, 2000, 2007): This view maintains that if the perceptual and cognitive apparatus of different observers is similar, and if different observers give similar reports of what they experience, it is reasonable to assume their conscious experience of a given stimulus is also similar. This approach does not aim for objectivity in the sense of measuring conscious experience independently from the observer. Instead, it investigates agreement between different first-person perspectives, i.e. intersubjectivity. However, inferences about first-person experiences based on subjective measures relies on the assumption that similar perceptual and cognitive apparatuses give similar experiences; and while this intuition may appear reasonable to many, the argument is not compelling to those who do not share this intuition.

1.2.2. Conscious access

According to the second concept of consciousness, termed conscious access or access consciousness, a representation is conscious if it is available for reasoning and direct control of action (including reporting) (Baars, 2002; Block, 2002; Dehaene et al., 2006). This means for visual consciousness that observers perceive a stimulus consciously if the contents generated by the visual system can be used to guide behaviour based on reason and if the contents can be reported. Conscious access is the concept of consciousness underlying global workspace theory (Baars, 2002; Dehaene et al., 2006; Dehaene & Naccache, 2001). The global workspace theory assumes that representations encapsulated in brain systems operating

in parallel are unconscious. A global workspace is instantiated by broadcasting representations to multiple other brain systems. The global availability of representations through the workspace is subjectively experienced as a conscious state (Dehaene & Naccache, 2001). The neural implementation of conscious access is an "ignition"-like spread of neural activity from sensory cortices to a fronto-parietal network resulting in a widely distributed NCC (Dehaene et al., 2006).

Are subjective measures eligible to measure conscious access to visual contents? As reportability is among the two cognitive functions that are included in the definition of conscious access, subjective reports are a natural choice of measuring conscious awareness from the viewpoint of conscious access and global workspace theory (Dehaene & Naccache, 2001; Dehaene, 2010). If visual contents are reportable, it means that visual contents cannot be encapsulated into the visual system; instead, it shows that the visual contents are available to decision making and language as required for rating decisions (Dienes, 2008). Overall, subjective measures are adequate measures of conscious access.

1.2.3. Higher-order consciousness

According to the third prominent concept of consciousness, an observer is conscious of a particular mental state if the mental state is accompanied by a higher-order mental state that represents the observer as being in a particular mental state (Carruthers, 2011; Lau & Rosenthal, 2011). According to different flavours of higher-order consciousness, this higher order mental state can be a thought (Rosenthal, 1986), a percept (Lycan, 2004), or a statistical inference (Lau, 2008a). In terms of visual consciousness, observers perceive a stimulus consciously only if they possess a mental state that represents them as perceiving the stimulus, i.e. observers need to know/sense/infer that they perceive the stimulus. As these higher-order mental states require metacognition, higher order theories predict a close connection between consciousness and metacognitive systems (Rosenthal, 2000). Consequently, the neural correlates of higher-order consciousness are assumed to be similar to those of metacognition and thus located in frontal and parietal cortex (Lau & Rosenthal, 2011).

Endorsing a higher-order consciousness is the most common theoretical background for using subjective measure of conscious awareness because subjective measures can be seen as a test of higher-order mental states (Dienes, 2004, 2008): In order to report that they saw

the stimulus, participants need to know, sense or infer that they saw the stimulus. The metacognitive aspect of subjective measures makes subjective measures the most valid measure of conscious awareness in the higher-order sense (Lau, 2008b).

1.3. Why objective measures of conscious awareness are not sufficient

Given the fact that objective measures have dominated cognitive psychology for the second half of the 20th century (Boring, 1953; Danziger, 1980; Eriksen, 1960), it may appear surprising why more and more researchers, even from different theoretical perspectives, have come to argue for using subjective measures in consciousness research (Dehaene & Naccache, 2001; Dienes, 2008; Lau, 2008b; Ramsøy & Overgaard, 2004). The main argument why objective measures need to be accompanied by subjective measures starts from the premise that conscious experiences are not necessarily in accordance with performance in discrimination tasks (Lau, 2008b; Seth et al., 2008). The reason is that all three major theories of consciousness allow for the possibility that conscious awareness is in disagreement with discrimination task performance (Dienes, 2008).

First, there may be cases where discrimination performance is above chance in absence of conscious awareness. The standard example is blindsight, which is caused by lesions to primary visual cortex. These patients report to be blind in the visual hemifield contralateral to the damaged brain area. Despite their apparent blindness to stimuli presented in their visual field corresponding to the lesion, these patients are able to perform well above chance in forced-choice tasks on stimuli presented in regions in space where they report to be blind (Weiskrantz, 1986). Proponents of phenomenal consciousness believe this is possible because the occipital lesion destroys pathways necessary for conscious experience, while discrimination performance is supported by neural circuits not part of the NCC (Lamme, 2006). Global workspace theory assumes that blindsight is mediated by a processing stream that does not trigger a global workspace and thus occurs in absence of conscious awareness (Dehaene & Naccache, 2001). Finally, higher-order theories assume that blindsight patients fail to have an accurate metarepresentation of their perceptual capacities (Lau, 2008a).

Second, there may also be cases where conscious experience exceeds the manifest discrimination performance. For example, when participants are presented with arrays of several letters, observers report they can see all or almost all letters, but typically are able to report no more than 3 to 4 of the letters (Block, 2011; Sperling, 1960). Conscious experience

without discrimination performance may occur when contents in visual short-term visual memory associated with phenomenology are overwritten by new stimuli before they can be transferred to working memory (Block, 2011; Vandenbroucke, Sligte, & Lamme, 2011). According to higher-order theories, conscious experience without performance can be explained by illusory metacognition of percepts at unattended locations (Lau & Rosenthal, 2011). Conscious experience without discrimination performance is however not consistent with global workspace theory (Kouider, de Gardelle, Sackur, & Dupoux, 2010). Nevertheless, all three major concepts of consciousness predict at least some instances where discrimination performance deviates from conscious awareness.

If discrimination performance alone is not sufficient to measure conscious awareness, can conscious awareness be assessed without asking the participant to make a response? So-called no-report paradigms promise to disentangle the NCC from the correlates of reports by instructing the participant not to make a response at all (Frässle, Sommer, Jansen, Naber, & Einhäuser, 2014; Pitts, Martínez, & Hillyard, 2012; Pitts, Metzler, & Hillyard, 2014). One flavour of these experiments relies on a modified inattentional blindness paradigm (Pitts et al., 2012): In the first phase of the experiment, participants perform a demanding perceptual task while task-irrelevant line segments forming a configuration are presented. In a following interview, participants are interrogated about their awareness of the configurations and thus their attention is directed towards the configurations when the task is repeated. Finally, awareness of the configurations is assessed by interview a second time. The idea is that participants are unaware of the configuration in the first phase of the experiment, but after the first interview they become aware of the stimuli. However, the whole "no-report" paradigm does rely heavily on subjective measures, because integral parts of the experiment are the two interviews. The participants' answers during the interview are subjective reports and thus subjective measures of consciousness. The second flavour of no-report paradigms tries to replace subjective reports during the experiment by physiological markers: As the alternations of conscious experiences during binocular rivalry are correlated with pupil size and opto-kinetic nystagmus, Frässle et al. (2014) suggested to replace subjective measures by supposedly more objective eye-tracking based measures. However, eye-tracking based measures may not be the best choice as physiological marker of awareness because multiple dissociations between conscious awareness and eye movements exist (Spering & Carrasco, 2015). In general, this method requires a validation phase for each physiological marker, where the close correlation between conscious experiences and the marker is established. For

this purpose, again subjective measures of conscious awareness are typically used. As discussed above, discrimination performance cannot be used to establish a physiological marker because discrimination performance does not necessary indicate that observers are also conscious of the stimulus. Overall, no-report paradigms avoid subjective measures only at the point in time when neural processes are measured, while subjective measures are still an integral part of the experiments.

When dissociations between subjective reports and objective measures are observed, many empirical researchers tend to place their trust on the objective measure (e.g. Eriksen, 1960; Hannula, Simons, & Cohen, 2005; Shanks & St. John, 1994). However, all existing concepts of consciousness imply that dissociations between objective measures and conscious awareness may exist. As a consequence, objective measures should always be accompanied by subjective measures (Dehaene & Naccache, 2001; Dehaene, 2010; Dienes, 2004, 2008; Lau, 2008b; Ramsøy & Overgaard, 2004).

1.4. Can subjective measures provide scientifically useful data?

The term "subjective" in subjective measures has the negative connotations of "biased" and "unscientific". Unsurprisingly, subjective measures have not been considered as serious scientific data for most of the 20th century (Boring, 1953; Danziger, 1980; Eriksen, 1960), and many still consider subjective measures as problematic (Hannula et al., 2005; Irvine, 2012; Schmidt & Vorberg, 2006). Opponents of subjective measures usually defend their view with variants of the following arguments:

(i) Language and words may be inadequate to reflect conscious experiences (Eriksen, 1960).

(ii) Subjective measures depend on applying criteria (Block, 2005; Irvine, 2012; Schmidt & Vorberg, 2006).

(iii) Subjective measures do not exhaustively detect all instances where participants are aware of the stimulation (Hannula et al., 2005; Persaud, McLeod, & Cowey, 2007; Shanks & St. John, 1994).

Critique (i) argues that words and language, and consequently subjective measures, are inadequate to reflect consciousness experiences. As Eriksen (1960) pointed out, words are highly abstract symbols with no physical resemblance to the relationships they denote. This argument seems to apply to a concept of consciousness as phenomenal consciousness

inaccessible to verbal report. After all, there is no principle reason why participants are unable to verbalize the characteristic relationships of conscious access and higher order consciousness. However, it seems questionable why inaccessible phenomenal consciousness is more adequately assessed by objective measures than by subjective ones. After all, the metrics of discrimination performance is also quite different from the metrics of phenomenal experience. If conscious experiences are not associated with subjective reports, there is little reason to believe that conscious experiences are associated with discrimination responses. If indeed the only way to know what a specific experience is like is to share it (Jackson, 1982), neither subjective nor objective measures are adequate measures of conscious experience; there will probably never be a satisfactory way to measure conscious experience at all. If a more optimistic view on the human ability to communicate their phenomenal experiences is adopted (Velmans, 2000, 2007), words and language appear to be a natural choice as medium of communication. After all, language is the primary mechanism of communication of humankind. Some theories even suggest that perceptual awareness evolved to enhance communication with other humans (Frith, 2011; Graziano & Kastner, 2011).

A more modest interpretation of argument (i) asserts that the translation of conscious experiences into a subjective measure is not impossible, but prone to errors. However, this is not a principle argument against the use of subjective measures; instead, it suggests that researchers should not blindly trust in their subjective measures (Velmans, 2007). At a consequence, a series of studies attempted to identify the most suitable subjective measures of conscious awareness empirically (Dienes & Seth, 2010; Rausch & Zehetleitner, 2014; Sandberg, Bibby, Timmermans, Cleeremans, & Overgaard, 2011; Sandberg et al., 2010; Szczepanowski, Traczyk, Wierzchoń, & Cleeremans, 2013; Wierzchoń, Asanowicz, Paulewicz, & Cleeremans, 2012; Wierzchoń, Paulewicz, Asanowicz, Timmermans, & Cleeremans, 2014). How can the eligibility of a specific subjective measure be evaluated? The first possibility is to assess the correlation between subjective measures and objective discrimination performance: If there is a correlation between subjective measures and task performance, it means that the sensory systems were able to provide some sensory evidence, which directly or indirectly has propagated to the processes engaged in the rating decision. Consequently, the degree to which subjective measures predict trial accuracy, a.k.a. type 2 sensitivity (Galvin, Podd, Drga, & Whitmore, 2003), can be used as a reference frame to interpret the credibility of subjective measures (Fleming & Lau, 2014). If participants report a conscious experience of a stimulus and their subjective reports are correlated with

discrimination performance about the stimulus, it is widely agreed that observers are conscious of the stimulus (e.g. Vandenbroucke et al., 2014). A second possibility is to evaluate the relationship to variations of stimulus quality (e.g. stimulus contrast, presentation time). If a subjective measure stands in no relation to the physical quality of the stimulus (e.g. its presentation time or its luminance), it appears unlikely that a representation of the stimulus was available at the rating decision. The final proposal was to consider the neural correlates of subjective measures: The credibility of subjective measures also increases when they are reliable predictors of brain activity (Charles, Van Opstal, Marti, & Dehaene, 2013). Overall, the difficulty of translating conscious experiences into a subjective measure does not rule out the usefulness of subjective measures per se; instead specific subjective measures need to be examined carefully if they are suitable measures of conscious awareness.

Argument (ii) maintains that subjective measures are not immediate measures of conscious awareness because they always involve comparison between the internal decision variable and a criterion for a subjective report (Block, 2005; Irvine, 2012). This response criterion depends on a variety of different factors such as motivation, training, and the aims of the participant, all of which are no perceptual factors (Irvine, 2012). Uncontrolled changes of the participants' criteria may result in different subjective reports between two conditions, although awareness is the same (Schmidt & Vorberg, 2006). As a consequence of the ubiquity of response criteria, Irvine argued objective measures should be used instead of subjective measures. However, the confound between internal signal and response criteria can be eliminated by the mathematical tools developed by type 2 signal detection theory (Galvin et al., 2003). Standard signal detection theory provides a rationale to distinguish between participants' sensitivity to distinguish between two response options and their criteria (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). Type 2 signal detection theory adds tools to distinguish between participants' ability to distinguish between correct and incorrect trials, a measure of metacognition, and the criteria applied in rating decisions (Fleming & Lau, 2014; Galvin et al., 2003). As a consequence, the dependency of rating decisions on criteria is no longer a decisive argument against the use subjective measures.

Critique (iii) refers the use of subjective measures in the implicit perception literature (Cheesman & Merikle, 1984). The argument asserts that subjective measures are limited because they do not exhaustively indicate all instances of conscious awareness. Critique (iii) comes in several variants: The first concern is that the statistical power of subjective measures

might be too low (Newell & Shanks, 2014; Shanks & St. John, 1994). Statistical power is an important issue because of the underlying logic of the majority of implicit perception experiments: A test of explicit (= conscious) processing is contrasted with a test of implicit (= unconscious) processing. If the test of unconscious processing is significant, but the test of conscious processing is not, it is usually concluded that perception is implicit (Schmidt & Vorberg, 2006). As the probability of obtaining a significant result depends on the statistical power of the test (Dienes, 2011), if the power of the implicit test is larger than the power of the explicit test, the experimental setup is prone to misclassify effects as implicit (Shanks & St. John, 1994). However, this reasoning rests on a misunderstanding of standard statistics: Non-significant p-values should never be interpreted as evidence for the absence of an effect; such a conclusion requires a power analysis based on a theoretically specified effect size (Dienes, 2011). Typically, many researchers indeed fail to conduct or report a power analysis, which creates the risk that the design is biased for implicit processing (Vadillo, Konstantinidis, & Shanks, 2015). If researchers ensure that the power of the explicit test is adequate, there is no reason why subjective measures cannot be used to test explicit processing. Alternatively, researchers can also pursue a Bayesian approach to hypothesis testing, which provides an even more elegant solution to the power problem in subliminal perception than power analysis (Dienes, 2015).

The second flavour of critique (iii) maintains that participants may be not sufficiently motivated to reveal all conscious experiences they in fact have (Persaud et al., 2007). Indeed, subjective measures depend strongly on motivated participants; if participants carelessly or maliciously choose to conceal their conscious experiences, there are not many possibilities to control for that. A potential solution to the motivational problem may be to ask participants to wager money on the outcome of their own performance. In this case, participants maximize their earnings only if they report all conscious experience that they have (Persaud et al., 2007). Unfortunately, post-decisional wagering is biased by participants' risk aversion (Dienes & Seth, 2010) and loss aversion (Fleming & Dolan, 2010). Nevertheless, it appears unlikely that lack of motivation will distort the results of standard subjective measures if participants are naïve to the research question: When participants are unmotivated to follow the instructions of the experimenter, some of them will fail to report exhaustively all conscious experiences they have. However, it is equally probable that some of them will report conscious experiences they in fact do not have. In summary, the effects of unmotivated participants are likely to cancel out between conditions.

The final variant of critique (iii) argues that subjective measures do not exhaustively test all conscious contents relevant for the task (Shanks & St. John, 1994). For example, in visual perception, there may be so-called *fringe experiences*, experiences where the subject is aware that an event occurred, but does not experience any characteristics of the stimulus (Mangan, 2001; Ramsøy & Overgaard, 2004). If a subjective measure only requires participants to report whether they had an experience of the stimulus, fringe experiences remain undetected. However, the content-specificity of subjective measures can also be interpreted as an advantage of subjective measures, because it is possible to tailor the content of a subjective measure to the conscious content of interest to a specific research question. If both brief glimpse experiences as well as content-specific experiences are of interest to a specific research question, a scale can be designed that offers response options for both of them. If all knowledge the participant may have is of relevance to the research question, one might just ask participants to rate their confidence in being correct at the discrimination task (Dienes, 2004, 2008). Overall, varying the contents of subjective measures allows for more flexibility in inquiring different conscious contents than discrimination tasks.

1.5. Milestones for creating subjective measures of consciousness

1.5.1. The content of subjective measures

The first characteristic of subjective measures investigated in the present work is the content of subjective measures of visual consciousness. For the purpose of the present discussion, the content of subjective measures can be defined as what participants are asked to make a report about. Subjective measures always have some content by virtue of the fact it is necessary to instruct participants what they should report. Notably, subjective measures exist where the first experience on the unaware-aware-spectrum is explained to the participants as experiences without content (Ramsøy & Overgaard, 2004) . However, even these first vague experiences do have content; possibly the experience is more of an intuition or a feeling that some stimulus is or had been present, without a visual experience of the separate features of the stimulus (Mangan, 2001).

As highlighted by Fig. 1-2, there are always at least two potential contents of subjective measures in a standard experimental situation: the stimulus and the discrimination response (cf. Chapter 2 and 3, Zehetleitner & Rausch, 2013). This raises the question which of the two should be the content of subjective measures of conscious awareness. Empirical

studies are excellently suited to investigate whether these two semantic contents are associated with different behaviours and neural correlations. Many theorists have considered reports of visual experience and decisional confidence to be interchangeable (Ko & Lau, 2012; Lau & Rosenthal, 2011; Seth et al., 2008), although others have argued for the opposite (Charles et al., 2013; Sahraie, Weiskrantz, & Barbur, 1998; Schlagbauer, Müller, Zehetleitner, & Geyer, 2012).

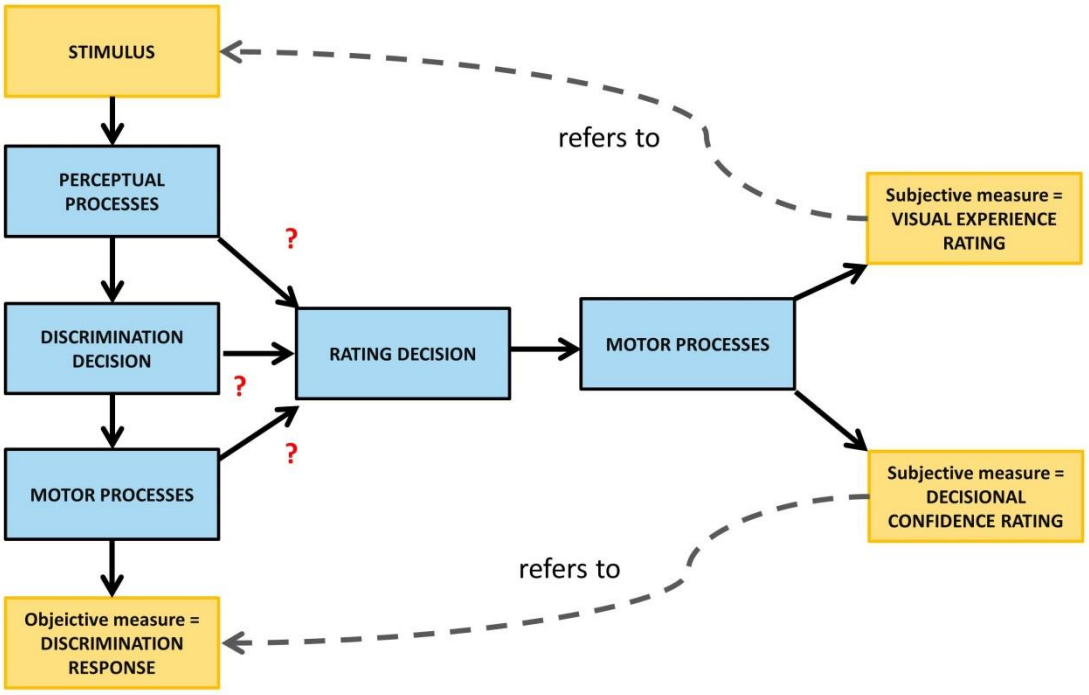

Figure 1-2. The content of subjective measures. Subjective measures may stand in a semantic relation to the discrimination response, for example at confidence ratings or post-decisional wagering. However, they may also refer semantically to the stimulus, for example at visibility or visual experience ratings.

While empirical investigations are able to show whether there are differences between subjective measures of visual experience and decisional confidence, it remains a conceptual issue what conclusions are to be drawn from the results. Many previous studies rested on the assumption that visual experience and decisional confidence are equally valid as measures of conscious awareness (Sandberg et al., 2011, 2010; Wierzchoń et al., 2012, 2014). Their rationale is that one measure should be preferred as subjective measure of conscious awareness that outperforms all the other measures in predicting trial accuracy. In contrast to this assumption, others have proposed that participants can be consciously aware of being

correct or wrong without being consciously aware of the stimulus (Carota & Calabrese, 2013; Charles et al., 2013; Sahraie et al., 1998). As a consequence, the question whether subjective measures should be about visual experience of the stimulus or about confidence in having made a correct discrimination response is again conceptual: Specific research questions may imply that either visual experience or decisional confidence is the relevant conscious content. To differentiate between these two interpretations, the present studies evaluate subjective measures not only by the correlation with discrimination performance, but also the correlation with stimulus quality as well as neural correlates. If either subjective measures that relate to the experience of the stimulus or to decisional confidence was corrupted from a great amount of unsystematic noise, it would be expected that the correlation with accuracy, stimulus quality, and brain activity should all be diminished. If the pattern of correlation varies across criteria, it would mean that subjective measures of visual experience and decisional confidence are different independently from the quality of measurement.

1.5.2. The granularity of subjective measures

The second characteristic of subjective measures investigated in the present work is the degree of granularity of subjective measures of visual consciousness: How many different scale steps should such a measure provide? The question of the preferred degree of granularity has both a theoretical as well as a methodological dimension. The reason for the theoretical importance is a prediction by global workspace theory, which asserts conscious awareness does not vary gradually, but in an all-to-none fashion (Dehaene et al., 2006; Dehaene & Naccache, 2001; Sergent, Baillet, & Dehaene, 2005). Other theories, for instance the radical plasticity thesis, assume that there are multiple degrees of conscious awareness (Cleeremans, 2008, 2011). If participants either were unconscious of the stimulus or fully conscious of the stimulus, i.e. if there were only two distinct states of conscious awareness, there would be no benefit of using scales with multiple response options. In contrast, if participants were able to consistently differentiate between more than two rating categories, it would follow that the number of degrees of conscious awareness is greater than two. The methodological aspect of investigating the granularity of subjective measures attempts to maximize the amount of information provided by subjective measures. If participants were able to differentiate between more conscious states than the number of scale steps provided by subjective measures, the consequence would be an unnecessary loss of information. However, if the number of scale steps is larger than the number of conscious states participants can

discriminate, there will be no benefit of adding more scale steps; participants might even get confused by the superfluous number of response options, resulting again in a loss of information (Overgaard, Rote, Mouridsen, & Ramsøy, 2006).

1.5.3. Quantifying type 2 sensitivity

The third and final research question addressed by the present work is to identify the preferred method to quantify the relationship between subjective measures and performance in visual discrimination tasks. Quantifying the relationship between subjective reports and discrimination performance is a very useful tool for consciousness research: First, it can be used as a reference frame to interpret subjective reports (Fleming & Lau, 2014). When a correction between subjective reports and discrimination performance is empirically observed, it shows that at least some conscious contents relevant for performance in the task are available for the rating decision. In contrast, when subjective measures and discrimination performance are unrelated, it can be concluded that the rating decision was not influenced by processes engaged in the discrimination task. Consequently, no correlation between discrimination performance and confidence was proposed as a criterion for implicit performance (Dienes, Altmann, Kwan, & Goode, 1995). Second, the association between discrimination performance and subjective reports can be used as a tool to validate subjective measures (Dienes & Seth, 2010; Sandberg et al., 2010; Wierzchoń et al., 2012, 2014): If the conscious experiences underlying participants' subjective reports are the same between two subjective measures, but the correlation with task performance is different between two subjective measures, it can be inferred that one measure is corrupted by a greater amount of noise or does not record the same amount of information than the other one (see Chapters 4 and 5; Rausch et al., 2015; Rausch & Zehetleitner, 2014). Finally, when the association between subjective measures and task accuracy is quantified by a measure of type 2 sensitivity, it can be used to counter arguments that subjective measures are inadequate to assess conscious awareness because they depend on a criterion (Irvine, 2012; see section 1.4). Type 2 sensitivity is defined as degree to which subjective reports differentiate between correct and incorrect trials irrespective of the criteria participants apply (Fleming & Lau, 2014; Galvin et al., 2003). Different measures of the association between subjective measures and task accuracy may vary in their eligibility as measure of type 2 sensitivity. Nevertheless, to fulfil this last purpose, it must be ensured that the control over criteria provided by the method to quantify the relation between subjective measures and task accuracy is effective.

Despite the theoretical importance and practical usefulness of the association between subjective reports and discrimination performance, there is no universally accepted method to quantify the relation between the two. Three competing approaches were proposed:

(i) logistic regression analysis, with subjective measure as predictor and trial correctness as dependent variable (Sandberg, Bibby, & Overgaard, 2013; Sandberg et al., 2010)
(ii) type 2 receiver operating characteristics (Fleming, Weil, Nagy, Dolan, & Rees, 2010)
(iii) meta-d_a (Maniscalco & Lau, 2012)

While (ii) and (iii) are measures derived from type 2 signal detection theory and thus were developed to account for criteria, logistic regression analysis is a method for binary dependent variables without special intent to control for criteria. However, up to now, it has never been investigated whether logistic regression is a suitable measure of the relation between subjective measures and discrimination performance. If the results obtained by logistic regression and signal detection theory based measures do not converge, the question arises what is the relationship between the results obtained by the two.

2. STIMULUS-RELATED VS. RESPONSE-RELATED SUBJECTIVE MEASURES[1]

by Michael Zehetleitner and Manuel Rausch[2]

2.1. 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 either reported 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 about the accuracy of the discrimination response, that response-related ratings outperformed stimulus-related ratings in predicting trial accuracy, and different response-related scales were more strongly associated with other response-related scales than with stimulus-related ratings. 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 the accuracy of discrimination response and discuss the relation to type 2 blindsight.

2.2. Introduction

The quest for neural correlates of consciousness relies typically on a comparison between two different types of measurements: those of neuronal processes on and those of consciousness (Crick & Koch, 1990; Rees et al., 2002). This approach critically relies on

[1] A version of this Chapter has been published as Zehetleitner, M., & Rausch, M. (2013). Being confident without seeing: What subjective measures of visual consciousness are about. *Attention, Perception, & Psychophysics, 75*, 1406–1426. doi: 10.3758/s13414-013-0505-2. Reproduced with permission by Springer.

[2] Michael Zehetleitner and Manuel Rausch share first authorship. Michael Zehetleitner conceived the research questions, Michael Zehetleitner and Manuel Rausch conceived the experiments, Manuel Rausch collected and analysed the data, Michael Zehetleitner and Manuel Rausch co-wrote the manuscript.

defining measures of consciousness, which presents a huge obstacle in empirical science (Chalmers, 1998). With respect to measures of consciousness, currently several operationalizations are proposed in the literature.

2.2.1. Objective vs. subjective measures

One prominent view distinguishes between objective measures and subjective measures (Seth et al., 2008). Measures of consciousness are considered objective if the participant's state of awareness is determined based on their 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 (Eriksen, 1960; Schmidt & Vorberg, 2006). When a subject 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 et al., 2005). Proponents of this view often make use of signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002), assuming that observers are conscious if their sensitivity to discriminate between signal and noise is above a pre-defined 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 of scientific study in the same way as other kinds of behaviour, such as button presses.

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, 1885). 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 response, either

with the possibility that the reward is lost if the wager is incorrect (Persaud et al., 2007) or without the risk of losing the wager (Dienes & Seth, 2010). As wagering is independent of speech, it has successfully been used 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 directly to make judgements 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).

2.2.2. 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 1 and type 2. In blindsight type 1, 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 2: These patients occasionally report a feeling or knowing that something happened in their blind visual field, although they insist 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 et al., 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 posterior cortex was only transiently disrupted via

transcranial magnetic stimulation: Occipital TMS between 86 -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 judgement, indicating that TMS affected the reports of subjective experience more than the reports of subjective confidence.

2.2.3. Stimulus-related vs. response-related ratings[3]

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-related rating), and measures relating to the accuracy of the discrimination response (response-related rating). In detail, we discuss whether stimulus-related and response-related ratings:

(i) might relate to different events in terms of signal detection theory

(ii) can be interpreted as measures of different processes within the cognitive architecture

(iii) might be associated with different experiences from the first-person perspective

First, it can be argued that the stimulus-related and response-related ratings mirror a distinction in SDT between type 1 tasks and type 2 tasks (Galvin et al., 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 judgement 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

[3] In the original version of the manuscript published at Attention, Perception, and Psychophysics, the two classes were called stimulus rating and decision rating. However, the latter is more accurately be characterized as "response-related rating": When participants are unconfident because they realize that they made an error due to accidently hitting the wrong button, their confidence is probably not based on their decision (which is correct in this event), but on their incorrect response instead. Consequently, the term "response-related rating" is used throughout the thesis.

existing subjective measures suggests such a correspondence, as they semantically reference either to the stimulus or to the response: “How clearly did you experience the stimulus?”, “how confident are you that your response was correct?” Thus, it seems reasonable to assume that the events in the world the two kinds of ratings refer to are different.

Concerning the second point, it is possible to connect stimulus- and response-related ratings to different functions within the cognitive architecture. Nelson and Narens’ 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 response, it is possible that both kinds of subjective reports are created by sub-sampling 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 for a response.

Third, stimulus- and response-related 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 response-related 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 response-related ratings, and likewise, a feeling of confidence is not the primary referent of stimulus-related ratings.

It should be noted that the distinction between stimulus- and response-related ratings proposed here overlaps with but is not identical to the distinction between introspective

reports and metacognitive reports proposed by Overgaard and Sandberg (Overgaard & 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 judgements 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- and response-related ratings is symmetrical in the sense that they are both based on a metacognitive judgement: 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 judgement, they assess those processes involved in selecting one out of several task alternatives. However, both stimulus- and response-related 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, as the cognitive functions of stimulus perception and decision making are closely connected, it is to be expected that the behavioural 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 behavioural 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 stimulus- and response-related subjective reports in healthy human participants.

2.2.4. Evaluation criteria for subjective measures of consciousness

The selection of criteria to evaluate measurements of consciousness is non-trivial given the fact we cannot observe another person's consciousness from the third person

perspective (Jackson, 1982; Nagel, 1974). As the extent to which a measurement "really" captures consciousness is impossible to determine, we will only consider three objective characteristics. Assuming that stimulus- and response-related 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 et al., 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 their variance is specific to them, or is shared by the other measures.

2.2.5. Empirical differences between subjective measures

Different subjective measures of consciousness have been previously compared to 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 confidence ratings imposed a more liberal criterion for ratings in terms of accuracy than the other scales (Wierzchoń et al., 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 et al., 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 et al., 2011).

2.2.6. 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 accuracy of the response, participants were instructed to either wager imaginary money on their response, to express their confidence in being correct, or to give an attribution of choice rating whether their orientation discrimination judgement was based on a guess or on knowledge. In Exp. 2-1 and 2-2, participants performed a masked orientation discrimination task, followed by one stimulus-related rating and one response-related rating (in Exp. 2-1) and two response-related ratings (in Exp. 2-2). In Exp. 2-3 and 2-4, observers performed a masked shape discrimination task with a stimulus- and a response-related rating (in Exp. 2-3) and three different response-related scales (in Exp. 2-4). Exp. 2-5 was conducted to compare stimulus- and response-related ratings in a motion discrimination task with random dot kinematograms (RDK). We collected ratings with visual analogue rating scales (VARS), 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 VARS 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 Exp. 2-1, 2-2, 2-3, and 2-4 and the proportion of dots moving coherently in one direction in Exp. 2-5, 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 centre 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. Two ratings after each single trial of the experiment 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 (ROC) and determined sensitivity and response criterion based on the area under the curve.

If subjective measures showed a similar pattern to type 2 blindsight, we hypothesized that stimulus-related and response-related ratings would exhibit different psychometric thresholds and different levels of difficulty where the zero correlation criterion is met: Response-related 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- and response-related ratings, we predicted that the association of stimulus- and response-related ratings is not as close as the association of two different response-related ratings. Third, as response-related ratings unlike stimulus-related ratings refer primarily to trial accuracy, we predict that response-related ratings exhibit a more pronounced SDT type 2 sensitivity than stimulus-related rating do. Response-related ratings should only be more efficient in predicting trial accuracy, not stimulus quality; consequently, we expect that the psychometric slope of stimulus-related ratings would be at least the same as the psychometric slopes of response-related ratings.

2.3. Experiment 2-1

Exp. 2-1 addressed the issue of comparing stimulus-related ratings against the three different response-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-related rating, and a response-related rating. There were three different response-related scales: Observers were either asked to wager imaginary money on the correctness of their discrimination response, to attribute whether the discrimination decision was rather based on a guess or on knowledge, or to give a confidence rating.

2.3.1. Methods

2.3.1.1. Participants

20 participants (2 male, 2 left-handed) participated in the experiment. The age of the participants ranged between 19 and 29, with a median of 23. All reported to have normal or corrected-to-normal vision, confirmed that that they did not suffer from epilepsy or seizures, and gave written-informed consent. The experiment has been 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 München. Participants received either €8 per hour or course credits in return for participation.

2.3.1.2. 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 24 inch 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, USA) 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/m², minimal luminance: 9.5 cd/m^2), presented in front of a grey (12.5 cd/m²) background. Squares subtended 3° x 3° degrees of visual angle. The mask consisted of a rectangular box (4° side length) with a black (1.3 cd/m²) and white (85.0 cd/m²) chequered pattern consisting 6 x 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 labelled boundaries underneath, all coloured black (1.3 cd/m^2). An index box was always initially located at the scale centre. Participants used a Cyborg V1 joystick (Cyborg Gaming, UK) to move the index along the scale and to select a location on the scale. The question of the stimulus-related rating was always “how clearly did you see the grating?” with the anchors “unclear” and “clear”. The three different response-related 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”.

2.3.1.3. Trial structure

Each trial began with the presentation of a fixation cross at screen centre 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 stimulus-onset-asynchronies 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 afterwards, the first question appeared on the screen. Participants were always asked to deliver both a stimulus-related rating and a response-related rating after each single trial. The scale type of the response-related 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-related rating or a response-related rating was asked first changed between sessions. When participants had given the first rating, they had to move the index back to the scale centre, 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 then, before the next trial started (see Fig. 2-1).

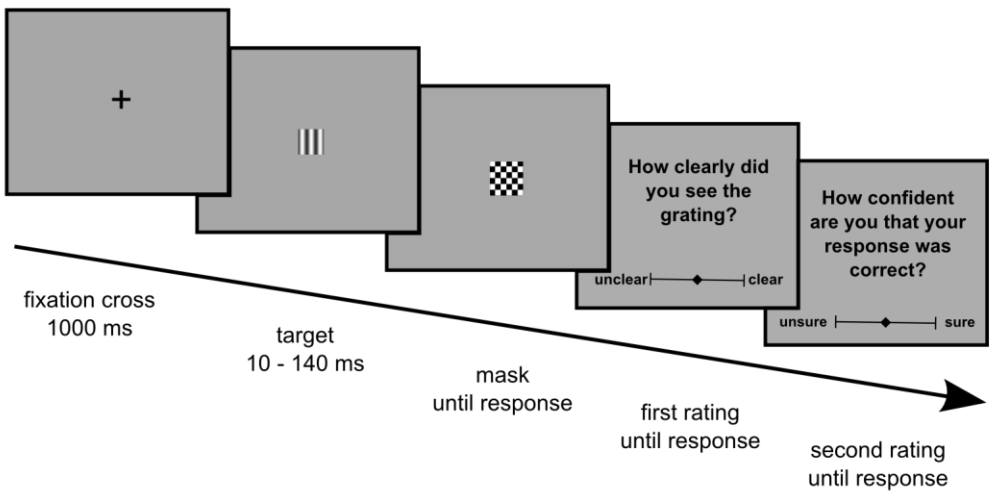

Figure 2-1. The trial sequence in Exp. 2-1, 2-2, 2-3, and 2-4.

2.3.1.4. Design and procedure

The experiment comprised 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-related 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 response-related ratings, participants were told that the ratings referred to their previous orientation discrimination response. 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 9 blocks with 40 trials each and took on average 45 minutes.

2.3.1.5. Analysis

All analysis 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-related 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).

2.3.1.5.1. Psychometric functions

To assess the relationship between stimulus- and response-related 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, 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 parameters sets of stimulus- and response-related ratings were compared by two-tailed paired t-tests.

2.3.1.5.2. SDT type 2 analysis

ROCs were constructed separately for each individual and for stimulus- and response-related ratings: For this reason, the rating data of each individual was 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 based on formulae provided by (Fleming et al., 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 the mean rating over all observers.

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

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

2.3.2. Results

2.3.2.1. Timing effects

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

2.3.2.2. Descriptive statistics

The mean error frequency in the discrimination task was .17 (SD = .08), and ranged between .41 for the shortest SOA and .01 for the longest SOA. Across the complete experiment, stimulus-related ratings averaged 46.8% of the scale range (SD = 10.0). For the response-related 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.

2.3.2.3. Psychometric functions

Within-subject ANOVAs revealed that there were no significant differences between the three response-related scales in terms of threshold, $F(2,38) = 1.2$, n. s., and slope, $F < 1$. Therefore, the rating data was pooled across different response-related scales. An estimation of psychometric functions on stimulus-related ratings aggregated across participants revealed a threshold of 4.05 (SE = .09), a slope of 2.81 (SE = .64), a lower asymptote of .10 (SE = .3), and an upper asymptote of .10 (SE = .07). For response-related ratings, the threshold was 3.93 (SE = .06), the slope 2.84 (SE = .54), the lower asymptote .10 (SE = .03), and the upper asymptote .03 (SE = .05, see Fig. 2-2).

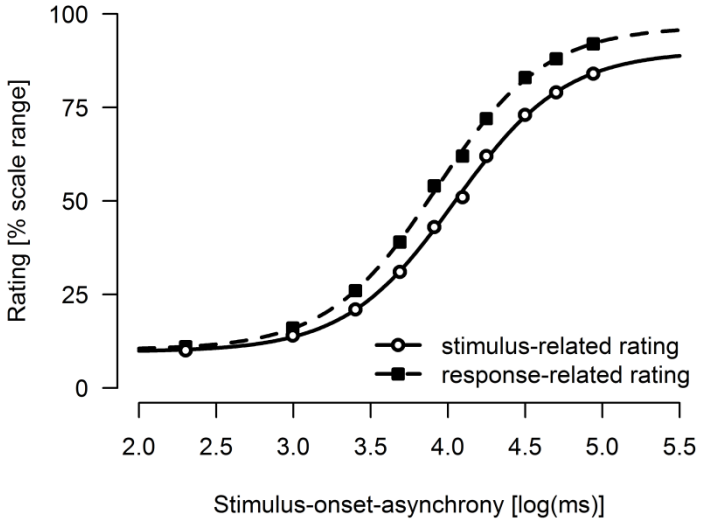

Figure 2-2. . Estimated logistic functions for stimulus-related ratings and response-related ratings. Points indicate the averaged ratings for each SOA, the solid line indicates the estimated psychometric function for stimulus-related ratings, and the dashed line the estimated psychometric function for response-related ratings.

Paired t-tests on coefficients estimated on the level of each individual revealed that the threshold for response-related ratings was lower than the threshold for stimulus-related ratings, $t(19) = 2.2$, $p < .05$, $d = .45$, and the upper asymptote was higher for response-related

ratings than for stimulus-related ratings, t(19) = 2.6, p < .05, d = .61. However, there were no significant differences in terms of slope, t(19) = 1.6, n. s., as well as lower asymptote t(19) = .8, n. s.

2.3.2.4. SDT type 2 analysis

The data was again pooled across different response-related scales as a within-subject ANOVA suggested there was no significant difference between the three response-related ratings in terms of A_{roc}, F(2,38) < 1, and B_{roc}, F(2,38) = 3.1, n. s. Fig. 2-3 displays the ROC-curves for stimulus- and response-related ratings for the whole sample.

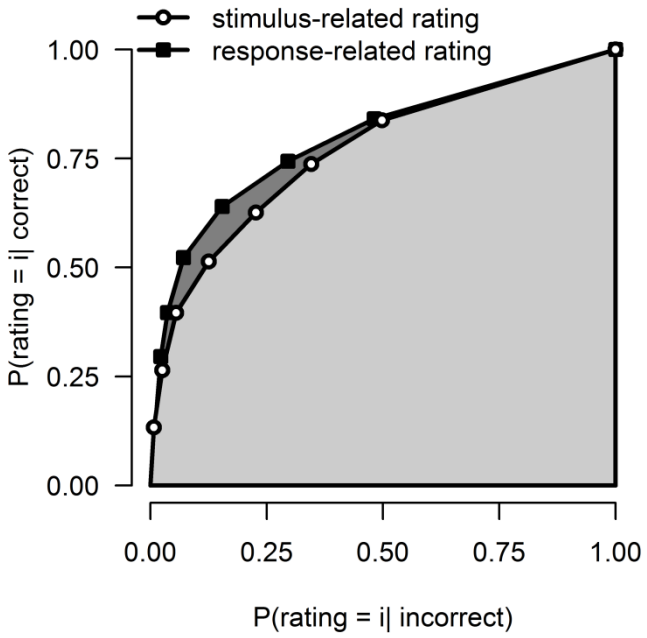

Figure 2-3. Receiver-operating-characteristics. On the x-Axis, there is the cumulative probability of each rating bin given that the trial was incorrect; on the y-axis, there is the cumulative probability for each rating given that the rating was correct. The area under the curve is used to determine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

The mean type-2- sensitivity as quantified by A_{roc} was .79 for response-related ratings (SD = .07) and .78 for stimulus-related ratings (SD = .07). Paired t-tests revealed that the difference A_{roc} between stimulus-related ratings and response-related ratings was significant, t(18) = 2.4, p < .05, d = .20. The mean type-2 criterion (B_{roc}) was -.15 (SD = .73) for response-related ratings and -.74 (SD = .82) for stimulus-related ratings. The difference between stimulus- and response-related ratings in terms of B_{roc} was significant as well, t(18) = 4.2, p < .001, d = 1.03.

2.3.2.5. Zero correlation criterion analysis

Multiple paired t-tests suggested that response-related ratings in correct trials were always greater than response-related ratings in incorrect trials at each single SOA, all p_{cor}'s < .05. By contrast, stimulus-related ratings were not significantly greater in correct trials than in incorrect trials at the shortest SOA, t(19) = .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 as indexed by Cohen's d were greater for response-related ratings than for stimulus-related ratings (see Table 2-1).

Table 2-1

T-*tests comparing ratings in correct vs. incorrect trials, separately for stimulus-related and response-related ratings and each SOA.*

		Stimulus-related ratings			Response-related ratings		
SOA	df	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

2.3.2.6. Within-trial regression

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

2.3.3. Discussion

Exp. 2-1 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 discrimination response. 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-related and response-related ratings with respect to their psychometric functions, the zero correlation criterion at different SOAs, and SDT-type 2 characteristics. It was observed that response-related ratings were associated with a lower psychometric threshold than stimulus-related ratings. We did not observe a substantial difference in the psychometric slope of stimulus and response-related ratings, indicating that both types of ratings had comparable relative sensitivities to changes of the quality of stimulation. Concerning the analysis of zero correlation criterion, response-related ratings were greater in correct trials than in incorrect trials for each SOA; while for stimulus-related 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 response-related ratings than for stimulus-related ratings at 9 out of 10 SOAs. Regarding SDT type 2 measures, response-related ratings significantly outperformed stimulus-related 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 certain stimulus conditions these patients express a high degree of confidence in their responses, although they report no visual experience (Persaud et al., 2011; Sahraie et al., 1998). In line with this, observers in the current experiment also exhibit higher thresholds towards reporting visual experience than reporting confidence. These data seem to suggest that a weaker level of stimulation is needed to elicit confidence in the response 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 or 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 a way they felt the two ratings had to be different. However, if this was the case, the bias would affect both stimulus-related and response-related ratings to the same extent and cannot account for the threshold offset between stimulus-related and response-related ratings or for the difference in SDT type 2 sensitivity.

2.4. Experiment 2-2

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

2.4.1. Methods

2.4.1.1. Participants

20 participants (6 male, 1 left-handed) participated in the Exp. 2-2. The age of the participants ranged between 20 and 40, with a median age of 27. All participants reported to have 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.

2.4.1.2. Apparatus, stimuli, design, and procedure

Apparatus, stimuli, design and procedure were identical to Exp. 2-1.

2.4.1.3. Trial sequence

The trials were identical to Exp. 2-1, except that instead of asking one stimulus-related rating and one response-related rating after each trial, there were always two out of the three possible response-related ratings. Each combination of ratings was presented for three blocks. The sequence of ratings was randomized and was opposite for the consecutive session.

2.4.1.4. Analysis

To ensure comparability between Exp. 2-1 and 2-2, the same analysis was performed for Exp. 2-2 than for Exp. 2-1, except that instead of comparing stimulus-related ratings

against response-related ratings, the three different response-related scales as confidence, attribution-of-choice, and wagering were compared against each other by an analysis of variance with scale type (confidence vs. wagering vs. attribution) as within-subject 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 his/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/her response criterion for all three scales was extremely conservative, so the B_{roc}-value could not be computed.

2.4.2. Results

2.4.2.1. Descriptive statistics

On average, the proportion of erroneous trials in Exp. 2-2 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).

2.4.2.2. Psychometric functions

Fig. 2-4 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-subject ANOVAs 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 asymptote, $F < 1$. Post-hoc t-tests revealed that the threshold for wagering was above the threshold for attribution of choice, $t(18) = 2.7$, $p < .05$, $d = .35$, and for confidence, $t(18) = 3.6 < .01$, $d = .22$, 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, $t(18) = 3.0$, $p < .05$, $d = .41$, 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.

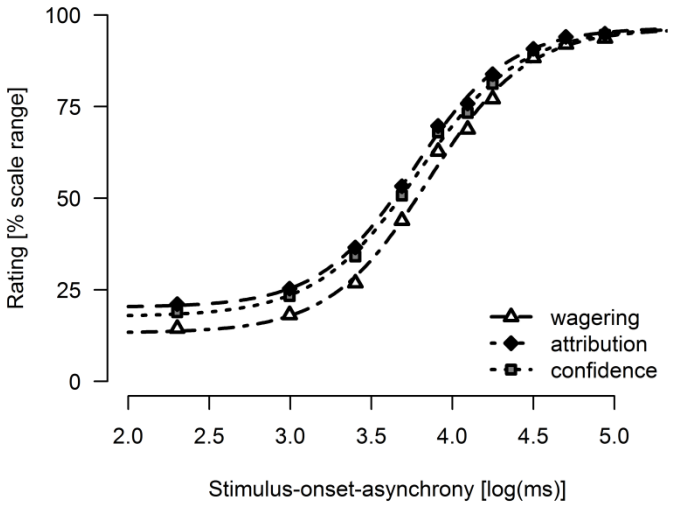

Figure 2-4. Estimated functions for confidence ratings, attribution of choice ratings, and wagering. Squares indicate mean confidence ratings for each SOA, diamonds indicate attribution of choice ratings, and triangles indicate wagering. Separate lines indicate the estimated psychometric curves.

2.4.2.3. SDT type 2 analysis

Fig. 2-5 displays the ROC-curves for the three different scales averaged across participants.

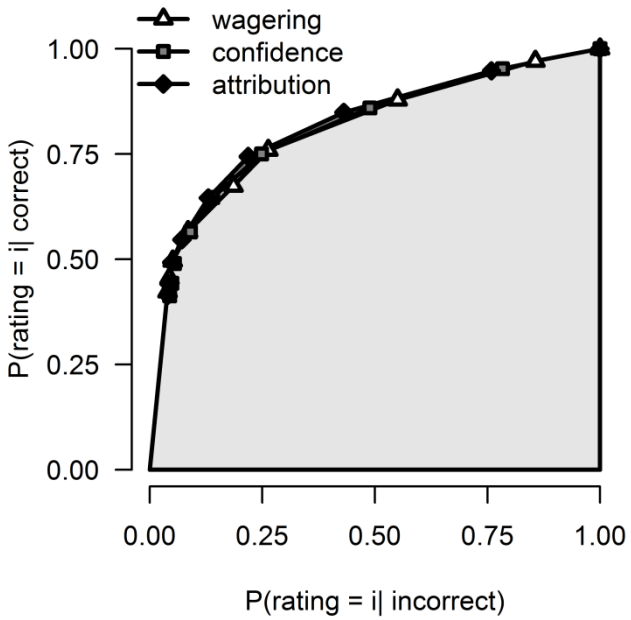

Figure 2-5. Receiver-operating characteristics in Exp. 2-2. The area under each curve indicates SDT type 2 sensitivity. Squares indicate confidence ratings, diamonds indicate attribution of choice ratings, and triangles indicate wagering.

The mean type-2- sensitivity as quantified by A_{roc} was .79 for confidence (SD = .07) and .80 for the 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 = .62) for confidence ratings, -1.05 (SD = .39) for attribution of choice ratings, and -.92 (SD = .55) for wagering. There was no significant effect of scale type on B_{roc}, $F(2,34) = 1.3$, n. s.

2.4.2.4. Zero correlation criterion analysis

As Table 2-2 shows, trial correctness predicted ratings in all three scale types starting with a 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, $t(19) = .6$, n. s., as well as confidence ratings, $t(19) = .7$, n. s. Effect sizes varied inconsistently between the different scales at different SOAs (see Table 2-2).

Table 2-2

Results of multiple t-tests comparing ratings in correct and incorrect trials in Exp. 2-2, separately for each different scale.

		Attribution of choice			Wagering			Confidence		
SOA	df	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

2.4.2.5. 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, $t(4770) = 89.4$, $p < .001$.

2.4.3. Discussion

Exp. 2-2 was conducted in order to investigate the relationship between three response-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 neither found any significant differences regarding SDT type 2 sensitivity, nor 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 response-related ratings in Exp. 2-2 seemed to be stronger than the association of response-related ratings with stimulus-related ratings as observed in Exp. 2-1.

To summarize, Exp. 2-2 revealed a considerable amount of similar empirical properties of confidence ratings, 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 for the other two ratings. A potential explanation for this finding is that wagering is not only a measure of the cognitive processes involved in the discrimination task, but might also 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 response-related ratings after Exp. 2-4.

2.5. Experiment 2-3

Exp. 2-3 investigated whether the differences between stimulus-related and response-related ratings as observed in Exp. 2-1 generalize to a masked object discrimination task. After each trial, observers indicated how clearly they experienced the shape of the stimulus (instead of the orientation of its grating as in Exp. 2-1) and how confident they felt about the accuracy of their discrimination choice.

2.5.1. Methods

2.5.1.1. Participants

16 participants (2 male, 1 left-handed) participated in the Exp. 2-3. The age of the participants ranged between 19 and 26, with a median of 22. All participants reported to have 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.

2.5.1.2. Apparatus and stimuli

The apparatus was the same as in Exp. 2-1 and 2-2, 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/m², minimal luminance: 9.5 cd/m^2), presented in front of a grey (12.5 cd/m²) background. Squares and circles subtended 3° x 3° degrees of visual angle. Mask and rating scales were identical to Exp. 2-1.

2.5.1.3. Trial structure

The trial structure was the same as 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 two-alternative forced-choice judgement 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?”, the anchors being “unsure” and “sure”. Answers were collected via VARS. If the shape judgement had been wrong, the trial ended with “error” displayed on the screen for 1,000 ms.

2.5.1.4. Design and procedure

Exp. 2-3 involved one session of approximately 1 hour. 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-related rating referred to their experience of the shape, and the response-related 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.

2.5.1.5. Analysis

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

2.5.2. Results

2.5.2.1. Descriptive statistics

The mean error frequency in Exp. 2-3 was .23 (SD = .05). On average, observers gave a stimulus-related rating of 41.1% (SD = 12.9) and a response-related rating of 52.2% (SD = 14.0).

2.5.2.2. Psychometric functions

Paired t-tests performed on individual parameters suggested that the response-related ratings were associated with lower thresholds than stimulus-related ratings, $t(14) = 2.0$, p(one-tailed) < .05, $d = .42$ (see Fig. 2-6a). In addition, we observed a marginal difference of lower asymptotes, $t(14) = 2.1$, $p = .06$, $d = .52$, but no difference between slopes, $t(14) = 1.5$, n. s., or upper asymptotes, $t(14) = 0.8$, n. s.

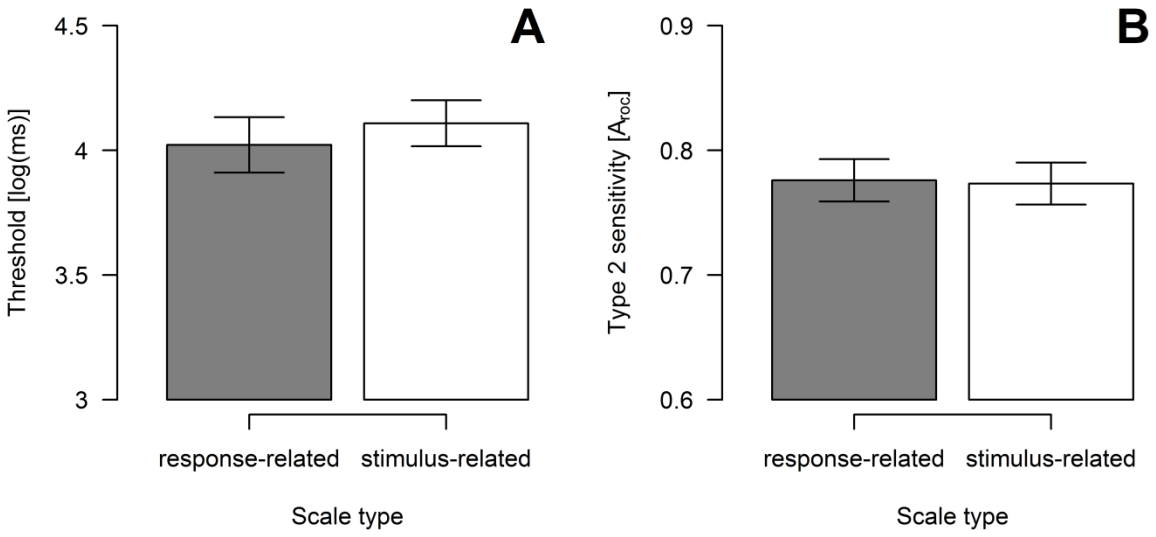

Figure 2-6. Results of Exp. 2-3. Panel A: Mean thresholds derived from stimulus-related ratings and response-related ratings. Panel B: Type 2 sensitivity of stimulus-related ratings and response-related ratings

2.5.2.3. SDT type 2 analysis

Analysis of the SDT type 2 sensitivity resulted in mean A_{roc} of .77 (SD = .08) for stimulus-related ratings and mean A_{roc} of .78 (SD = .08) for response-related ratings. For the response criterion, B_{roc} was -.93 (SD = 1.19) for stimulus-related ratings and -.39 (SD = .78) for response-related ratings. Paired t-tests suggested that there was no significant difference between A_{roc}, t(15) = .9, n. s. (see Fig. 2-6b), but the response criterion of response-related ratings was more liberal, t(15) = 2.6, p < .05, d = .61.

2.5.2.4. Zero correlation criterion analysis

Multiple t-tests suggested that both stimulus-related and response-related ratings were greater in correct trials than in incorrect trials at SOAs of 25.0 ms or greater. At shorter SOAs, no significant effects were observed (see Table 2-3).

Table 2-3

Multiple t-tests comparing ratings in correct and incorrect trials in Exp. 2-*3, separately for each different scale*

SOA	Stimulus-related ratings				Response-related 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

2.5.2.5. Within-trial regression

The hierarchical linear regressions revealed that response-related ratings could be efficiently predicted by stimulus-related ratings. The regression coefficients was .79, SE = .01, t(7400) = 104.7, p < .001.

2.5.3. Discussion

Exp. 2-3 investigated whether a pattern of subjective reports similar to type II blindsight, i.e. 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-related ratings and response-related ratings showed different characteristics in terms of psychometric functions, SDT type 2 measures, and shared variance within trials.

Regarding psychometric functions, we observed that the threshold of response-related ratings was significantly higher than the threshold of stimulus-related 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 response-related ratings was more liberal, but there was no reliable difference in sensitivity. In contrast to our prediction, while response-related ratings were associated with higher effect sizes than stimulus–related 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 behaviour of normal participants, observers in Exp. 2-3 had a higher threshold for response-related ratings than for stimulus-related ratings, meaning they would report confidence in being correct about the discrimination task already at a level of stimulation where their reports of visual experience was 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 response is consistent across tasks.

Concerning the classification of subjective measures of consciousness into two classes, the results of Exp. 2-3 are more divergent than those of Exp. 2-1. We observed differences between stimulus-related and response-related 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-related and response-related ratings is relevant: If the focus is on the correlation between subjective reports and objective performance (e.g. on zero correlation criteria), stimulus- and response-related 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 response-related ratings might lead to opposite conclusions.

2.6. Experiment 2-4

Exp. 2-4 was conducted to explore whether the lag of psychometric curves between wagering and the other response-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 response, whether they guessed or knew their discrimination response, or how much money they would place as wager that their response was correct.

2.6.1. Methods

2.6.1.1. Participants

16 participants (6 male, 1 left-handed) participated in the Exp. 2-4. The age of the participants ranged between 20 and 40, with a median age of 25. All participants reported to have 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.

2.6.1.2. Apparatus and stimuli

The apparatus and stimuli were the same as in Exp. 2-3, expect that the refresh rate was set to 160 Hz.

2.6.1.3. 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 were used. After onset of the mask and a delay period of 600 ms, participants gave a two-alternative forced-choice judgement 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 response-related scales were presented on the screen.

2.6.1.4. Design, procedure, and analysis

Design, procedure, and analysis were the same as in Exp. 2-2.

2.6.2. Results

2.6.2.1. Descriptive statistics

The mean error frequency in Exp. 2-4 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 wager of 49.1% (SD = 15.7).

2.6.2.2. Psychometric functions

Fig. 2-7a displays mean psychometric thresholds of each scale in Exp. 2-4. A comparison of the estimated parameters via a within-subject ANOVAs revealed no effects of scale type on thresholds, slopes, upper asymptotes, or lower asymptotes, all F's < 1.

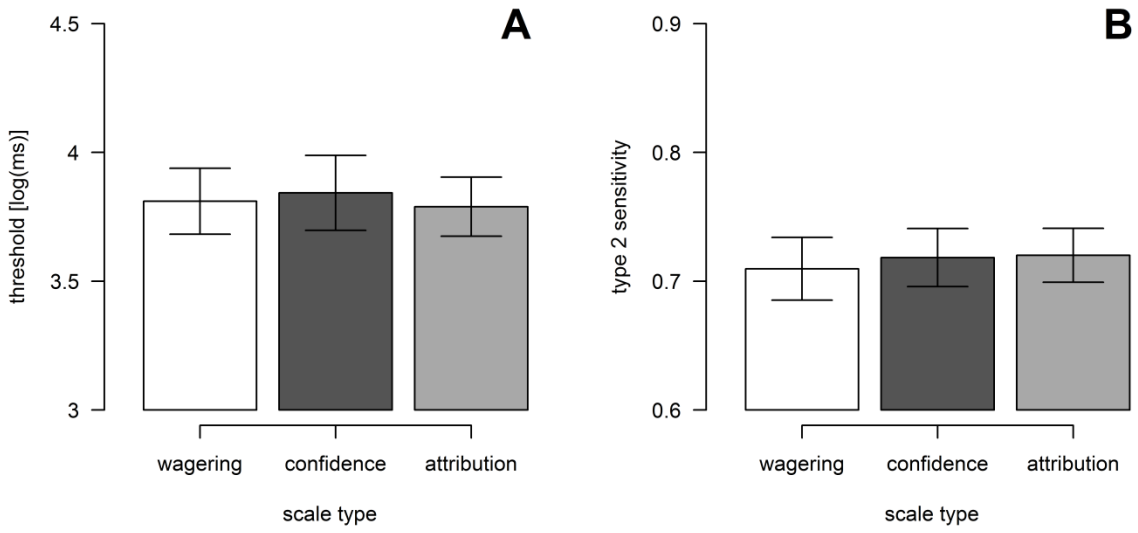

Figure 2-7. . Results of Exp. 2-4. Panel A: Thresholds for confidence ratings, attribution of choice ratings, and wagering. Panel B: SDT type 2 sensitivities.

2.6.2.3. 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 .22 (SD = 2.46) for confidence ratings, -.17 (SD = 1.81) for attribution of choice ratings, and .05 (SD = 1.54) for wagering. There was no significant effect of scale type on B_{roc}, F < 1, see Fig. 2-7b.

2.6.2.4. Zero correlation criterion analysis

As shown by Table 2-4, ratings in correct trials were significantly larger than in incorrect trials for all three scales at the SOA of 31.25 ms, all p_{cor}'s < .05. At shorter SOAs, all t-tests were not significant.

2.6.2.5. 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.

Table 2-4

Multiple t-tests comparing ratings in correct and incorrect trials in Exp. 2*-4, separately for each different scale*

	Attribution of choice				Wagering				Confidence			
SOA	df	t	p_{cor}	d	df	t	p_{cor}	d	df	t	p_{cor}	d
6.3	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
18.8	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
31.3	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

2.6.3. Discussion

Exp. 2-4 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 Exp. 2-2 also emerged at 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 at 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 response-related ratings or stimulus-related ratings, the association between two different response-related ratings in Exp. 2-4 seemed to be stronger than the association between a stimulus-related rating and a response-related rating in Exp. 2-3.

Overall, the Exp. 2-1, 2-2, and 2-4 concurrently indicate that verbal reports that refer to the discrimination response 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 Exp. 2-2, but did not replicate in Exp. 2-4. Thus, our experiments provide converging evidence that attribution of choice ratings, confidence ratings, and wagering form one coherent category of subjective measures of consciousness.

2.7. Experiment 2-5

In Exp. 2-1, 2-2, 2-3, and 2-4, sensory evidence was always manipulated by short presentation of the stimulus in conjunction with backwards masking. In Exp. 2-5, we investigated whether the discrepancy between subjective reports about the stimulus and subjective reports about the discrimination response can be replicated when sensory evidence is varied by another manipulation, i.e. the proportion of coherently moving dots of RDKs. After indicating the direction of motion of the coherently moving dots, observers delivered both a rating of the subjective clarity of motion and of confidence in the motion discrimination response.

2.7.1. Methods

2.7.1.1. Participants

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

2.7.1.2. 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 refresh rate of 120 Hz driven by a Mac with OS X 10.7 as 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 x 2 pixels, luminance 78.5 cd/m^2) in from of a black background (0.0 cd/m^2), which appeared in a circular aperture (diameter: 5) centred 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 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 leftwards and rightwards motion by pressing the left and right arrow button on the keyboard. Subjective reports were collected in the same way as in the previous experiments. The stimulus-related rating was "How clearly did you see the coherent motion?" with the anchors "unclear" and "clear"; the response-related rating was "how confident are you that your response was correct?" with the anchors "unsure" and "sure".

2.7.1.3. Trial structure

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

2.7.1.4. Design and procedure

Exp. 2-5 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 subjective reports, it was ensured that participants understood that the stimulus-related rating referred to motion experience created by the coherently moving dots, and the response-related 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 7 blocks with 49 trials each.

2.7.1.5. Analysis

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

2.7.2. Results

2.7.2.1. Descriptive statistics

The mean error rate in Exp. 2-5 was .22 (SD = .53). On average, observers gave a confidence rating of 59.7% of the scale range (SD = 11.0), and a stimulus-related rating of 52.0% (SD = 12.6).

2.7.2.2. Psychometric functions

Two-tailed paired t-tests of the estimated parameters revealed that the offset of thresholds between stimulus-related ratings and response-related ratings was significant $t(20) = 4.0$, $p < .001$, $d = .73$ (see Fig. 2-8a); 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.

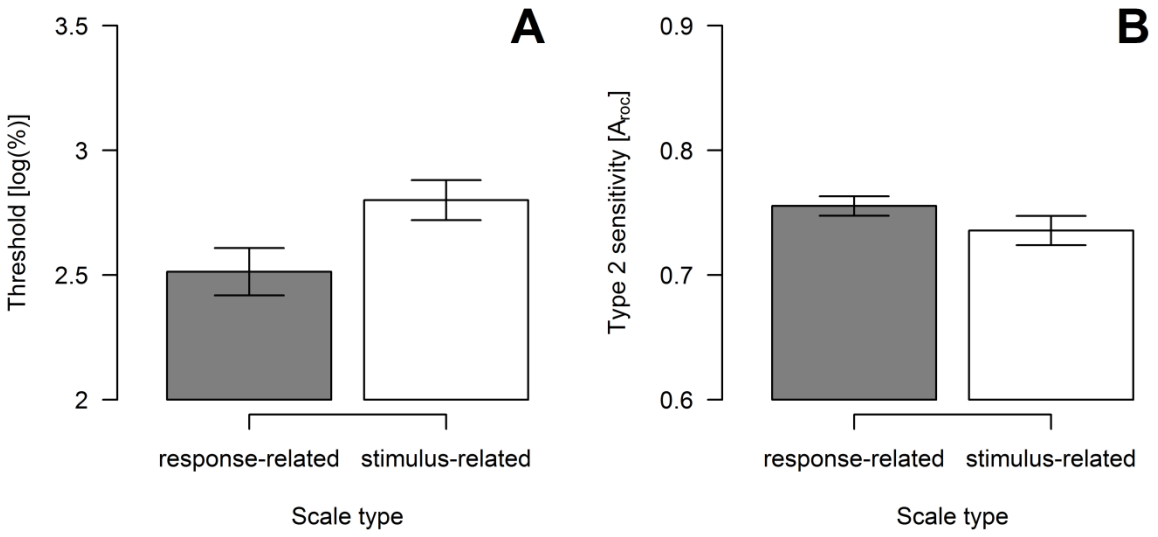

Figure 2-8. Results of Exp. 2-5. Left panel: Thresholds derived from response-related ratings and stimulus-related ratings. Right panel: Type 2 sensitivities of response-related ratings and stimulus-related ratings. Error bars indicate 1 SEM.

2.7.2.3. SDT type 2 analysis

For SDT type 2 sensitivity, the mean A_{roc} was .73 (SD = .05) for stimulus-related ratings, compared to .74 (SD = .03) for response-related ratings. Two-tailed paired t-tests suggested that the difference was significant, t(20) = 2.2, p < .05, d = .41 (see Fig. 2-8b). For the response criterion, B_{roc} was -.63 (SD= .74) for stimulus-related ratings and 0.10 (SD = 1.0) for ratings of the response. T-tests suggested that B_{roc} was different between stimulus-related and response-related ratings as well, t(20) = 5.0, p < .001, d = .82.

2.7.2.4. Zero correlation criterion analysis

Table 2-5 shows overviews t-tests performed between correct and erroneous trials at each level of coherence. Both stimulus- and response-related ratings were significantly different between correct and incorrect trials at the coherence of 2.7%. At a coherence of 1.3 %, the effect of trial correctness on response-related ratings was marginally significant, t(20) = 1.3, p = .06, d = .2, but could not be observed for stimulus-related ratings, t(20) = 0.4, n. s.

Table 2-5

Multiple t-tests comparing ratings in correct and incorrect trials in Exp. 2*-5, separately for each different scale*

Coherence	Stimulus-related ratings				Response-related 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

2.7.2.5. Within-trial regression

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

2.7.3. Discussion

Exp. 2-5 was conducted to test whether the observed discrepancy between stimulus-related and response-related 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-related ratings required a higher proportion of coherently moving dots than response-related ratings, although the relative sensitivities of both kinds of ratings were not substantially different. In addition, we found that response-related ratings outperformed stimulus-related ratings in predicting trial accuracy, and was associated with a more liberal type 2 response criterion. Concerning the zero correlation criterion, response-related ratings were marginally greater in correct trials than in incorrect trials at a coherence level of 1.3%, while stimulus-related ratings were associated with trial accuracy at a coherence of at least 2.7%. The magnitude of this effect was greater for response-related ratings than for stimulus-related ratings for 6 out of 7 levels of coherence. The association between stimulus-related and response-related ratings was comparable to Exp. 2-1 and was considerably smaller than the association between confidence, wagering, and attribution of choice ratings in Exp. 2-2 and 2-4. Overall, the results of Exp. 2-5 support nicely the distinction between stimulus- and response-related ratings, which has thus been shown for masked orientation discrimination, shape discrimination, and random dot motion discrimination.

2.8. General discussion

The five experiments presented here addressed two research questions: First, we investigated whether reports of high confidence and low visual experience, as it is reported for type 2 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 participant's discrimination response.

We compared ratings of the stimulus with ratings of the response in a masked orientation discrimination task (Exp. 2-1), a masked shape discrimination task (Exp. 2-3) and a motion discrimination task (Exp. 2-5). Concerning psychometric functions, the thresholds of response-related ratings were substantially lower than the thresholds of stimulus-related 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, response-related ratings were associated with a more liberal response criterion in all experiments and a greater sensitivity in two out of three experiments. Concerning the analysis of the zero correlation criterion, the results were more diverse: In Exp. 2-1 and 2-5, response-related 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 Exp. 2-3, we observed no differences in the zero correlation criterion analysis at short SOAs.

Confidence ratings, attribution of choice ratings, and wagering were compared during a masked discrimination task with respect to orientation (Exp. 2-2) and shape (Exp. 2-4). Regarding psychometric functions, wagering was associated with a lower threshold than the other two scales in Exp. 2-2, but no differences appeared in Exp. 2-4. 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 the patterns of the ratings are quite similar. However, beyond that similarity, response-related ratings were more efficient in predicting one of the other response-related ratings in Exp. 2-2 and 2-4 than predicting the stimulus-related ratings in Exp. 2-1, 2-3, and 2-5, suggesting there is a proportion of variance not shared between the two types of measures.

2.8.1. Type 2 blindsight in normal observers?

The current 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 responses 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 (Overgaard, Fehl, Mouridsen, Bergholt, & Cleeremans, 2008; Zeki & Ffytche, 1998). In our data, the threshold for response-related ratings was lower than for stimulus-related 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 Exp. 2-1 and 2-5, but not Exp. 2-3, response-related ratings predicted trial accuracy at a weaker level of stimulation than stimulus-related ratings did. Although the discrepancy between reported confidence and experience seems to be considerably stronger for blindsight patients, it seems as if our data shows 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.

2.8.2. Stimulus vs. response-related 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 response: The psychometric threshold for response-related ratings was lower than for stimulus-related ratings in all three experiments. With regards to SDT type 2 characteristics, response-related 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 response-related ratings in type 2 sensitivity over stimulus-related ratings because response-related ratings refer semantically to the accuracy of the trial. Moreover, wagering, confidence, and attribution of choice ratings were more strongly associated with other response-related scales within single trials than with stimulus-related ratings for both orientation discrimination in Exp. 2-1 and 2-2 and for shape discrimination in Exp. 2-3 and 2-4. Thus, consistent with our classification of subjective measures as stimulus-related ratings or response-related 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-related and response-related ratings as measurements of the strength of overlapping but not identical neural signals, although our data only supports a distinction at the level of measurements, not at the level of mechanisms. We have speculated that stimulus-related ratings might constitute a measurement of neural signals during sensory processing; while response-related 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 the response, 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-related ratings was more prone to noise than the translation into response-related ratings, it could be explained why response-related 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-related ratings. However, as noise is unsystematic, this account would predict that the correlation of stimulus-related ratings with all other events would be corrupted by noise, not only the correlation with trial accuracy. Contrary to this prediction, we observed no substantial differences between stimulus- and response-related 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 response-related ratings are only more closely related to the accuracy of discrimination responses than stimulus-related ratings, but there is no difference between stimulus- and response-related 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. It supports the view that the characteristics of subjective measures influence the events subjective measures refer to.

2.8.3. A continuum of multiple thresholds?

The discrepancy between stimulus- and response-related 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, i. e. 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 response- and low stimulus-related ratings). If the stimulation is stronger, a more stable or a different kind of representation emerges and participants report some confidence in being correct (response-related ratings increase), but

they still claim to have little experience of the stimulus (stimulus–related ratings lower than response-related ratings). Only with even greater stimulation performance, response-related ratings, and stimulus-related 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 to have visual experiences. Consequently, if a participant reports a visual experience it is very likely that s/he will also be able to discriminate the stimulus and report confidence in the discrimination response. The reverse is not the case: If a participant reports confidence in the discrimination response, there is still uncertainty whether s/he 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 to have seen all the items on display, although memory performance is restricted three to five items (Sperling, 1960). To investigate the relationship between thresholds derived from stimulus-related ratings and response-related ratings, more studies employing different paradigms and different stimulus modalities are required. Therefore, we recommend always considering stimulus-related and response-related ratings in consciousness research.

2.8.4. Relation to previous studies

The results reported here are in line with a previous artificial grammar study which 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, 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-related rating. In the study by Sandberg and colleagues, participants rated their experience on the PAS, a four-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, e. g. the 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-related ratings because it requires participants to report experiences without content as well, which could also be non-visual intuitions. However, this reasoning is entirely post-hoc; 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.

2.9. Conclusion

In summary, the present experiments indicate that participants' subjective reports when being asked to rate their perception of the stimulus vs. their discrimination response – although being similar in many ways – show reliable and important differences. Similar to type 2 blindsight patients, subjective ratings that referred to a discrimination response had lower thresholds than subjective measures which referred to the percept of the stimulus, i.e., 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, response-related ratings exhibited different SDT type 2 characteristics and different response-related scales were more strongly correlated with other response-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 response.

2.10. Acknowledgements

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 Michael Zehetleitner).

3. ELECTROPHYSIOLOGICAL CORRELATES OF CONFIDENCE AND EXPERIENCE[4]

by Manuel Rausch, Agnieszka Wykowska, and Michael Zehetleitner[5]

3.1. Abstract

The quest for the neural correlate of consciousness (NCC) is one of the biggest challenges to contemporary cognitive neuroscience. This quest is complicated by the fact that consciousness is a multidimensional construct where different dimensions imply different behavioural measurements. Our study is the first to examine the time courses of the neural correlates of two different types of subjective reports of key relevance to consciousness: reports of confidence in perceptual decisions and reports of visual experience of the stimulus. Our EEG results show that the early ERPs predicted if participants were going to report being confident in discrimination decisions, but were not yet predictive whether participants reported a clear experience over and above being confident. The strongest correlate of clear visual experiences was relatively late and only in close temporal proximity to the perceptual discrimination response. We conclude that subjective reports of visual experience and decisional confidence are associated with partially separate processes; and that research on NCC should differentiate between the different types of subjective reports.

3.2. Introduction

The quest for the neural correlates of human consciousness (NCC) is one of the most prominent and debated problems in cognitive neuroscience (Crick & Koch, 1990; Rees et al., 2002). One reason why a solution to this problem is still pending is that competing concepts

[4] unpublished manuscript submitted for publication.

[5] Manuel Rausch, Agnieszka Wykowska, and Michael Zehetleitner conceived the experiment, Manuel Rausch collected the data, Manuel Rausch and Agnieszka Wykowska performed the EEG analysis; Manuel Rausch and Michael Zehetleitner performed the statistical analysis; Manuel Rausch, Agnieszka Wykowska, and Michael Zehetleitner co-wrote the manuscript.

of consciousness imply different behavioural markers according to which researchers should ascribe conscious awareness to a participant: Some theorists have defended a concept of consciousness where conscious experience may inform decision making but is still inaccessible to verbal report (Block, 2005; Lamme, 2006); consequently, decisions in objective tasks should be used to decide whether an observer was conscious of a stimulus or not (Hannula et al., 2005; Irvine, 2012; Schmidt & Vorberg, 2006). According to a different view, participants' reports about their experience are the key phenomena for an empirical science of consciousness (Dennett, 2003, 2007), hence they are the primary raw data that needs to be recorded (Dehaene & Naccache, 2001; Dehaene, 2010). Finally, some theories proposed consciousness is associated with metacognitive processes (Carruthers, 2011; Lau & Rosenthal, 2011; Timmermans et al., 2012): according to these theories, confidence judgments are the most valid measure of conscious awareness (Dienes, 2004, 2008; Lau & Rosenthal, 2011). Given all these diverse concepts of consciousness, it seems necessary that an empirical science of consciousness assesses more than just one behavioural measure. Our study is to our knowledge the first study to compare the timing of the neural correlates of subjective reports visual experience in comparison to the neural correlates of confidence about the accuracy of task decisions.

Using more than just one single behavioural marker of conscious awareness is informative only if different markers fail to converge to the same results. Indeed, there is empirical evidence that a distinction should be made between subjective reports about visual experience and decisional confidence: In a series of psychophysical experiments, the majority of observers reported feelings of confidence in perceptual discrimination judgments at a level of stimulation where they not yet report a visual experience of the task-relevant feature of the stimulus; only when the stimulation is stronger, they would report a visual experience in addition to feeling confident in being correct (see Chapter 2, Zehetleitner & Rausch, 2013). Similarily, participants were shown to be able to detect their own errors even in absence of consicous visual experiences (Charles et al., 2013). Extreme dissociations between visual experience and decisional confidence have been reported with neuropsychological patients: After lesions to primary visual cortex, so-called blindsight patients report to be blind in the visual field contralateral to the impaired brain area, although they are able to discriminate visual stimuli presented in their seemingly blind visual field in forced-choice tasks with remarkable accuracy (Weiskrantz, 1986). Some blindsight patients report a considerable degree of confidence that judgments about a stimulus presented in their blind hemifield were

correct (Sahraie et al., 1998), and wager the same amount of money on judgments on stimuli in the blind as in the intact hemifield when performance is balanced (Persaud et al., 2011). Similarly, there is a case of an achromatic patient, who feels being colour-blind after occipital brain damage but performs well in colour discrimination tasks, and his confidence in being correct in the task strongly correlates with task performance (Carota & Calabrese, 2013). In spite of accurate discrimination performance and high levels of confidence, these patients report no experience of the task-relevant stimulus characteristics. In several of these experiments, decisional confidence was also more closely associated with task performance than visual experience (Rausch et al., 2015; Zehetleitner & Rausch, 2013), although others have reported the reverse relationship (Sandberg et al., 2010; Wierzchoń et al., 2014).

These dissociations between subjective reports of experience and confidence raise the question what is the mechanism underlying these effects. Three non-exclusive hypotheses were proposed: independent access to different sets of stimulus features (Rausch et al., 2015), distinct metacognitive processes involved in decisional confidence (Charles, King, & Dehaene, 2014; Charles et al., 2013; Overgaard & Sandberg, 2012), and placement of different sets of criteria (Wierzchoń et al., 2012). Concerning the feature hypothesis, stimuli may be represented by a hierarchy of features, of which conscious reportability varies independently (Kouider et al., 2010). While confidence may be primarily based on the stimulus feature relevant for selecting a response to the objective task (Dienes, 2008), reports of visual experience may require participants to consider other stimulus features in addition to the task-relevant one, which is why the condition to report a visual experience is less frequently fulfilled that the condition to report decisional confidence (Rausch et al., 2015). Concerning the metacognitive hypothesis, two distinct metacognitive processes involved exclusively in decisional confidence have been suggested: First, both reports of experience and confidence may depend on participants' conscious visual experiences, but decisional confidence requires an additional metacognitive process that relates performance in the current task (Overgaard & Sandberg, 2012). Second, a separate metacognitive system, which operates in parallel to conscious processing, may be involved in decisional confidence judgements (Charles et al., 2014, 2013). The final hypothesis asserts that the difference between experience and confidence can entirely be explained by participants applying different criteria to the same dimension of evidence, only reporting one's experience imposes a more conservative reporting strategy than reporting one's decisional confidence (Wierzchoń et al., 2012).

Importantly, the feature hypothesis and the metacognitive hypothesis can be disentangled by the *time courses of the neural correlates* of subjective reports regarding visual experience and decisional confidence: If experience and confidence depend on different sets of stimulus features, subjective reports of experience and confidence might be associated with different time courses of sensory neural activity. For example, the visual system is organized as a hierarchy, where neurons tuned to basic features provide input to neurons tuned to more complex features (Hochstein & Ahissar, 2002; Riesenhuber & Poggio, 1999). Consequently, if participants consider features of different complexity for experience and confidence, the neural correlates of a subjective report requiring more complex features should be later in time. In contrast, if the effect of experience vs. confidence is due to additional metacognitive processes involved in decisional confidence only, the sensory correlates of experience and confidence should be same, and specific correlates of decisional confidence should occur only *after* the features of the stimulus have been extracted.

The most convenient method to assess the time courses of neural correlates of experience and confidence is EEG – and specifically ERPs – due to their excellent temporal resolution (Luck, 2005). Importantly, the ERP correlates of visual awareness reported in previous studies can be classified depending on whether they presumably relate to sensory or post-perceptual functions: The cognitive processes associated with earlier ERP correlates are widely assumed to be sensory in nature, e.g. amplification of the signal by attentional mechanisms for an early occipital positivity around 100 ms after presentation of a stimulus (Koivisto & Revonsuo, 2010; Railo, Koivisto, & Revonsuo, 2011; Verleger, 2010), and construction of visual content (Railo et al., 2011), feature-based attention (Pitts et al., 2014), or object-based attention (Verleger, 2010) for a mid-range negative deflection recorded at posterior and temporal electrodes around 200 ms. In contrast, the hypotheses concerning a later ERP correlate, a positive deflection on centroparietal electrodes around 400 ms, all imply that visual content already exists at that point in time, e.g. broadcast of visual content within a global workspace (Del Cul et al., 2007; Lamy, Salti, & Bar-Haim, 2009; Sergent et al., 2005), update of working memory (Koivisto & Revonsuo, 2010), decision between task alternatives (Verleger, 2010), or confidence (Eimer & Mazza, 2005).

As Table 3-1 shows, only few previous studies detected a correlate of subjective reports in the time range of the early positivity, while many studies found such a correlate at a medium latency, and all studies observed an effect at the late positivity. However, as none of

these studies assessed decisional confidence in addition to visual experience, it is unclear if these ERPs are associated with both visual experience and decisional confidence, or specifically with one of them.

Table 3-1.

Previous studies reporting ERP correlates of subjective reports of visual experience and decisional confidence using identical stimuli.

Study	Paradigm	Content of subjective reports[6]	Timing of detected ERP effect		
			Early positivity	Mid-range negativity	Late positivity
Pins and Ffytche (2003)	low contrast	Experience	sig.	sig.	sig.
Koivisto and Revonsuo (2003)	change detection	Experience	n.s.	sig.	sig.
Sergent et al. (2005)	attentional blink	Experience	n.s.	sig.	sig.
Eimer and Mazza (2005)	change blindness	Confidence	n.s.	n.s.	sig.
Pourtois, De Pretto, Hauert, and Vuilleumier (2006)	change blindness	Experience	n.s.	sig.	sig.
Del Cul et al. (2007)	masking	Experience	n.s.	n.s.	sig.
Schankin and Wascher (2007)	change blindness	Experience	n.s.	sig.	sig.
Koivisto et al. (2008)	masking/ contrast	Experience	n.s.	sig.	sig.
Lamy et al. (2009)	masking	Mixture	n.s.	n.s.	sig.
Genetti, Britz, Michel, and Pegna (2010)	degraded stimuli	Mixture	sig.	sig.	sig.
Salti, Bar-Haim, and Lamy, (2012)	masking	Mixture	n.s.	n.s.	sig.

In addition, the interpretation of the absence of significant effects in early time windows in these studies is limited by the nature of significance testing: When P-values are not significant, it is not legitimate to infer the effect does not exist without appropriate power

[6] We classify the content of subjective report as "visual experience" if participants were instructed to report if they had seen the stimulus, and as "confidence" if participants reported their subjective confidence in the discrimination decision of the trial being correct. „Mixture" means that the subjective report was instructed in a way it referred to both the subjective experience of the stimulus as well as to the confidence. "Sig" indicates "significant" and means an effect in this time range was detected; "n.s." stands for "not significant" and means that it is unknown whether there is an effect or not.

analysis (Dienes, 2011). Bayesian hypothesis testing exceeds P-values in so far as it allows to quantify the evidence for both presence and absence of an effect (Dienes, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009). The present study is to our knowledge the first study that assesses both the presence and the absence of early ERP correlates.

To summarise, our study aimed at dissociating ERP correlates of experience of visual stimulus vs. ERP correlates of confidence about the discrimination decision and at investigating their time courses, in order to determine what is the plausible underlying mechanism of situations in which human observers report to be confident in their discrimination decision, but do not yet report an experience of the stimulus. We hypothesized that if visual experience depends on the extraction of additional stimulus features on top of those features relevant for decisional confidence, early and mid-range ERPs associated with decisional confidence and visual experience might follow different time courses. On the contrary, if the behavioural effects of experience and confidence as content of subjective reports were due to metacognitive processes alone, we expect identical ERP correlates of experience and confidence in the early and mid-latency time range, while decisional confidence should be more strongly associated with the late ERP positivity.

To meet the aims of our study, we recorded EEG while observers performed a 2AFC masked orientation discrimination task. Participants delivered two subjective reports after each (objective) discrimination response, one of which was about their experience of the stimulus, and the other about their confidence in their orientation discrimination response (see Fig. 3-1). To increase the number of trials available for ERP averaging, participants' subjective thresholds were determined in a screening session, and a stimulus at threshold was used for each individual participant throughout the whole EEG recording session. This procedure resulted in three different combinations of subjective reports while stimuli were physically identical: (i) participants reporting that they had a clear experience of the stimulus and simultaneously were confident about their discrimination decision, (ii) participants reporting unclear experience and nevertheless confidence about the discrimination response, and (iii) participants reporting unclear experience as well as admitting to have chosen the orientation response based on guessing. As participants reported to be confident regarding their discrimination response in (i) and (ii), but the level of reported experience was different between these trial categories, the comparison between (i) and (ii) can be used to investigate the correlates of participants reporting a clear visual experience over and above reporting to

be confident in a discrimination decision. Consequently, we refer to this comparison as *experience contrast*. Correspondingly, the comparison between (ii) and (iii) is informative of the correlates of participants reporting to be confident, and is thus referred to as *confidence contrast*. A sample to analyse combination (iv) - participants reporting a clear experience and low confidence - was not collected, because only few participants would report this combination (see Fig. 3-2 for results of the screening session).

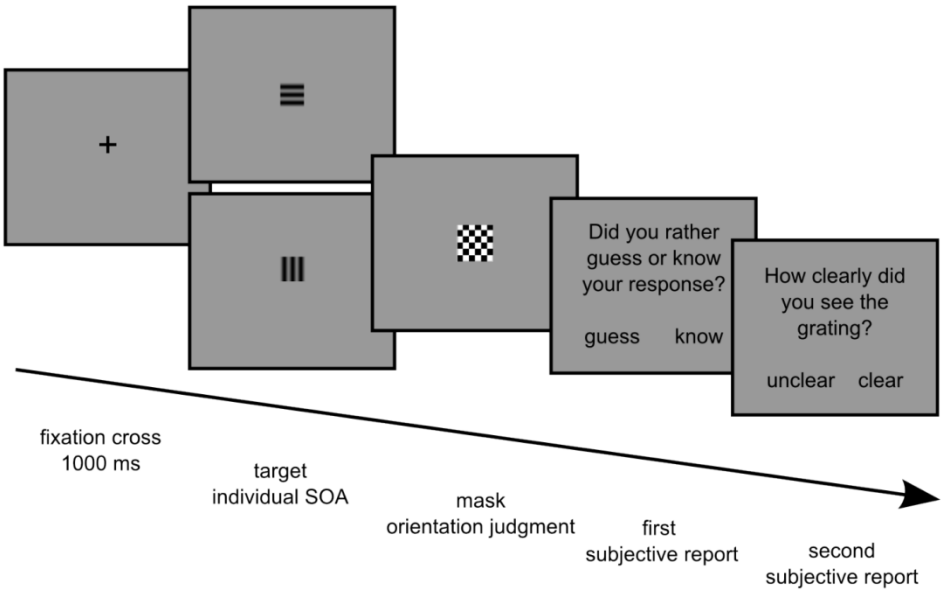

Figure 3-1. Trial structure. The order of the visual experience and confidence judgments was counterbalanced across participants.

3.3. Experiment

3.3.1. Material and Methods

3.3.1.1. Participants

20 participants took part in the EEG experiment. Mean age was 23.9 (SEM = .6) and all were right-handed. All participants reported normal or corrected-to-normal vision, no history of neuropsychological or psychiatric disorders and no psychoactive medication. Participants gave written informed consent and received either course credits or €8 per hour for participation. The experiment was conducted according to the principles expressed in the Declaration of Helsinki, and the study protocol was approved by the ethics committee of the Deutsche Gesellschaft für Psychologie.

3.3.1.2. Apparatus and stimuli

The stimuli were presented on and CRT-monitor with 17 inch screen size and a 100 Hz refresh rate, placed in a distance of approximately 75 cm in front of a participant, located in a sound-attenuated and electrically shielded cabin. The experiment was conducted using a PC with Windows XP, MATLAB and Psychtoolbox (Brainard, 1997; Pelli, 1997). The target stimulus was a square, which subtended 3° x 3° degrees of visual angle textured with either a horizontal or a vertical oriented sinusoidal grating (frequency: 1 cycle/deg, maximal luminance: 16.7 cd/m²; minimal luminance: 4.6 cd/m²), presented over a grey (16.7 cd/m²) background. The mask consisted of a rectangular box (4° side length) with a black (4.6 cd/m²) and white (85.8 cd/m²) chessboard pattern consisting 6 x 6 equal squares. One half of the participants responded to the orientation task by pressing "A" or "S" on the keyboard with their left hands and pressed either "K" or "L" for the first and "N" or "M" for the second subjective report with their right hands. The other half responded to the orientation task with their rights hands by pressing "K" and L" and pressed "A", "S", and "Y" and "X" to deliver their subjective reports.

3.3.1.3. Trial structure

As Fig. 3-1 shows, each trial began with the presentation of a fixation cross at the screen centre for 1000 ms. Subsequently, the target stimulus was presented for a period of time until it was replaced by the mask. The mask onset was timed individually for each participant based on a threshold estimated during a separate screening session (see below). Both stimulus and mask were located at fixation. The mask remained on the screen until participants indicated by button press whether the orientation of the grating had been horizontal or vertical. To prevent premature responses, participants could not respond until 600 ms after mask onset. Immediately afterwards, the first of the two questions was presented. The questions were "how clearly did you see the grating?" with the possible answers "unclear" vs. "clear", and "did you guess or know your answer?" with the possible answers "guess" and "know". Our previous study suggested that asking observers whether they attribute their own choice to guessing or to knowledge is equivalent to a confidence rating (see Chapter 2; Zehetleitner & Rausch, 2013). Subjects always responded to both questions after each trial, and which of the two questions was first was counterbalanced across subjects. If the response to the task had been erroneous, "error" was displayed on the screen for 1,000 ms after the last subjective report, before the next trial started.

3.3.1.4. Design and procedure

The experiment involved a behavioural screening session as well as an EEG recording session. For both sessions, participants were instructed to report the orientation of the grating of the target stimulus as accurately as possible and to deliver the subsequent reports as carefully as possible. Prior to the main experiment, participants were instructed to fixate at the cross at the screen centre and to avoid blinking. First, participants performed 20 trials of training, and then 12 blocks of 72 trials each. After each block, the percentage of errors was displayed to provide participants with feedback about their accuracy.

3.3.1.5. Screening session

The screening session consisted of 9 blocks with 40 trials each, with the same task as in the maim experiment, except there were six different SOAs between target stimulus and mask of 20, 30, 50, 60, 80 and 100 ms. The screening session data was analysed by estimating psychophysical functions on subjective reports separately for each participant. For this purpose, we fitted psychometric functions quantifying the relationship between SOA and the probability of reporting a clear experience and to be confident about their response respectively using the R package gnlm (Lindsey, 2010). The psychometric function was defined by the formula

$$f(x) = \frac{1}{1 + e^{\frac{-(x-\alpha)}{\beta}}}$$

where x is the logarithm of the SOA, β denotes the slope of the psychometric function, and α is its centre. The threshold was defined as the SOA where the probability of reporting a clear experience or being confident was 50%. Error trials were omitted from analysis. The results of the psychophysical function analysis are found in Fig. 3-2. As the present study was designed to investigate the ERP correlates of situations where participants report to be confident about the response but not yet report to have a clear experience of the stimulus, the EEG recording session was performed only with participants who had a higher threshold to report a clear experience of the stimulus than to be confident about the response (21 out of 29). In addition, participants were excluded if their performance did not exceed change level (1 participant) and if their reports were too conservative so thresholds could not be determined with precision because one of the thresholds fell far outside the range of SOAs presented in the experiment (2 participants). Overall, 20 out of 29 participants of the screening session met the inclusion criteria. For those participants, one psychometric function

was fitted on the combined rating data of both verbal reports, of which the threshold was used as SOA in the EEG recording session.

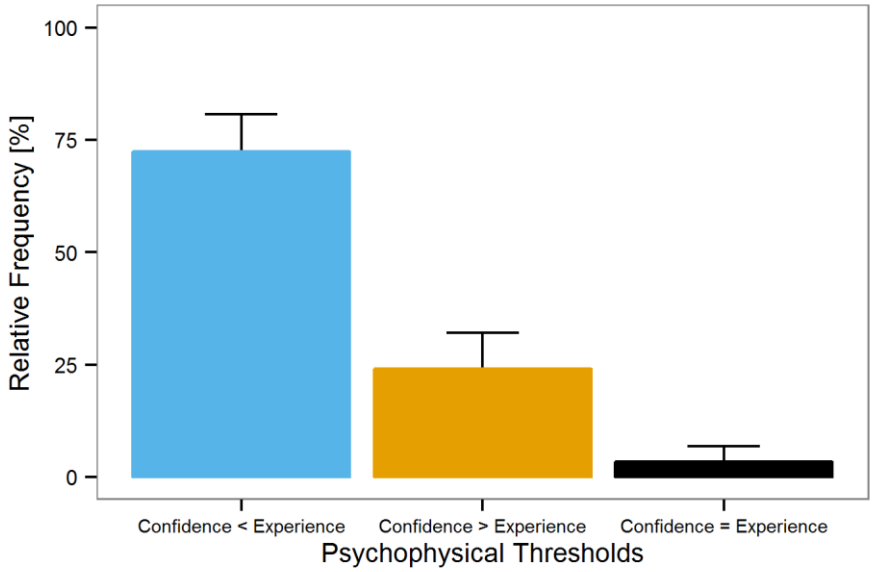

Figure 3-2. Relative frequency of participants with a higher threshold to report a clear experience than to report to be confident about the discrimination response (blue), of participants who apply lower thresholds for reports of experience than for confidence (orange), and of participants with the same thresholds for experience and for confidence (black). A Bayes factor confirmed that more participants apply more conservative thresholds for visual experience than for confidence than vice versa, BF_{10} = 8.28, posterior distribution of the probability of a lower threshold for confidence than for experience: mean = .70, 95% credible interval = [.54 .84].

3.3.1.6. EEG recordings

The EEG was recorded at a digitization rate of 500 Hz from 64 Ag/AgCl active electrodes (ActiCAP, Brain Products, GmbH, Munich, Germany), positioned according to the International 10-10-system. Horizontal eye movement were monitored by means of electrodes placed 1 cm lateral to the outer canthi of the eyes, and vertical eye movements by electrodes placed below and above the left eye. The EEG signals were amplified by BrainAmp amplifiers (BrainProducts, Munich) with a high cut-off filter at 250 Hz and low cut-off filter at 0.1 Hz. All electrodes were referenced to Cz and re-referenced offline to the averaged activity across all electrodes. Electrode impedances were kept below 5 kΩ.

3.3.1.7. Analysis

All data sets, analysis scripts, and supplementary results are available for download at the Open Science Framework to facilitate reproduction of the present study and replication of

its results (Ince, Hatton, & Graham-Cumming, 2012; Morin et al., 2012; Simonsohn, 2013)[7]. EEG filtering, epoching, artefact rejection, and ERP averaging were performed using Brain Vision Analyzer software 1.05 (Brain Products GmBH, Munich); all other analyses were performed in R (R Core Team, 2014).

3.3.1.7.1. EEG analysis

First, the data was filtered off-line with a 40-Hz high cut-off filter (24 dB/Oct). Second, the EEG was epoched into segments either locked to the onset of the stimulus or to the discrimination response. The stimulus-locked segments lasted from 200 ms before stimulus onset until 600 ms poststimulus, with the interval between 200 ms pre-stimulus interval used for baseline correction. The response-locked segments were from 1300 ms prior to response until 50 ms after response, with the first 200 ms again used as baseline. Eye movement and blink detection was performed on F9, F10, and vEOG electrodes: Segments with an absolute voltage difference exceeding 80 μV or a voltage step between two sampling points exceeding 50 μV at one of these electrodes were excluded. In addition, we excluded channels with amplitudes exceeding ±80 μV, or an activity lower than 0.1 μV within intervals of 100 ms. Trials with incorrect responses were excluded. A separate ERP waveform was constructed for each of the three possible combinations of subjective reports, i.e. we compared (i) trials, in which subjects reported to have clearly seen the stimulus and were confident about discrimination response, (ii) trials, in which subjects reported their experience of the stimulus was rather unclear, but they were confident about discrimination response, and (iii) trials in which subjects reported their experience was unclear, and they performed the orientation response based on guessing. In four participants, the number of trials after artefact rejection was too low (< 5 trials) to compute stable ERPs, so these four participants were excluded from EEG analysis. To stay uncommitted about the timing of ERP correlates of subjective reports, the data was divided into a series of time windows of equal duration. For stimulus-locked ERPs, each window had a duration of 50 ms, and the first and last window began 100 ms and 300 ms after stimulus onset, respectively. For response-locked ERPs, each window had a duration of 100 ms, and the first and last window started 400 and 100 ms before the response. ERPs were quantified by the mean amplitude of each time window to avoid bias from varying signal-to-noise ratios at unequal numbers of trials between conditions (Luck, 2010). As the existing literature suggests an occipital topography of early effects, a

[7] Link to the full material:
https://osf.io/ghfwj/?view_only=19c269713cfc425da5772850bca36f91

posterior-temporal topography of mid-range effects, and a central-parietal topography of late effects (Koivisto & Revonsuo, 2010; Railo et al., 2011), the analysis was performed on mean amplitudes from electrodes PO7, PO8, O1, O2, and Oz for the time windows between 100 and 150 ms after stimulus onset, PO7, PO8, O1, O2, P7 and P8 for the time windows between 150 ms and 350 ms, and P1, P2, Pz, CP1, CP2, CPz, C1, C2, and Cz for the response-locked time windows (cf. maps in Fig. 3-4 and Fig. 3-5). To determine the timing of the mid-range negative and late positive effect, we constructed differences waves for both the experience and confidence contrast. Onset and peak latency were determined using the Jackknife-based scoring method with the 25% and 50% area criterion in the time windows 150-350 ms poststimulus and 300-0 ms before response (Kiesel, Miller, Jolicoeur, & Brisson, 2008; Ulrich & Miller, 2001).

3.3.1.7.2. Statistical analysis

As both the presence as well as the absence of effects are relevant to the current study, we base our interpretation on Bayes factors, which provide a continuous measure of how the evidence supports the alternative hypothesis over the null hypothesis and vice versa (Dienes, 2011; Rouder et al., 2009). Bayes factors were computed using the R library BayesFactor, where default priors are placed on standardized effect sizes (Morey & Rouder, 2014). 95% credible intervals, the intervals that include the true parameter with a probability of .95, were computed based on 10^6 samples from posterior distributions. Despite the merits of Bayesian statistics over P-values, analogous conventional statistics are available at open science framework to improve comparability with previous studies.

Type 2 sensitivity, the association between subjective reports and discrimination performance (Fleming & Lau, 2014; Galvin et al., 2003), was quantified by meta-d_a (Maniscalco & Lau, 2012) using a maximum likelihood procedure implemented in R (Rausch et al., 2015) and compared between experience and confidence by the Bayesian equivalent of a t-test (Rouder et al., 2009). Trials were participants made subjective reports more quickly than 200 ms after presentation of the scale were considered as premature response and were excluded from analysis.

To examine effects on ERP amplitudes by Bayes factors, we fitted Bayesian linear regression models (Rouder & Morey, 2012) for each time window with mean amplitude as dependent variable. Each model involved the experience contrast, the confidence contrast, hemisphere, site (electrodes O1, O2, Oz: occipital, PO7; PO8: parieto-occipital, P1, P2, P3,

P4, P7, P8, Pz: parietal, CP1, CP2, CPz: central) and a random effect of participant as predictors. Experience and confidence contrasts were tested by dropping each of them out of the model and comparing model without experience or confidence against the full model. Consequently, each Bayes factor reflects the evidence that a specific contrast explains variance in ERP amplitudes over and above the other contrast, hemisphere, and site. The experience contrast was coded in a way that the estimated effect can be directly interpreted as the difference between trials when observers reported to be confident and to have a clear experience, and trials when observers reported they were confident but their experience was unclear. Likewise, the confidence contrast reflects the difference between trials when observers reported they had no clear experience and they guessed the response, and trials when observers reported they had no clear experience but they felt confident about the accuracy of the discrimination decision.

For latencies of the experience and confidence contrasts, we computed Bayes factors by transforming t-values from standard statistics into Bayes factors. Analogous to the analysis of amplitudes, t-values were obtained by a linear regression model with the experience contrast, confidence contrast, hemisphere, site, and a random intercept effect of participant using the R libraries lme4 (Bates, Maechler, Bolker, & Walker, 2014) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2014). The t-values were corrected as ERP latencies were determined by the jackknife-based scoring method (Kiesel et al., 2008; Ulrich & Miller, 2001).

3.3.2. Results

3.3.2.1. Behavioural results

During the experiment, participants made on average 17.4 % errors (SEM = 3.2 %). Moreover, they reported to be confident about the discrimination response and to have a clear experience in 29.3 % of the trials (SEM = 5.1), to be confident without a clear experience in 36.2 % of trials (SEM = 4.2), and to have guessed the orientation in combination with an unclear experience in 34.2 % of all trials (SEM = 5.1). The association between subjective reports and performance quantified by meta-d_a was greater for reports of decisional confidence (M = 1.5, SEM = .2) than reports of visual experience (M = 1.1, SEM = .2, see Fig. 3-3). The Bayesian analysis indicated strong evidence for different meta-d's of experience and confidence, BF_{10} = 60.38, posterior distribution of the difference between confidence and experience: M = .39, 95% credible interval = [.18 .61].

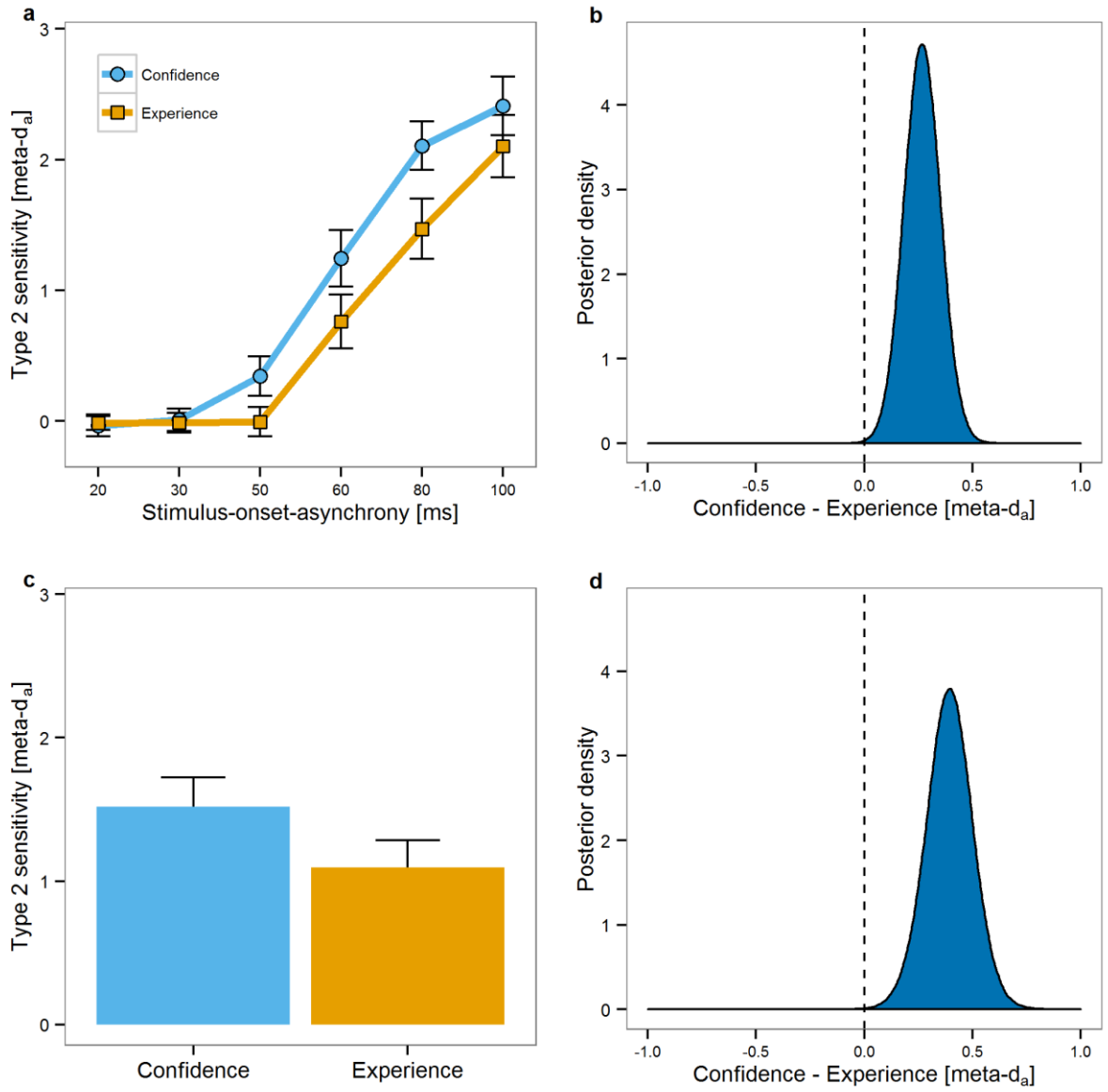

Figure 3-3. Type 2 sensitivity measured by meta-d_a depending on if subjective reports were about decisional confidence or visual experience. The greater meta-d_a, the more efficiently subjective reports differentiate between correct and incorrect task responses. (a) Mean and standard errors of meta-d_a of decisional confidence (blue) and visual experience (orange) of participants of the screening experiment, as a function of stimulus-onset asynchrony. Bayes factors revealed effects of SOA, $BF_{10} = 4.82 \cdot 10^{52}$, and experience vs. confidence, $BF_{10} = 24.81$, but no interaction, $BF_{10} = 0.02$. (b) Posterior distribution of the effect of confidence vs. experience during the screening experiment assuming a default JZS prior (Morey & Rouder, 2014). Mean of the posterior distribution: 0.27, 95% credible interval: [0.10 0.43]. Positive values indicate that Type 2 sensitivity of confidence is greater than of experience, and vice versa. (c) Meta-d_a of confidence and experience in the main experiment. (d) Posterior distribution of the effect of confidence vs. experience during the main experiment.

3.3.2.2. ERP results

3.3.2.2.1. Stimulus-locked ERPs

As Fig. 3-4 shows, both contrasts were associated with a mid-latency negative shift over posterior electrodes, although a correlate of the contrast of confidence (light grey boxes, blue vs. black line) emerged somewhat earlier than of the contrast of experience (dark boxes, orange vs. blue line). In detail, for the earliest time windows between 100 and 150 ms after the stimulus onset, the Bayes factors indicated there was evidence for the null hypothesis (i.e. there is no difference in mean amplitudes) for both the experience contrast, $BF_{10} = 0.24$, as well as the confidence contrast, $BF_{10} = 0.23$. For the time windows between 150 and 200 and between 200 and 250 ms, the Bayes factor indicated positive support for an effect of confidence, BF_{10}'s = 4.76 and 8.32, while the support for an effect of visual experience was only anecdotal and thus not conclusive, BF_{10}'s = 0.68 and 2.11. Only at the following time windows between 250 and 300 ms as well as 300 – 350 ms, we observed evidence for an effect associated with the experience contrast, BF_{10}'s = 15.04 and 4.54, and with the confidence contrast, BF_{10}'s 28.02 and 17.28. Posterior means and 95% credible intervals for all experience and confidence contrasts can be found in Table 3-2.

Table 3-2
Means and credible intervals of the posterior distributions of the experience and confidence contrasts in µV for each time window.

Time Window		Experience contrast			Confidence contrast		
		M	Credible interval		M	Credible interval	
			2.5	97.5		2.5	97.5
Stimulus-locked time windows	100 - 150	0.13	-0.37	0.63	-0.11	-0.61	0.39
	150 - 200	-0.55	-1.23	0.13	-0.87	-1.56	-0.19
	200 - 250	-0.74	-1.41	-0.08	-0.93	-1.60	-0.26
	250 - 300	-1.08	-1.81	-0.36	-1.16	-1.88	-0.44
	300 - 350	-0.81	-1.45	-0.18	-0.97	-1.60	-0.34
Response-locked time windows	-400 - -300	0.35	-0.06	0.78	0.14	-0.27	0.55
	-300 - -200	1.06	0.62	1.50	0.83	0.39	1.27
	-200 - -100	1.06	0.62	1.50	0.79	0.35	1.23
	-100 - 0	0.87	0.44	1.30	0.63	0.20	1.06

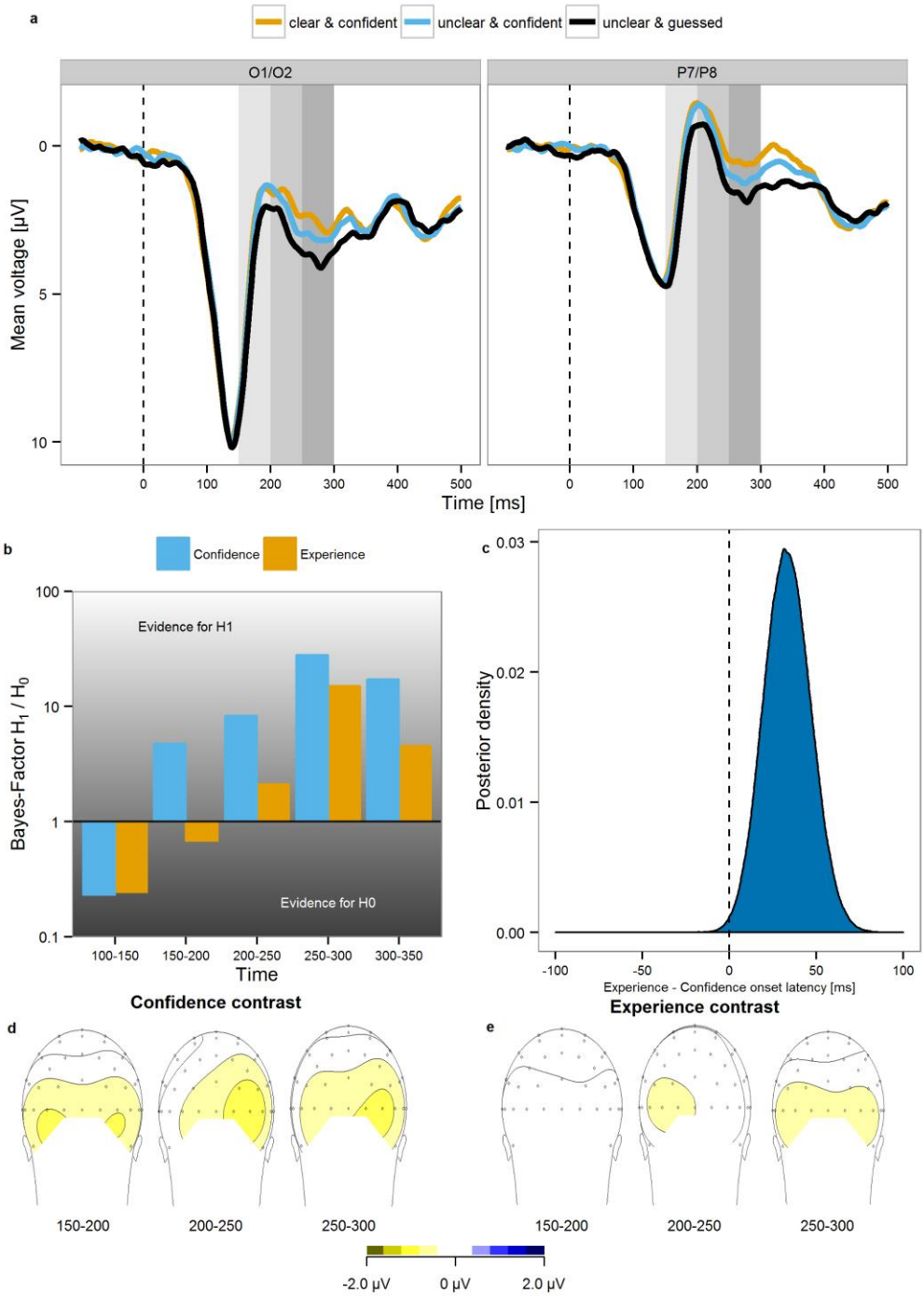

Figure 3-4. Time course of stimulus-evoked electrophysiological activity. (a) Grand average ERPs plotted as a function of time and subjective reports. Left panel: occipital electrodes O1 and O2. Right panel: temporal electrodes P7 and P8. Orange line: ERP average when participants reported clear experience and being confident. Blue line: ERP average when participants reported unclear experience despite being confident. Black line: ERP average when participants reported an unclear experience and were unconfident about the discrimination response. (b) Bayes factors of the experience contrast (orange) and confidence contrast (blue) in time windows of 50 ms each. (c) Posterior distribution of the difference in 25% area latency between the experience contrast and confidence contrast. (d) Scalp distribution of a difference wave of the confidence contrast for the time windows 150-200, 200-250, and 250-300 ms after stimulus onset. (e) corresponding voltage maps for the experience contrast.

The latency analysis revealed that the mean onset of the effect related to the confidence contrast averaged across electrodes was at 198.3 ms and the average peak was at 251.8 ms, while the mean onset at the experience contrast was not until 237.5 ms with the peak at 265.9 ms. A Bayes factor analysis confirmed that the effects related to the experience and confidence contrast differed in onset, $BF_{10} = 8.37$ (see posterior distribution of the effect in Fig. 3-4c), but not in peak, $BF_{10} = 0.33$.

3.3.2.2.2. Response-locked ERP

As Fig. 3-5 shows, response-locked central and parietal ERPs were associated with subjective reports in the time between 250 ms and 100 ms before the response, although the experience effect (medium and dark grey boxes, orange vs. blue line) seemed to be greater in magnitude and more broadly distributed than of the confidence effect (black vs. blue line). In addition, Fig. 3-5a shows an early posterior effect of confidence maximal around 850 ms before the response, which may reflect the same effect as observed at stimulus-locked ERPs (light grey box, black vs. blue line, and Fig. 3-4d). The Bayesian analysis indicated support for the null hypothesis in the time windows 400-300 before the response for the confidence contrast, $BF_{10} = 0.24$, while the evidence for the experience contrast was not more than anecdotal in favour of the null hypothesis, $BF_{10} = 0.83$. In all three following time windows from 300 to 200 ms, from 200 to 100 ms, and from 100 ms until the response, the Bayes factors indicated there were effects related to both contrasts. For the confidence contrasts, the evidence in favour of an effect was very strong at 300-200 ms, strong at 200-100, and positive at 100-0, BF_{10}'s = 177.46, 94.89, and 11.53. For the experience contrasts, the evidence was always very strong, BF_{10}'s = $1.23 \cdot 10^4$, $1.07 \cdot 10^4$, and 484.70.

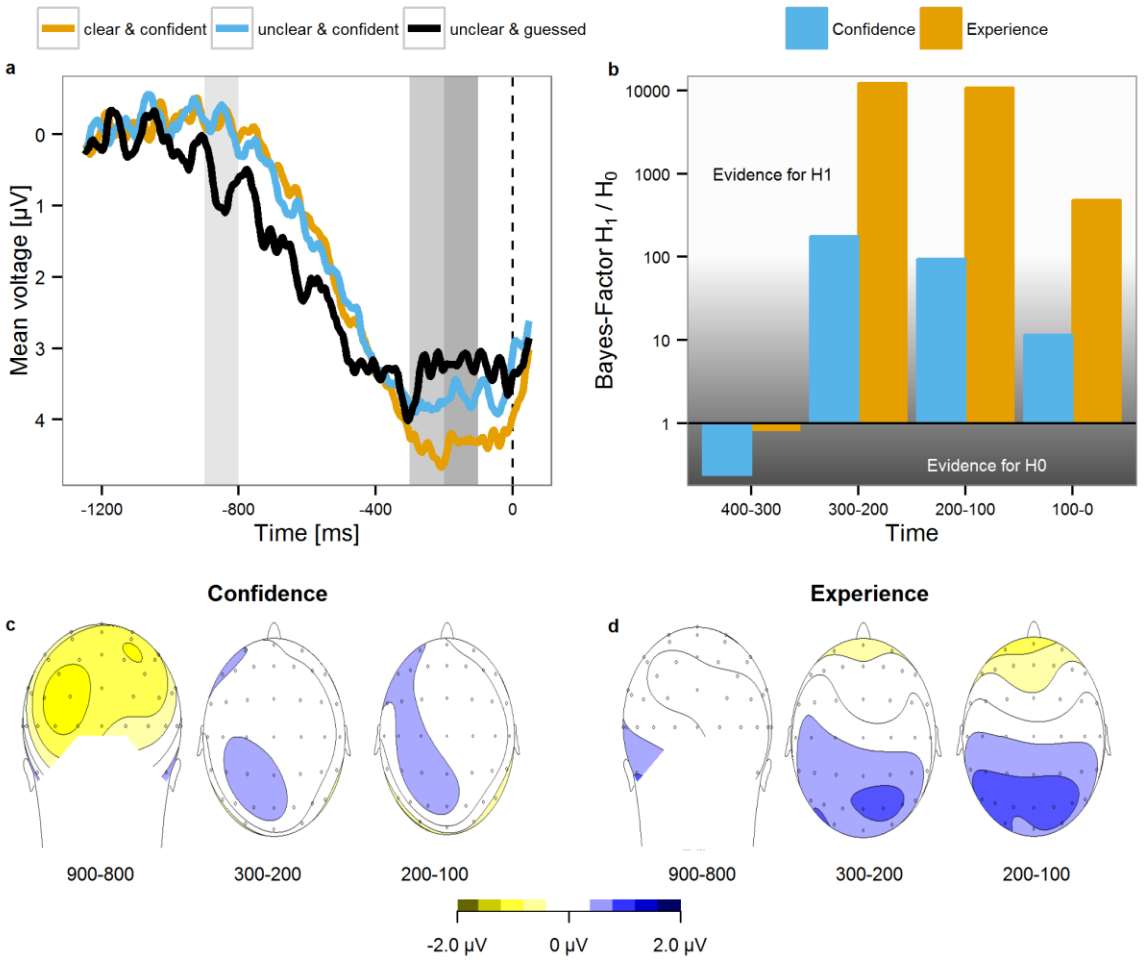

Figure 3-5. Time course of electrophysiological activity locked to the response. (a) Grand average ERPs of electrodes Pz and CPz plotted as a function of time and subjective reports. Orange line: ERP average when participants reported clear experience and being confident. Blue line: ERP average when participants reported unclear experience despite being confident. Black line: ERP average when participants reported an unclear experience and were unconfident about the discrimination response. (b) Bayes factors of the experience and confidence contrast in time windows of 100 ms each. (c) Scalp distribution of a difference wave of the confidence contrast for the time windows 900-800, 300-200, and 200-100 ms prior to the response. (d) corresponding voltage maps for the experience contrast.

The latency analysis of the parietal effects of experience and confidence revealed that the onsets and peaks of both effects were almost simultaneous with respect to the response. 25% area latency of the effect related to the experience contrast was 218 .3 ms prior to response, compared to 220.6 ms for the confidence contrast. 50% of the area was reached at 155.0 ms before the response for the experience contrast and 154.1 ms before the response for the experience contrast. The Bayes factor confirmed there were no differences between experience and confidence contrasts in onset, $BF_{10} = 0.26$, and peak, $BF_{10} = 0.33$.

3.4. Discussion

The present experiment was conducted to identify the mechanisms underlying subjective reports in situations when participants report being confident in discrimination decisions, but not yet report a clear experience of the stimulus. We predicted that if decisional confidence depends only on the stimulus feature relevant for the task response but visual experience requires additional stimulus features, ERPs associated with specifically experience and confidence might follow distinct time courses already at sensory time ranges. In contrast, if the behavioural differences between visual experience and decisional confidence were due to additional metacognitive processes specific to decisional confidence, sensory ERP correlates of experience and confidence should be the same, but decisional confidence should be more strongly associated with later ERPs. We observed that decisional confidence was more predictive for the accuracy of the task response than visual experience. ERPs suggested that both subjective reports of experience and decisional confidence were associated with a mid-latency negative shift over posterior electrodes; but the correlates of decisional confidence emerged about 40 ms earlier than those of visual experience. The strongest correlate of clear visual experience was not observed until about 200-150 ms before the objective discrimination response at centroparietal electrodes.

3.4.1. Why is confidence earlier than experience?

As perceptual decisions seem to logically depend on the outcome of stimulus perception, the neural correlates of subjective reports of experience are intuitively expected earlier in time than the neural correlates of decisional confidence. In contrast to this intuition, our study suggests the inverse temporal relation: ERPs associated with confidence occur earlier in time than those associated with a clear visual experience.

The most plausible explanation for the temporal delay between the correlates of experience relative to confidence lies in the nature of sensory evidence participants take into account when they make a subjective report about a discrimination decision, in contrast to subjective reports about their visual experience (Rausch et al., 2015): When participants report their confidence in an orientation discrimination judgment, they have to evaluate only those stimulus characteristics relevant for task (Dienes, 2008), which is the orientation in the present study. However, when participants report their experience of the grating, they may take more stimulus characteristics into account in addition to its orientation, although these

features were not relevant for the response to the task, for example the luminance of the stimulus, or the strength of figure-ground separation. According to the partial awareness hypothesis, conscious access to different stimulus features may vary independently (Kouider et al., 2010). If conscious access to features varies independently, and the set of features required for visual experience is larger than the set of features for confidence, it can also be understood why the conditions to report a visual experience is less frequently fulfilled (Sahraie et al., 1998; Schlagbauer et al., 2012; Zehetleitner & Rausch, 2013): In some situations, participants may have conscious access to the task-relevant stimulus feature(s), and thus report confidence about the task, while the additional features required for reporting visual experience are not accessible, and thus they report to have no visual experience. In the very same situations, confidence would also predict trial accuracy even in absence of visual experience (Charles et al., 2013; Rausch et al., 2015). In addition, as the task-relevant feature in the current task was orientation, a fairly basic feature, the extraction of additional features may require a longer period of time after the visual system has already determined the orientation, which is why confidence about the accuracy of the orientation response can be predicted from sensory ERPs earlier in time than reports of visual experience.

An explanation of the temporal delay between the correlates of experience and confidence based exclusively on metacognitive processes appears unlikely because this delay emerged already during a mid-latency negative shift. Although the cognitive functions associated with this ERP activity have not yet been fully identified, there is a consensus that it fulfils some sensory role (Koivisto & Revonsuo, 2010; Pitts et al., 2014; Verleger, 2010). However, distinct metacognitive processes are possibly engaged later on in evaluating the different sets of features relevant to experience and confidence (Charles et al., 2014, 2013; Overgaard & Sandberg, 2012).

Concerning the hypothesis that participants only place more conservative criteria for experience than for confidence on the same dimension of sensory evidence (Wierzchoń et al., 2012), the temporal delay between the correlates of experience and confidence could be explained if coarse evidence created in an early time range is sufficient for decisional confidence, and more refined evidence created later is required for visual experience. However, if identical evidence underlay subjective reports of experience and confidence, both should be equally efficient in predicting trial accuracy when criteria are controlled, which is not the case (Rausch et al., 2015; Sandberg et al., 2010; Zehetleitner & Rausch, 2013).

Overall, the differences between experience and confidence appear to be more fundamental than just placement of criteria.

3.4.2. The timing of neural markers of consciousness

The timing of neural markers of conscious awareness was proposed as a test of theories of the NCC (Lau, 2011), in particular if these neural markers occur rather early (during sensory processing) or late in time (i.e. after sensory processing). We observed that both subjective reports of experience and confidence can be predicted already from mid-range negative shift over posterior and temporal electrodes, suggesting that a substantial part of the neural processes that determine the contents of participants' reports coincide already with sensory processes. This timing of ERP correlates is consistent with theories predicting that conscious experience is associated with activity in sensory cortex (Block, 2005; Lamme, 2006; Zeki, 2003). This mid-range effect might reflect recurrent feedback along the visual ventral stream (Railo et al., 2011), which plays a key role as the substrate of consciousness in several theories (Block, 2005; Lamme, 2006). Other theories hold that consciousness depends on post-perceptual activity in parietal and frontal cortices (Baars, 2005; Dehaene & Changeux, 2011; Lau & Rosenthal, 2011). In the context of ERPs, advocates of global workspace theory argued that conscious awareness is associated with distributed activity starting 300 ms after stimulus onset, while earlier ERP correlates reflect only task performance. Consequently, the absence of an association between subjective reports and earlier ERPs has been interpreted as evidence for global workspace theory (Del Cul et al., 2007; Lamy et al., 2009). However, the sensory ERPs in the present study predicted verbal reports although stimulation was physically identical and only correct trials were taken into account. Nevertheless, the association between subjective reports and sensory ERPs can be accounted for by late theories of the NCC if it is assumed that the mid-range negativity reflects only the potential of the stimulus to become conscious, while conscious experience is instantiated only at later links of the causal chain that leads to a subjective report. As it is impossible to observe first-person experiences from a third-person-perspective (Jackson, 1982; Nagel, 1974), it is unlikely that these views can ever be tested empirically. What the present data does demonstrate though is that a substantial proportion of the variability of reports is determined already at the time of sensory processing, highlighting the importance of mid-range sensory processes for creating the content of subjective reports.

3.4.3. Confidence and experience are not interchangeable

In contrast to a prominent view in consciousness research (Lau & Rosenthal, 2011; Seth et al., 2008), subjective reports do not form one coherent category of measurements of consciousness. The qualitatively different time course of the correlates of subjective reports of experience and confidence suggests in line with previous experiments that these two are not interchangeable (Charles et al., 2013; Sahraie et al., 1998; Schlagbauer et al., 2012; Zehetleitner & Rausch, 2013): If participants were asked to report their visual experiences only, ERP correlates would be delayed compared to if participants were asked to report their confidence in their task responses. The temporal offset between the correlates of experience and confidence cannot be explained by the quality of the measurement: It might be argued that subjective reports of experience may be compromised with noise, thus explaining why reports of confidence detect effects in time ranges missed by reports of experience. However, as the degree of association always depends on the amount of noise in the measurement, it would follow that all associations of the noisy measurement with other variables were smaller. Consequently, if confidence reports were more reliable than reports of visual experience, the ERP effects of confidence would be greater than the ERP effect of visual experience in all time windows. On the contrary, reports of experience were at least as efficient as reports of confidence in predicting later ERPs. Overall, for a comprehensive theory of the NCC, we suggest consciousness research needs to investigate the neural correlates of confidence in relation to the neural correlates of experience.

3.5. Acknowledgements

This work was supported by the German-Israeli Foundation for Scientific Research and Development (1130-158) and the Deutsche Forschungsgesellschaft (ZE 887/3-1).

4. VISUAL ANALOGUE AND DISCRETE SCALES AS MEASURES OF VISUAL EXPERIENCE[8]

by Manuel Rausch and Michael Zehetleitner[9]

4.1. Abstract

Can participants make use of the large number of response alternatives of visual analogue scales (VAS) when reporting their subjective experience of motion? In a new paradigm, participants adjusted a comparison according to random dot kinematograms with the direction of motion varying between 0 and 360°. After each discrimination response, they reported how clearly they experienced the global motion either using a VAS or a discrete scale with four scale steps. We observed that both scales were internally consistent and were used gradually. The visual analogue scale was more efficient in predicting discrimination error but this effect was mediated by longer report times and was no longer observed when the VAS was discretized into four bins. These observations are consistent with the interpretation that VAS and discrete scales are associated with a comparable degree of type 2 sensitivity, although the VAS provides a greater amount of information.

4.2. Introduction

The lack of an established measurement for conscious experience is a key challenge to the prosperity of an empirical science of consciousness (Chalmers, 1998). The choice of an adequate measure is delicate because different theoretical perspectives on consciousness can imply different measurements. Some theorists are critical about the use of subjective reports because they assume participants might have conscious experiences they are unable to report (Block, 2005; Eriksen, 1960; Lamme, 2006) or they do not report because their criterion is too conservative (Hannula et al., 2005). In contrast, proponents of higher-order thought theories

[8] A version of this chapter has been published as 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

[9] Manuel Rausch and Michael Zehetleitner conceived the experiment, Manuel Rausch collected and analysed the data, Manuel Rausch, and Michael Zehetleitner co-wrote the manuscript.

often argue that subjective reports are more valid than objective measures because unconscious processes might drive objective performance as well (Dienes, 2004, 2008; Lau, 2008b). However, as subjective experiences cannot be observed from the third-person point of view (Jackson, 1982; Nagel, 1974), it is impossible to test empirically whether subjective measures of consciousness leave out conscious experiences that observers are unable to report, or whether objective measures suggest falsely that performance in a task is conscious. However, some researchers decide a priori to adopt a perspective that requires the use of subjective reports, either because they endorse a higher-order perspective on consciousness (Carruthers, 2011; Lau & Rosenthal, 2011; Timmermans et al., 2012), or because they consider subjective reports themselves as the subject of their scientific investigations (Dennett, 2003, 2007); if they do so, the empirical question arises how a scale needs to be designed given the metacognitive abilities of humans to obtain as much information from participants as possible.

4.2.1. The content of subjective scales

Subjective scales designed to measure conscious experience are constituted out of at least two components: (i) the question participants are instructed to answer and (ii) the way participants deliver their subjective report. Concerning the question, we proposed a classification of subjective scales on the event in the world subjective reports refer to, specifically whether subjective reports refer to the stimulus or to the discrimination response (cf. Chapter 2 and 3, Zehetleitner & Rausch, 2013). Examples for stimulus-related scales would be to ask participants how visible the stimulus was (Sergent & Dehaene, 2004), to rate clarity of the response defining feature (Zehetleitner & Rausch, 2013), or to report both the experience of specific features as well as feelings of something being shown (Ramsøy & Overgaard, 2004). Response-related scales may ask participants to report how confident they are about the preceding objective task response (Peirce & Jastrow, 1885), whether they attribute their objective task response to guessing, intuition, memory, or knowledge (Dienes & Scott, 2005), how much money they would wager on the accuracy of the objective task response (Persaud et al., 2007), or whether they experienced a “feeling-of-warmth” with respect to the previous task response (Wierzchoń et al., 2012).

Several studies compared subjective scales with different questions participants were asked to respond to: Dienes and Seth (2010) reported that wagering was biased by the participants’ risk-aversion, but there were no differences between confidence and wagering

after the possibility of loss had been eliminated from wagering. Sandberg, Timmermans, Overgaard, and Cleeremans (2010) observed in a masked object identification task that the perceptual awareness scale (PAS) predicted task performance more efficiently than confidence and wagering did. In an artificial grammar task, it was reported that confidence ratings predicted objective performance more efficiently than ratings of awareness of the artificial grammar rule (Wierzchoń et al., 2012). Szczepanowski, Traczyk, Wierzchoń, and Cleeremans (2013) reported that confidence ratings were more closely correlated with performance than ratings of subjective awareness and wagering, although a recent reanalysis of the data found no significant differences between subjective awareness and confidence (Sandberg et al., 2013). Finally, subjective reports of visual experience were less strongly correlated with objective performance in masked orientation discrimination tasks or random motion discrimination tasks, but no substantial differences were observed in a masked form discrimination task. In addition, confidence ratings were associated with more liberal thresholds than reports of visual experience across all three visual tasks, and confidence and wagering were more strongly correlated with each other than with reports of visual experience (Zehetleitner & Rausch, 2013).

Four different lines of interpretation for empirical differences between subjective scales with different questions have been suggested: First, it has been assumed (at least for the purpose of a comparison between measurements) that different kinds of subjective reports are equal except the sensitivity (Dienes & Seth, 2010) and the exhaustiveness of the scale (Sandberg et al., 2010). The second suggestion was that different scales might encourage participants to access their conscious contents in different ways: In introspective judgments, participants just directly report their conscious experiences a s they have them; in metacognitive judgments however, participant use their conscious experiences to make more complex cognitive judgments about processes engaged in the objective task (Overgaard & Sandberg, 2012). Third, it has been proposed that different subjective scales might alter the quality of conscious experience itself: Some scales such as wagering might be more motivating for the participants, making them more attentive, and thus cause participants to experience the stimulus more distinctively (Szczepanowski et al., 2013). Finally, it was suggested that different questions may relate to different processes during the task: Stimulus-related reports may be informed by processes involved in stimulus representation, and response-related reports by processes involved in decision making (Chapter 2, Zehetleitner & Rausch, 2013).

4.2.2. Visual analogue vs. discrete scales

The present study investigated the response format as the second component of subjective scales, specifically whether responses to the same question are more conveniently recorded by a discrete scale or a visual analogue scale (VAS). From the viewpoint of information theory (Shannon, 1948), subjective reports should be collected with a maximum number of scale steps because the maximal amount of information recorded by one report is bounded by number of options provided to the participant. Specifically, as the maximum information is computed as the binary logarithm of the number of options, a binary scale records the information of 1 bit in one trial, 4 scale points 2 bits, 8 scale points 3 bits, etc. The information conveyed by a VAS, where the response is selected along a continuum, would theoretically depend on the number of scale positions differentiated by the equipment (between 2^8 and 2^{16} with custom joysticks), but is in practice limited by the number of positions that participants can differentiate on the continuum, which classical studies estimated to be at least 10 positions (Hake & Garner, 1951).

From the viewpoint of signal detection theory (SDT) (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002), however, the use of a high number of scale steps is only feasible if two requirements are met:

(i) Participants need to be able to maintain a sufficient number of criteria.
(ii) Participants' type 2 sensitivity (Galvin et al., 2003), i.e. their degree of access to their own task performance, should not be impaired by a great number of options.

The recent literature has raised doubts about both requirements for high-precision usage of VASs: Overgaard, Rote, Mouridsen, and Ramsøy (2006) proposed that VASs tend to be used like binary judgments: As only the extreme ends of the scale are labelled, reports may be dragged towards the extremes, reducing the number of criteria participants effectively use to two. In addition, they argued as there are no definitions for each experience along the continuum of the VAS, VAS could confuse participants and result in less accurate reports.

Only one study so far has empirically compared a VAS and discrete scale: Wierzchoń et al. (2012) compared subjective reports of rule awareness with four scale steps against a VAS of rule awareness in a 2AFC artificial grammar classification task and observed a tendency that the four-point scale predicted performance more efficiently than the VAS (irrespective of whether the VAS was binned into four scale steps or not), although the

statistics were not significant. Wierzchoń et al. (2012) also found that rule awareness measured by a VAS was worse than wagering and feeling-of-warmth both measured by a discrete scale, although there was no significant difference between discrete rule awareness and these two scales; however, these findings are hard to interpret because the content of the scales and the response format are confounded in these comparisons. In domains other than awareness, VASs have been demonstrated to be adequate measurements for state anxiety (Davey, Barratt, Butow, & Deeks, 2007), vertigo (Dannenbaum, Chilingaryan, & Fung, 2011), quality of live (de Boer et al., 2004), group cohesiveness (Hornsey, Olsen, Barlow, & Oei, 2012), mood (Kontou, Thomas, & Lincoln, 2012), thermal perception (Leon, Koscheyev, & Stone, 2008), and depression (Rampling et al., 2012), indicated by a strong correlation with an established multi-item questionnaire or by a high reliability of VASs, suggesting that participants are in principle able to make meaningful reports using VASs (although it should be noted that these studies did not compare VASs and discrete scales directly). As VASs were shown to be adequate measurements for a considerable number of different psychological constructs, it is reasonable to hypothesize that a VAS might be a convenient measurement of visual experience as well. Apart from that, it was argued that a VAS may induce more careful responses because it signals to the participant that an exact response is important, while a discrete scale might convey the message that a rough answer is sufficient (Funke & Reips, 2012).

In summary, although VASs are in principle suited to record a large amount of information, it is an open empirical question whether participants are able to use a VAS with a sufficient number of criteria and without loss of type 2 sensitivity, so employing a VAS is feasible.

4.2.3. Continuous vs. binary discrimination task

While the study by Wierzchoń et al. (2012) contrasted subjective reports and objective performance in a 2AFC discrimination task, the recent development of continuous discrimination tasks (Bays & Husain, 2008; Zhang & Luck, 2008; Zokaei, Gorgoraptis, Bahrami, Bays, & Husain, 2011) offers the opportunity to conduct a more powerful test of the amount of information recorded by a VAS. For example, in a typical 2AFC task, participants might be instructed to report whether a previously presented bar is tilted towards left or right. The set of possible stimulus features is two (left or right) and so is the set of possible responses. This paradigm can be changed into a continuous discrimination task by allowing

the bar to have any of all possible orientation and asking the participant to indicate the orientation of the bar via a response set of the same cardinality. Errors, defined as the deviation of stimulus and response, are binary in a 2AFC paradigm: either the response corresponds to the stimulus (i.e., is "correct"), or it does not (i.e., is "incorrect"). For continuous tasks however, the deviance between stimulus and response is a continuous variable: When for instance the stimulus consists of a vertical bar, the response may deviate from the true orientation by any angle between 0° and 90°.

The number of task response alternatives is relevant for comparing different scales because the information recorded by a scale depends on the entropy of metacognition, which in turn depends on the entropy of discrimination performance: When there are only two levels of accuracy, i.e. "correct" and "incorrect", there will be a comparably small number of metacognitive states, and consequently, a smaller number of scale steps might perform well to categorize these states. In contrast, when participants are required to adjust a comparison continuously according to a specific stimulus feature, there is a large number of different possibilities how accurate discrimination performance can be, and thus a large number of possible metacognitive states. Consequently, a scale with a larger number of response alternatives might perform better than a discrete scale when the number of response alternatives is large.

In general, performance in a continuous adjustment task can be described mathematically by a combination of a von Mises and a uniform distribution (Bays, Catalao, & Husain, 2009; Zokaei et al., 2011): If participants had to rely completely on guessing, their responses should be evenly distributed across the whole range of possible responses. However, if performance is better than chance, their responses would form a bell-shaped distribution centred at the correct response, with the spread of the distribution indicating the precision of the response. A continuous task for the purpose of the current study would be characterized by a continuous relationship between task difficulty and the precision parameter as well as the guessing parameter. Previous studies suggested that subjective reports are associated with both the precision parameter as well as the probability of guessing in working memory tasks (Rademaker, Tredway, & Tong, 2012), but to our knowledge, no study has so far introduced continuous tasks in the study of visual consciousness.

4.2.4. Criteria to evaluate subjective scales

As the current experiments entails a comparison between scales with a different number of scale steps, special attention should be paid to the choice of operationally defined criteria to evaluate the scales. We propose to employ three criteria of comparison:

(i) the correlation with discrimination performance
(ii) the internal consistency
(iii) the distribution of ratings.

The correlation with discrimination performance as well as internal consistency come with two very different interpretations depending on whether the amount of information collected with one report is controlled or not. When VAS judgements are binned into the same number of scale steps as the discrete scale and thus the amount of information recorded by the two scales is balanced, the correlation of subjective reports with discrimination performance is indicative of type 2 sensitivity (Galvin et al., 2003), the ability to discriminate between correct and incorrect trials. This is the rationale of numerous previous studies (Dienes & Seth, 2010; Sandberg et al., 2010; Szczepanowski et al., 2013; Wierzchoń et al., 2012) and is analogous to the term resolution in the confidence literature (Baranski & Petrusic, 1994). In contrast, under the assumption that the type 2 sensitivity of participants is comparable, a comparison between the association of the full VAS and objective performance on the one hand and the association between the discrete scale and performance shows whether the VAS is able differentiate between levels of performance that fall equally on the same scale step with the discrete scale and is thus indicative of the amount of information recorded by the scale.

The second criterion we took into account was the internal consistency of subjective reports within experimental conditions: A scale should provide maximally stable estimates of averages of the subjective reports across a number of data points. Again, the comparison between the discretized VAS and a discrete scale shows whether one scale is corrupted from noise unrelated to the number of scale steps; while a comparison between the internal consistency of full VAS and discrete scales shows whether participants can make use of the additional resolution provided by the VAS, i.e. it examines whether VAS reports differentiate between trials that fall on the same scale step at the discrete scale.

Third, another characteristic of subjective scales that has been extensively discussed is the distribution of subjective reports when collected with different scales: Are subjective scales of consciousness used gradually or are they used in a binary fashion? While some scales might be designed in a way that all scale steps are used with relatively equal probability, other scales might induce binary responses (Overgaard et al., 2006). This empirical question is related to the theoretical proposals that consciousness is either dichotomous (Dehaene & Changeux, 2011; Dehaene, Sergent, & Changeux, 2003) or a gradual phenomenon (Cleeremans, 2008, 2011). If stimulus consciousness varies binarily (i.e. stimuli are always either fully conscious or completely unconscious), an observers would only use the ends of the scale, resulting in a U-shaped distribution of ratings. If stimuli however can be more or less conscious, all points of the scale are potentially used, when stimulus strength increases, resulting in a uniform distribution when averaged across stimulus strength. However, in order to investigate the issue whether consciousness varies gradually or binarily, a scale is required where participants in principle use the intermediate scale steps as well; otherwise a U-shaped distribution would be observed no matter whether consciousness in a specific task in fact gradual or dichotomous (Sergent & Dehaene, 2004).

4.2.5. Rationale of the present study

The aim of the present study was to investigate whether participants can make use of the high resolution offered by VASs when measuring visual experience of motion. To address this issue, we compared a VAS and a discrete scale with respect to the criteria discussed in 4.2.4. As stimuli, we presented random dot kinematograms (RDKs), because RDKs allow for a fine-grained manipulation of task difficulty on a metric scale (by manipulating the percentage of coherently moving dots). For the objective task, we assessed objective performance as a continuous variable rather than just correct or false, a procedure that ensured a binary use of subjective reports was not due to binary task performance. To obtain a continuous measurement of task performance, we asked participants to report the orientation of motion by adjusting a clock-hand to point into the direction of the perceived motion, and measured the discrimination error as the angle between clock-handle and direction of motion. For the subjective scales, we asked participants always to report their degree of experience of the coherent motion, which was the same instruction as we used in the previous experiments (Zehetleitner & Rausch, 2013), and different from the established Perceptual Awareness Scale (PAS, Ramsøy & Overgaard, 2004) in that no instruction to report feelings of something

being shown was given. The experiment was designed to investigate the following three hypotheses:

(i) If the participants are able to make use of the additional resolution provided by VASs, the full VAS should predict the discrimination error more efficiently than the discrete scale. In addition, the internal consistency of the full VAS should be better, because the larger amount of data transmitted by each single subjective report would allow for more reproducible statistics based on the same number of trials.

(ii) If VAS reduced the type 2 sensitivity of subjective reports, we would expect that the discrete scale would be more efficient in predicting discrimination error and would produce more consistent estimates than the discretized VAS.

(iii) If participants are biased by the anchors of the VAS in a way that reports are given binarily, the ratings on the VAS but not on the discrete scale should form a U-shaped distribution. In addition, the discrete scale should outperform both the full and the discretized VAS in predicting discrimination error.

4.3. Experiment

4.3.1. Material and Methods

4.3.1.1. Participants

20 participants (5 male, 1 left-handed) took part in the experiment. The age of the participants ranged between 19 and 32 years, with a median age of 24. All participants reported to have normal or corrected-to-normal vision, confirmed that that they did not suffer from epilepsy or seizures and gave written-informed consent

4.3.1.2. Apparatus and stimuli

The experiment was performed with a Mac with OS X 10.7 as operating system and a Diamond Pro 2070 SB (Mitsubishi) monitor with 24 inch screen size. Stimuli were presented at a refresh rate of 120 Hz controlled by MATLAB and Psychtoolbox 3.0.10 (Brainard, 1997; Pelli, 1997); code adapted from http://www.shadlenlab.columbia.edu/Code/VCRDM). The stimuli were random dot kinematograms, consisting of on average 150 small white squares (sized 2 x 2 pixels, luminance 85.0 cd/m^2) in from of a black background (1.3 cd/m^2), which appeared in a circular aperture (diameter: 5°) centred 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 among 1.6, 3.1, 6.2, 12.5, 25, and 50%. The direction of movement was randomly chosen out of each possible direction. To record the orientation judgment, 12 circles (diameter: 0.2°, 2.2 cd/m^2) were displayed on the screen, forming one large circle centred at the screen with a diameter of 10°. Participants indicated the direction of motion and their rating on the VAS by a Cyborg V1 joystick (Cyborg Gaming, UK). The clock-hand consisted of a bar (length: 5°, width: 0.1°, 2.2 cd/m^2) and a circular head (diameter: 0.2°, 2.2 cd/m^2).

4.3.1.3. Trial structure

The trial structure is shown in Fig. 4-1. Each trial began with the presentation of a fixation cross at screen centre for 1,000 ms. Then a RDK was presented for 2,000 ms. Next, the circle around the screen centre appeared. As the participants started to move the joystick, the clock-hand appeared, pointing to the direction the joystick was moved to. The circle continued to be displayed on the screen until participants had pulled and released the trigger of joystick. Next, the subjective scale appeared, with either the four response categories from the discrete scale or the VAS. If the error of the orientation judgment had been larger than 45°, the trial ended with the display of "please indicate the direction more carefully" for 1,000 ms.

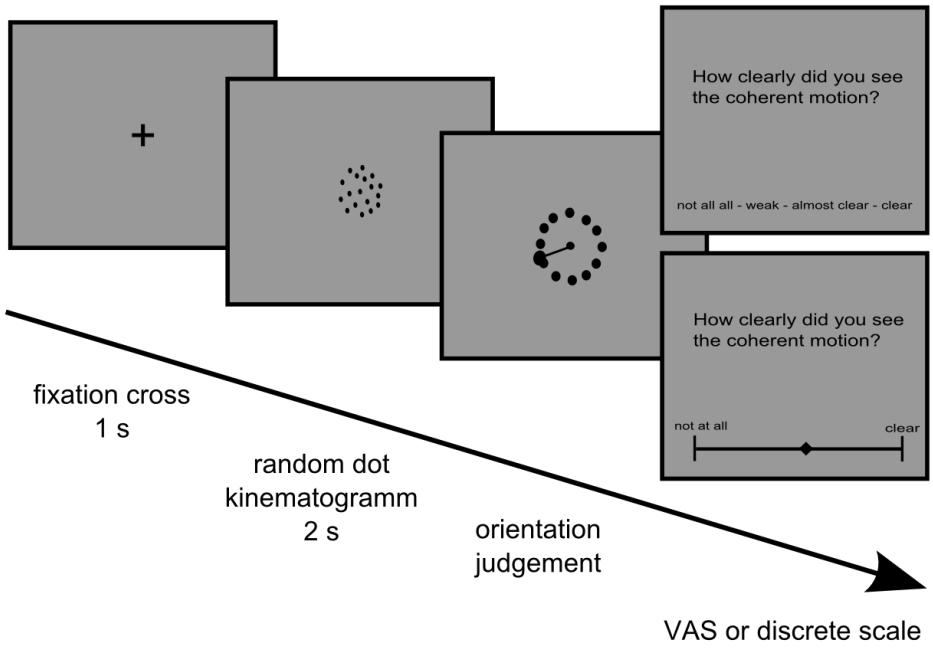

Figure ´4-1. Experimental procedure.

4.3.1.4. Procedure

The experiment lasted 1 hour on average. Participants were instructed perform the motion discrimination task as carefully as possible, with accuracy being more important than speed. For the subjective reports, participants were told that the subjective scale referred to the global motion experience created by the coherently moving dots. Again, participants were instructed to their ratings as carefully and as accurately as possible.

Participants indicated the direction of motion by using the joystick to move a bar that looked like a clock-hand. When the participants had moved the clock-hand in the direction they saw the dots moving, they confirmed their response by pulling the trigger of the joystick. The clock-hand consisted of a bar (length: 5°, width: 0.1°) and a circular head (diameter: 0.2°). To collect the subjective report, the question "how clearly did you see the coherent motion?" was displayed on the screen. In case of VAS, a continuous scale was shown underneath the question, with the ends labelled as "not at all", and "clear". Participants moved an index on the continuous scale by moving the joystick horizontally, and confirmed a position on the scale by pulling the trigger. In case of four point scales, the same question was displayed on the screen, but underneath the question, four response categories were shown, which were "not at all", "weak", "almost clear", and "clear". Participants responded to the discrete scales by pressing the keys 1, 2, 3, and 4 on the keyboard. At the beginning of the experiment, participants performed a training block with 24 trials. The main experiment involved 10 blocks with 45 trials each. During training, VAS and discrete scale trials were randomly intermixed each of the six possible coherences was presented six times in a from-easy-to-difficult order. During the main experiment, the two subjective scales alternated after each block and the levels of coherence varied randomly between trials.

4.3.1.5. Analysis

All analysis were performed in R 2.15.2 (R Core Team, 2012). For both the distribution analysis as well as the regression analysis, fast responses (defined as faster than 200 ms) and slow responses (defined as 2.5 standard deviations slower than the individual average) to the discrimination task or to the scale were omitted. Other exclusion criteria such as 2 or 3 standard deviations gave essentially the same results.

4.3.1.5.1. Distribution analysis of the discrimination responses

Discrimination responses were analysed by fitting a combination of a von Mises distribution and a uniform distribution to the data (Bays et al., 2009; Zokaei et al., 2011). The

uniform distribution models the distribution of responses in trials when participants relied on guessing, because when participants guessed, each orientation between 0 and 360 ° was equally probable. The von Mises (circular Gaussian) distribution centred at the true motion direction represents the distribution of responses in trials where participants were not guessing. The better participants performed the orientation judgment, the less responses jittered around the true motion direction; therefore, the concentration parameter of the von Mises distribution can be interpreted as the precision of orientation judgments. The model is described by the following equation:

$$\hat{\theta} = (1-\gamma)\phi_K\left(\hat{\theta}-\theta\right) + \gamma\frac{1}{2\pi}$$

where θ is the stimulus motion direction, $\hat{\theta}$ is the motion direction indicated by the participant, γ is the proportion of trials when participants were guessing, ϕ denotes the von Mises distribution with mean of zero and the concentration parameter K. Fitting was performed on the aggregated data across all participants and scales but separately for each level of coherence using maximum likelihood estimation and confidence intervals around each parameter were estimated using 10,000 bootstrap samples. Pooling over participants and scales was necessary to obtain a sufficient number of trials for the fitting algorithms to reach convergence. The purpose of this analysis was a manipulation check if performance in the current task was continuous or binary. As the hypotheses tested in the current study equally apply to metacognition of the precision as well as the guessing aspect of performance, it was legitimate to analyse the relationship between subjective reports and performance without differentiating between guessing and precision (see section 4.3.1.5.2.)

4.3.1.5.2. Relationship between scales and discrimination error

The relationship between the two scales and discrimination error was analysed by means of mixed model regression analysis based on the cumulative proportional odds model as implemented in the R library ordinal (Christensen, 2013), the ordinal equivalent to the analysis in previous studies (Sandberg et al., 2013, 2010; Wierzchoń et al., 2012). The dependent variable, the discrimination error, was determined by the absolute difference between the true motion direction and the reported motion and binned into 12 equal bins between 0 and 90° and a thirteenth bins for errors larger than 90° to allow computation of a proportional odds model. Non-parametric statistics were used to account for the fact that the discrimination error was bounded and strongly skewed. Inter-subject variance was modelled

by a random effect on the intercept. Scale (VAS vs. discrete scale), coherence (1.6 vs. 3.1 vs. 6.2 vs. 12.5 vs. 25.0 vs. 50.0) and subjective report and all interactions were treated as fixed effects. Significance of each fixed term was assessed by likelihood ratio tests between the full model and a model where the term was dropped. Confidence intervals were obtained from the likelihood root statistic. Subjective reports given by VAS and discrete scales were standardized separately. To investigate the effects of number of scale steps, two separate models were computed, one with the full VAS included as predictor, and one model where the VAS was binned into four equal partitions. We interpret a comparison between the discretized VAS and the discrete scale as indicative of type 2 sensitivity (i. e. the degree to which participants can access to their own performance) because when the VAS reports are binned to four, the amount of information in discretized VAS and discrete scale are the same, although we acknowledge that ordinal statistics do not provide any means of control over the influence of discrimination bias (Masson & Rotello, 2009). Given that type 2 sensitivity of VAS and discrete scale are the same, a comparison between the full VAS and the discrete scale is indicative of whether participants apply more criteria in the VAS than in the discrete scale and thus the full VAS discriminates between levels of performance that fall on the same scale step with the discrete scale. In addition, we analysed the effects of feedback and reporting time by computing two additional models comparing full VAS and the discretized scale with feedback and report time as additional fixed effect, respectively.

4.3.1.5.3. Internal consistency

Internal consistency was assessed by computing Cronbach's alpha (Cronbach, 1951) separately for each level of coherence using the R library ltm (Rizopoulos, 2006). Confidence intervals were estimated around Cronbach's alpha values based on 10,000 Bootstrap samples.

4.3.1.5.4. Distribution of subjective reports

To analyse the distribution of subjective reports, the ratings of VAS was again binned into four categories each covering a fourth of the scale range. The frequency of each bin was then compared against frequency of the corresponding response alternative of the discrete scales using an ANOVA with the factors rating category, coherence, and scale type (VAS vs. discrete scale). When sphericity did not hold, we adjusted the degrees of freedom according to the Greenhouse-Geisser correction. To resolve interactions, post-hoc t-tests were conducted comparing the frequency of each VAS bin with the corresponding response category of the

discrete scale separately for each level of coherence. P-values were adjusted by the Holm-correction to account for multiple comparisons.

4.3.2. Results

4.3.2.1. Discrimination performance

The mean discrimination error was 55.6° (SEM = 2.2) when participants were using the VAS and 56.3° (SEM = 2.2) when the discrete scale was used and ranged from 87.7° (SEM = 1.7) for the lowest to 13.7° (SEM = 1.6) for the highest level of coherence. The relative frequencies of orientation responses and the estimated distributions are shown in Fig. 4-2. The estimated parameters as well as bootstrapped confidence intervals are shown in Fig. 4-3.

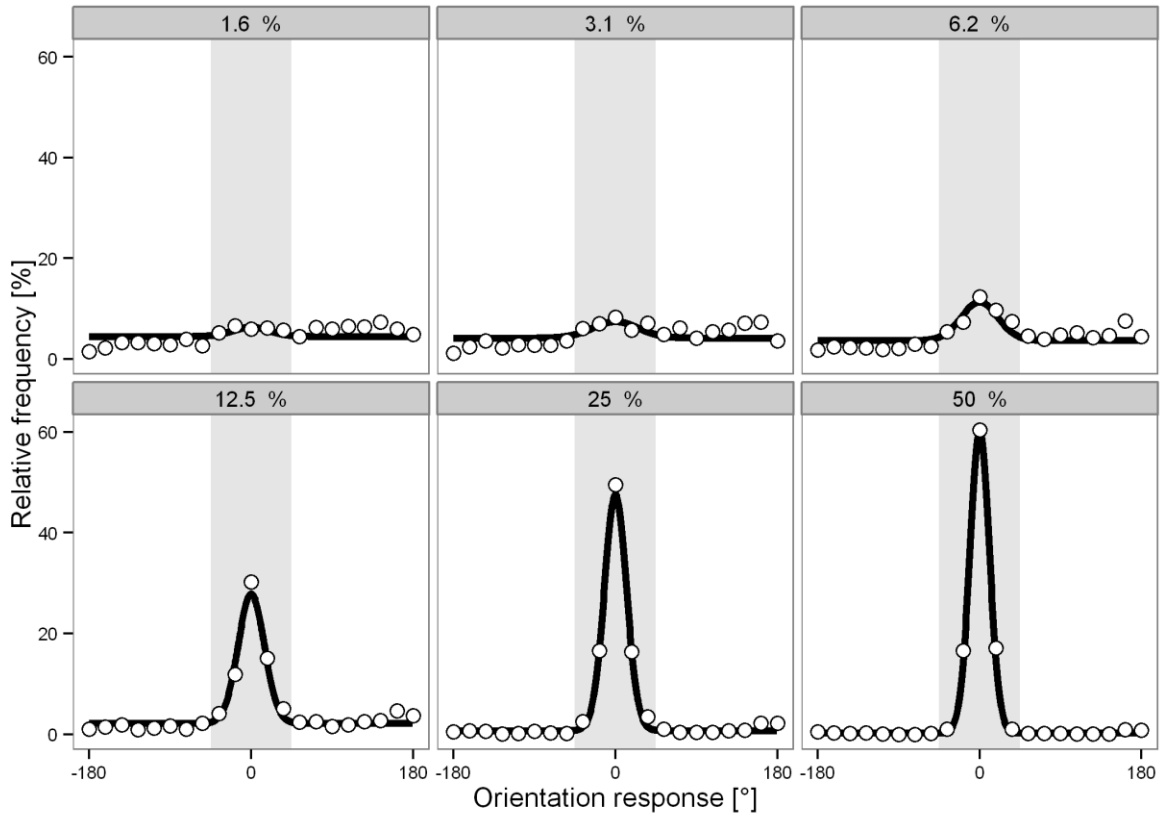

Figure 4-2. Distribution analysis of discrimination responses. Dots indicate the relative frequency of orientation responses with 0 as the true motion direction with different levels of coherence in each panel. Lines indicate the distribution of responses estimated from the fitted guessing and precision parameters. The grey highlighted area indicates the degree of accuracy between -45° and 45° where no error feedback was given.

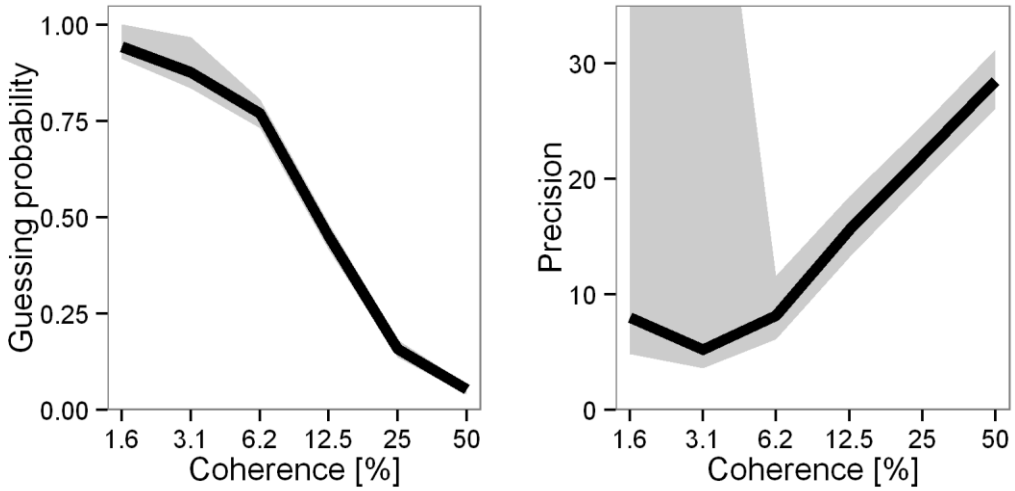

Figure 4-3. Estimated parameters from the distribution analysis plotted as a function of Coherence. Left Panel: Guessing probability. Right Panel: Precision. The grey areas indicate 95% bootstrapped confidence intervals.

The probability of guessing trials ranged between .94 at the lowest and .05 at the highest level of coherence. Confidence intervals indicated the guessing probability continuously decreased across all levels of coherence. The precision ranged between 5.2 at a Coherence of 3.1% and 28.5 at the maximum level of coherence. Confidence intervals suggested that there was a continuous increase of precision starting at a coherence of 6.2%, while the estimation of the precision parameter was not reliable for coherence levels of 1.3% and 2.6% (due to the low number of non-guessing trials).

4.3.2.2. Relationship between discrimination error and subjective reports

The regression weights and confidence intervals of the ordinal mixed model regression comparing the full VAS against the discrete scale as predictors of discrimination error can be found in Table 4-1. Likelihood ratio tests suggested significant main effects of subjective report [$\chi^2(1) = 195.0$, $p < .001$] and coherence [$\chi^2(5) = 1522.0$, $p < .001$], no effect of scale [$\chi^2(1) = 2.1$, n. s.], significant interactions between subjective reports and scale [$\chi^2(1) = 4.3$, $p < .05$] and between subjective reports and coherence [$\chi^2(5) = 50.2$, $p < .001$], and no three-way interaction [$\chi^2(5) = 6.8$, n. s.]. A regression model fitted on VAS ratings only revealed a regression coefficient for subjective reports of -.44 with a 95 % confidence interval of [-.52 -.36]. For the discrete scale, the same analysis revealed a coefficient of -.32 within a confidence interval of [-.40 -.24].

Table 4-1

Results of the mixed-effects ordinal regression model with discrimination error as dependent variable

Predictor	β	95 % CI		Likeli-hood ratio	df	p
		lower	upper			
Subjective report	-0.38	-0.43	-0.32	195.0	1	< .001
Coherence level				1522.0	1	< .001
– 1.6% vs. 50%	1.28	1.17	1.40			
– 3.1% vs. 50%	1.06	0.95	1.18			
– 6.2% vs. 50%	0.68	0.58	0.79			
– 12.5% vs. 50%	-0.39	-0.48	-0.29			
– 25% vs. 50%	-1.16	-1.27	-1.05			
Scale type	-0.01	-0.06	0.04	2.1	1	n. s.
Subjective report * coherence level				50.2	5	< .001
– Subjective report * 1.6% vs. 50%	0.26	0.14	0.38			
– Subjective report * 3.1% vs. 50%	0.21	0.09	0.33			
– Subjective report * 6.2% vs. 50%	0.01	-0.10	0.13			
– Subjective report * 12.5% vs. 50%	-0.26	-0.37	-0.16			
– Subjective report * 25% vs. 50%	-0.15	-0.26	-0.04			
Subjective report * scale type	0.06	0.01	0.11	4.3	1	< .05
Coherence level * scale type				10.3	5	n. s.
– 1.6% vs. 50% * scale type	0.10	-0.02	0.21			
– 3.1% vs. 50% * scale type	0.06	-0.06	0.17			
– 6.2% vs. 50% * scale type	0.10	-0.01	0.20			
– 12.5% vs. 50% * scale type	-0.05	-0.15	0.04			
– 25% vs. 50% * scale type	-0.03	-0.14	0.07			
Subjective report * Coherence level * scale type	-0.06			6.8	5	n. s.
– Subjective report * 1.6% vs. 50% * scale type	-0.02	-0.14	0.10			
– Subjective report * 3.1% vs. 50% * scale type	0.01	-0.11	0.12			
– Subjective report * 6.2% vs. 50% * scale type	-0.09	-0.21	0.02			
– Subjective report * 12.5% vs. 50% * scale type	-0.06	-0.17	0.04			
– Subjective report * 25% vs. 50% * scale type	0.08	-0.03	0.19			

Discrimination error as a function of coherence, scale, and subjective report (with discretized VAS ratings) are depicted in Fig. 4-4. The ordinal regression model comparing the discretized VAS and the discrete scale revealed significant main effects subjective report [$\chi^2(1) = 178.6$, $p < .001$] and coherence [$\chi^2(5) = 1586.2$, $p < .001$], no effect of scale [$\chi^2(1) = 2.4$, n. s.], a significant interaction between subjective reports and coherence [$\chi^2(5) = 47.4$, $p < .05$], but no interaction between subjective report and scale [$\chi^2(1) = 1.5$, n. s.], and no three-way interaction [$\chi^2(5) = 6.8$, n. s.].

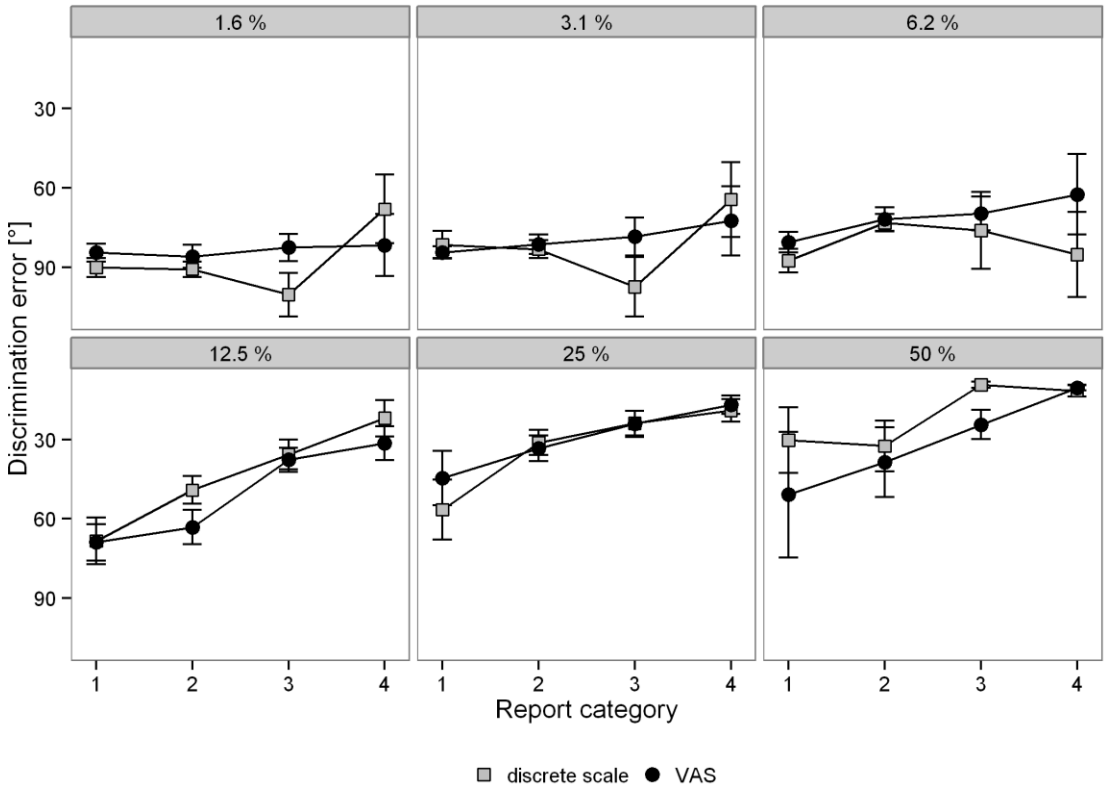

Figure 4-4. Discrimination error as a function of subjective reports, scale, and levels of coherence. The ratings on the visual analogue scale were discretized into four bins based on individual quartiles. A discrimination error of 90° indicates chance performance.

The frequency of feedback, which was provided after discrimination responses with an error greater than 45°, did not substantially differ between VAS trials (M = 40.3, SEM = 1.7) and discrete scale trials (M = 41.0, SEM = 2.1) [$t(19) = .7$, n. s.]. Including feedback on the previous trial into the ordinal regression analysis as an additional predictor revealed no effect of feedback [$\chi^2(1) = 0.1$, n. s.], no interaction between subjective reports and feedback, [$\chi^2(5) = 0.8$, n. s.], between scale and feedback [$\chi^2(1) = 2.1$, n. s.], or between scale, subjective reports, and feedback [$\chi^2(1) = 2.1$, n. s.]. Importantly, the interaction between scale and

subjective report was still significant when feedback was included into the analysis [$\chi^2(1) = 3.9$, $p < .05$].

For the VAS, the mean report time, i.e. the time between the orientation judgment and the subjective report, was 1329 ms (SEM = 95.6), compared to 944 ms (SEM = 73.3) with the discrete scale. As can be seen from Figure 4-5, ordinal regression slopes increased with report time for the VAS, while no such a relation was apparent for the discrete scale. The regression model with report time as additional predictor revealed a significant main effect of report time [$\chi^2(1) = 4.0$, $p < .05$], no interaction between report time and scale [$\chi^2(1) = 0.1$, n. s.], and between subjective report and time [$\chi^2(1) = 1.1$, n. s]. There was however a three-way interaction between subjective reports, scale, and report time [$\chi^2(1) = 5.5$, $p < .05$]. When response times were included into the model, the interaction between subjective reports and scale was no longer significant, [χ^2 (1) = 2.7, n. s]. Separate analyses of the impact of the report time on discrete scales and VAS revealed that the predictive efficiency of subjective reports made with the VAS interacted with rating time [$\chi^2(1) = 6.1$, $p < .05$], while subjective reports on the discrete scale were not influenced by rating time [$\chi^2(1) = 0.4$, n. s]. Overall, this pattern indicates that the differences in predictive power for discrimination error between the VAS and the discrete scale are mediated by longer report times.

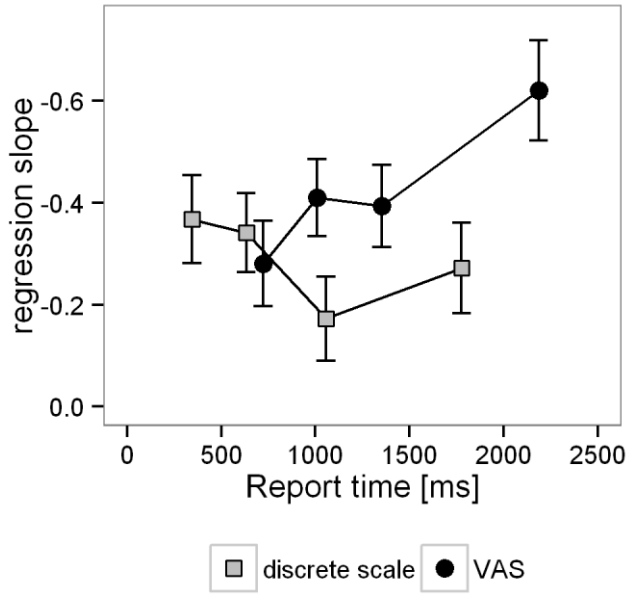

Figure 4-5. Ordered logistic regression slope of discrimination error predicted by subjective report depending on report time, i. e. time between objective task response and subjective report, and scale. To allow fitting separate regression models, report time is discretized into four bins based on the .25, .5, and .75 quantile.

4.3.2.3. Internal Consistency of subjective reports

Cronbach's alphas ranged between .83 and .93 for the discrete scale, between .84 and .93 for the discretized VAS, and .85 and .93 for full VAS (see Table 4-3). There was a numeric trend that alphas were larger for both the full and the discretized VAS than for the discrete scale at four out of six levels of coherence, but confidence intervals indicated the only substantial difference between the two scales was at a coherence of 6.2%, where the internal consistency of the VAS was greater. The internal consistency of the discretized VAS was always within the confidence intervals around the full VAS.

Table 4-3

Cronbach's alpha of VAS and discrete scales separately each level of coherence.

Coherence	Full VAS			Discretized VAS			discrete scale		
	alpha	CI 2.5	CI 97.5	alpha	CI 2.5	CI 97.5	alpha	CI 2.5	CI 97.5
1.6	.91	.82	.95	.87	.77	.92	.85	.68	.91
3.1	.91	.81	.95	.89	.78	.93	.86	.73	.91
6.2	.92	.87	.95	.91	.85	.94	.83	.63	.88
12.5	.85	.67	.91	.84	.67	.90	.85	.69	.90
25.0	.93	.80	.96	.92	.79	.96	.90	.76	.95
50.0	.93	.83	.96	.92	.83	.96	.93	.83	.96

4.3.2.4. Distribution of subjective reports

The mean subjective experience reported on the VAS was 49.2% of the scale range (SEM = 2.2) and 2.3 (SEM = 0.1) on the discrete scale ranging between 1 and 4 (which corresponds to a mean of 41.3% of the scale range and a standard error of 2.2 %). As can be seen from Fig. 4-6, the second scale step of the discrete scale was the dominant response even at a coherence of 1.3% when performance was effectively at chance. The ANOVA on response frequencies revealed a significant main effect of rating category [$F(2.0,38.8) = 5.3$, $p < .001$], significant interactions between scale and rating category [$F(3,57) = 17.4$, $p < .001$], and between rating category and coherence [$F(3.6,68.0) = 53.4$, $p < .001$], as well as a three-way interaction between coherence, scale type, and rating category [$F(5.3,99.9) = 7.2$, $p < .001$]. Post-hoc tests assessing whether the frequency of responses was different between the discrete scale and the VAS separately for each level of coherence and each response category are shown in Table 4-2. While there was no significant difference between reports of no experience on the discrete scale and the corresponding scale part of the VAS at each

coherence, reports of weak experiences occurred more often with the discrete scale than with the VAS at 5 out of 6 coherences, reports of almost clear experiences were more frequently reported with the VAS at lower coherences, and reports of clear experiences were more often with the VAS at a coherence of 25%.

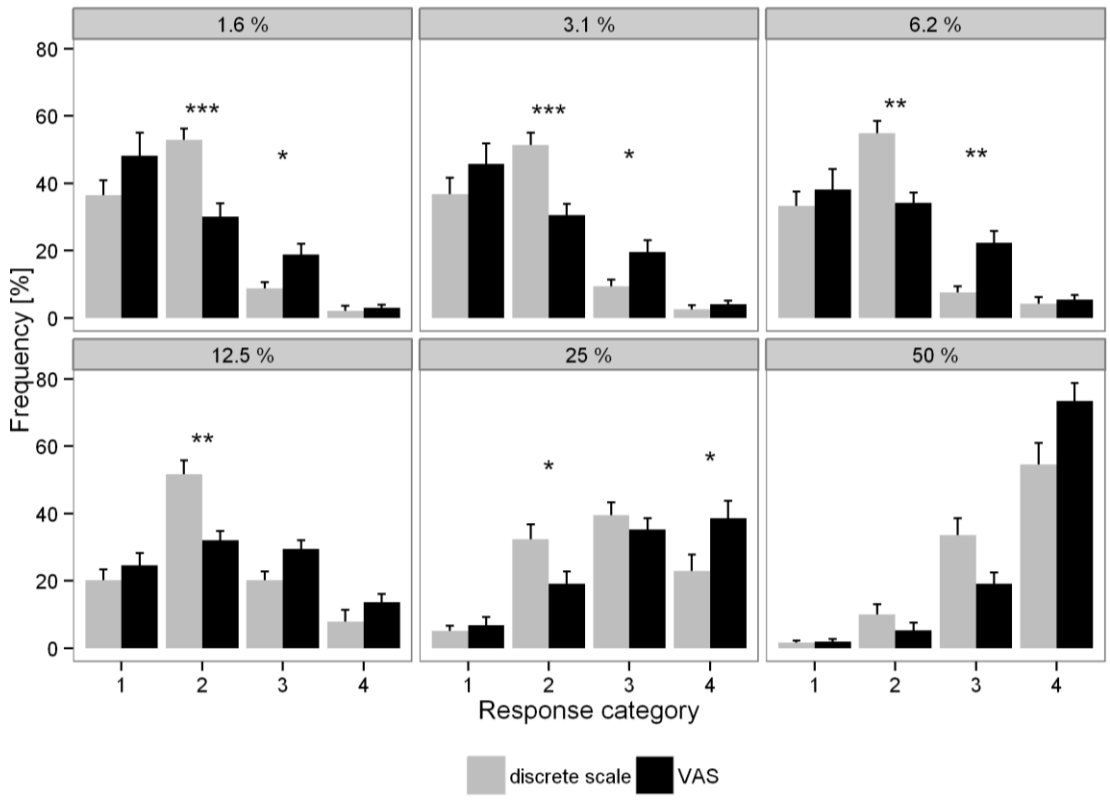

Figure 4-6. Frequency of each scale step of the discrete scales and the frequency of the corresponding scale parts of the VAS. Black bars indicate the VAS and grey bars the discrete scale. Error bars indicate 1 standard error of the mean.

4.4. Discussion

The present experiment investigated whether participants are able to use the high number of response alternatives provided by visual analogue scales appropriately when reporting visual experience of motion. We hypothesized that if a VAS allowed to retrieve a larger amount of information from participants' reports than discrete scales, the full VAS should be more efficient in predicting the discrimination error, and should be more internally consistent. Second, if a VAS reduced the type 2 sensitivity of subjective reports, we would expect that the discretized VAS should be less efficient in predicting the discrimination error than the discrete scale. Finally, if participants tended to use VASs in a binary way, ratings on

the VAS should form a U-shaped distribution, and the discrete scale should correlate more closely with discrimination error no matter whether the VAS is discretized or not.

Concerning the relationship between subjective reports and discrimination error, the full VAS predicted discrimination error more efficiently than the discrete scale, while there were no substantial differences between the discretized VAS and the discrete scale. The difference between the full VAS and the discrete scale was mediated by the response time to the scale. The analysis of internal consistency revealed no substantial differences for five out of six coherences, while both the full and the discretized VAS were more consistent at a coherence of 6.2%. Concerning the distribution of subjective reports, we observed that the VAS and the four-point scale were both not used in an all-or-nothing fashion, although participants had a tendency to report weak experiences in the discrete scale while they would report almost clear and clear experiences in the VAS.

4.4.1. The amount of information in VAS and discrete scales

According to a standard interpretation of differences between scales measuring subjective awareness, subjective reports are created by the same mechanisms, and differences between scales occur due to different qualities of the scale. A key aspect of the quality of the scale is the amount of information transmitted by each rating. According to information theory (Shannon, 1948), subjective reports collected by VAS should provide a larger amount of information that discrete scales, because 4 scale steps allow to record 2 bits of information, while the number of bits collected by a continuous scale is limited only by the number of positions participants are able to differentiate, and was estimated to be at least 10 positions (Hake & Garner, 1951), i.e. at least 3.32 bits. Consistent with the predictions from information theory, the full VAS was more closely correlated to the discrimination error than the discrete scale. We did not detect any substantial differences between the discrete scale and the discretized VAS in terms of type 2 sensitivity, suggesting that the additional alternatives participants have to consider when using a VAS did not add substantial amounts of noise to the subjective reports. Concerning internal consistency, there were no substantial differences between the two scales in five out of six coherences, although the VAS was more reliable at a coherence of 6.2%. Overall, it seems that a VAS indeed provides a larger amount of information than discrete scales, although the amount of information recorded by discrete scales is sufficient to provide reliable estimates as well.

4.4.2. The impact of report time

The difference between reports on the full VAS and on discrete scale in predicting discrimination error was mediated by the time of rating: While VAS ratings became more and more efficient in predicting trial accuracy with time, we observed no such a relation for the discrete scale. The first interpretation to these results is that a VAS provides a larger number of response alternatives, and selecting one out of this multitude of options could be more difficult and thus require a longer period of time. Second, it should be noted that also the motor response required by a VAS is more time-consuming than a simple button press: The association of the rating-accuracy relationship and report time at the VAS could also reflect the additional time demand of using a joystick and a decrease of rating precision when participants did not invest enough time to operate the joystick carefully. Third, an alternative explanation to these findings may be based on the dynamics of decision making: While standard SDT models of subjective reports assume that the evidence used for subjective reports is fixed at the time when observers respond to the task (Kepecs et al., 2008; Ko & Lau, 2012; Vickers, 1979), others have proposed that subjective reports are based on evidence participants continue accumulating after the objective decision is made (Pleskac & Busemeyer, 2010). Given that VAS judgements were associated with prolonged time participants needed to give a subjective report, post-decisional accumulation of evidence might be an alternative explanation why ordinal regression slopes are higher with VAS than with a discrete scale, because the additional 400 ms that it takes to make a judgement on the VAS might give participants more time to accumulate evidence. However, we observed a large overlap in the report times between the two scales in the current experiment where VAS regression slopes were larger although the time of the report was the same. In addition, while ordinal regression slopes seemed to increase almost linearly for the VAS, we found no indication of post-decisional accumulation for the discrete scale at all. What is possible is that participants keep accumulating sensory evidence after the decision when using the VAS only, either because they need the additional evidence to make fine-grained VAS ratings, or because they might be more motivated when using a VAS (Funke & Reips, 2012). The (cognitive and motor) cost of precise reporting and on-going accumulation accounts cannot be distinguished on grounds of the current data set. Given that a previous study failed to find any association with report time for both the VAS and discrete scales (Wierzchoń et al., 2012), future studies may be necessary to investigate the dynamics of metacognition.

4.4.3. Are visual analogue scales used binarily?

VAS received criticism because the continuum in combination with the labelled scale ends might result in a bimodal distribution of subjective reports, with scale extremes being chosen more frequently than the centre of the scale (Overgaard et al., 2006). First, we observed that intermediate scale steps were chosen frequently for both scales. Second, there was no difference between the frequency of the smallest scale step of the discrete scale and the lowest quarter of the VAS, indicating that VAS and discrete scales both applied the same minimal criteria for subjective reports. However, the second smallest scale step of the discrete scale (labelled as "weak") was more often chosen than the corresponding part of the VAS, while stronger experiences were more frequently reported with the VAS than with the discrete scales. There might be several explanations why more distinct experiences are more frequently reported with the VAS: First, participants could be biased by the labelled extremes of the scale (Ramsøy & Overgaard, 2004). Second, participants might suffer more strongly from an error of central tendency when they respond to discrete scales, and therefore the second scale step was the dominant response in the discrete scale. Finally, it is also possible participants are more motivated when using the VAS, being more attentive, and therefore have in fact clearer experiences (Szczepanowski et al., 2013). Concerning the impact of motivation, the two scales were associated with a comparable discrimination error, suggesting that the scale did not alter the way participants performed the task in general. Concerning a potential bias towards extremes, it should be noted that intermediate positions on the VAS were the most frequent responses for medium levels of coherence, suggesting that participants do use the centre of the scale when they consider it to be appropriate. In contrast, the second scale step of the discrete scale was the dominant response even at the lowest level of coherence when discrimination accuracy was effectively at chance, suggesting that the error of central tendency might be a factor in the distribution of discrete scales. The distribution of VAS is more plausible in a way that low ratings are dominant at low levels of coherence, intermediate ratings at medium coherences, and high ratings at high levels of coherences.

4.4.4. Discussion of methodology

It should be noted that the current experiment differs from previous studies addressing the topic of subjective reports in several ways. As this task is new to the field of metacognition, future studies are desirable to explore whether the findings obtained with this method are corroborated in more standard experiments. Most importantly, we quantified

discrimination error as a continuous variable rather than binary in the current study. In general, such an approach seems promising for the field of consciousness research because some theories of consciousness make specific predictions whether consciousness is gradual (Cleeremans, 2008, 2011) or dichotomous (Dehaene & Changeux, 2011; Dehaene et al., 2003), and recording performance in a non-binary way ensures that binary task performance does not artificially cause binary metacognition. Unfortunately, up to know, there is no proposal for a SDT-grounded measure of type 2 sensitivity equivalent to the measures applicable for binary tasks, so our analysis of type 2 sensitivity by ordinal regression does not provide the same control of response bias and confidence thresholds than it is possible for binary tasks. For the purpose of the current study, these potential confounds do not change the interpretation of the data because they would either affect the discrete scale, the full VAS, and the discretized VAS in the same way (response bias), or would affect the full VAS and the discretized VAS to the same degree (confidence thresholds), so it cannot be explained why only the full but not the discretized VAS provides more predictive power. Future studies however need to carefully consider the conceptual advantages of continuous tasks against the methodological disadvantages of the analysis methods available.

It may also be objected that the current task was not as continuous as it could have been, since all responses at above chance performance were concentrated between 45 and -45 degrees (where no error feedback was given), and thus the feedback might have motivated participants to perform at least as accurate as +/- 45 degrees. However, the precision of orientation judgements increased almost linearly although participants no longer received feedback, indicating participants did not perform the task in a binary fashion. As feedback might also have altered performance and type 2 sensitivity in the current task, parameters and coefficients estimated from the current experiment should not be naïvely expected to be the same in standard subliminal perception tasks where error feedback is suspended after a training period or is completely missing. Nevertheless, we did not observe any evidence that feedback on the previous trial influenced any contrast of interest for purpose of the current experiment, suggesting that feedback did not have a major impact on performance in the current study.

4.4.5. Equivalent conscious access?

Another interpretation of differences between various subjective scales is that different scales might encourage participants to use different mechanisms of conscious access to report

their conscious experiences (Overgaard & Sandberg, 2012). Indeed, it is plausible to assume that subjective reports in VASs and discrete scales are accomplished in parts by different processes. Discrete scales rely strongly on verbal categorization, because observers need to have a concept of each of the scale steps, while VAS need only an abstract understanding of the dimension as a whole. In contrast, VAS may depend on visuo-motor coordination, because participants need to translate their experience into spatial coordinates and have to move the joystick accordingly. This might be an explanation for the effects in the current study, although a previous study reported that five scale points cannot convey more information about subjective experiences than four scale points (Ramsøy & Overgaard, 2004). The number of scale steps participants can make use of in labelled scales depends on the participants' ability to categorize their percepts verbally, which might be limited to four. VASs do not depend to the same degree on verbal categorization; therefore, the amount of information transmitted by a VAS can be greater.

4.4.6. Conceptual reasons to prefer VASs or discrete scales

Finally, deciding between VASs and discrete scales is not a question that can be addressed entirely by empirical methods, but needs to be informed conceptually as well. First, a VAS is only feasible if the subjective reports can be given along one dimension. However, the study of visual awareness may require the assessment of several qualitative different patterns of subjective experience: For instance, it has been suggested that observers report "feelings that something has been shown" or "experiences without any content" (Ramsøy & Overgaard, 2004) or even to be confident about the discrimination judgement (Zehetleitner & Rausch, 2013) at low levels of stimulation, and report that they had an experience of a specific stimulus quality only at higher levels of stimulation. These discontinuities in the pattern of subjective reports along the unaware/aware continuum cannot be measured by one single VAS, so other measures are required if the full set of experiences during visual perception is of theoretical interest to a specific study. For example, an established measure that captures qualitatively different experiences is the Perceptual Awareness Scale (Ramsøy & Overgaard, 2004), where participants are asked to differentiate between the absence of an experience, experiences without any content, almost clear experiences of a specific stimulus feature, and full clarity of the specific stimulus feature. Alternatively, different dimensions can be assessed by combining two VASs with different content in one trial (Zehetleitner & Rausch, 2013).

Second, some theorists strongly focus the connection between consciousness and language (Vygotsky, 1962), and such a view might imply verbally categorized scale steps to be more valid than a continuous scale. However, other concepts of consciousness endorse a view where perceptual consciousness is not easily verbalized, and such a view may prefer VASs as they rely less heavily on verbal categorization.

4.5. Conclusion

We present data that both visual analogue scales as well as discrete scales are reliable measures of subjective reports of global motion experience. We found no evidence that the type 2 sensitivity is decreased or the pattern of reports is binary when participants are provided with a large number of scale steps. The data is consistent with the interpretation that participants are able to maintain a sufficient large number of meaningful criteria so that a VAS retrieves a larger amount of information than a discrete scale with four scale steps, provided that participants take their time to make the more subtle judgements. At least when the number of response alternatives of the objective discrimination task is large, subjective reports of motion experience may be recorded more conveniently by a VAS than by a discrete scale with the same content.

4.6. Acknowledgements

This research was supported by the German-Israeli Foundation for Scientific Research and Development (GIF) grant 1130-158 and the Deutsche Forschungsgesellschaft (DFG, i. e. German Research Council) grant ZE 887/3-1 (both to Michael Zehetleitner).

5. TYPE 2 SENSITVITY OF DECISIONAL CONFIDENCE AND VISUAL EXPERIENCE[10]

by Manuel Rausch, Hermann J. Müller, and Michael Zehetleitner[11]

5.1. Abstract

Previous studies provided contradicting results regarding type 2 sensitivity[12] 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 type 2 sensitivity. Comparing logistic regression and meta-d in a masked orientation task, a masked shape task, and a random-dot motion task, we observed type 2 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 type 2 sensitivity. We argue that a science of consciousness would benefit from the assessment of both visual experience and decisional confidence, and recommend meta-d_a as measure of type 2 sensitivity for future studies.

5.2. 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 (Eriksen, 1960; Hannula et al., 2005; Schmidt & Vorberg, 2006).

[10] A version of this chapter has been published as 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

[11] Manuel Rausch conceived the research questions and conducted the analysis; Manuel Rausch, Hermann J. Müller, and Michael Zehetleitner co-wrote the manuscript.

[12] In the version published in Consciousness and Cognition, the terms metacognitive sensitivity and metacognitive bias, synonyms of type 2 sensitivity and type 2 bias, were used throughout the whole text. However, this was changed to maintain consistency of terms throughout the different Chapters of this Thesis.

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; Timmermans et al., 2012). 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). 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 & Naccache, 2001; Dehaene, 2010).

5.2.1. 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, 1885), 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 et al., 2011, 2010; Szczepanowski et al., 2013; Wierzchoń et al., 2012, 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 (see Chapter 2; 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 et al., 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.

5.2.2. 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. 5-1a).

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 et al., 2003). Type 2 tasks allow the assessment of sensitivity and bias just as in type 1 tasks (see Fig. 5-1b): *Type 2 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 (also called metacognitive sensitivity). *Type 2 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 type 2 sensitivity is challenging because type 2 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.

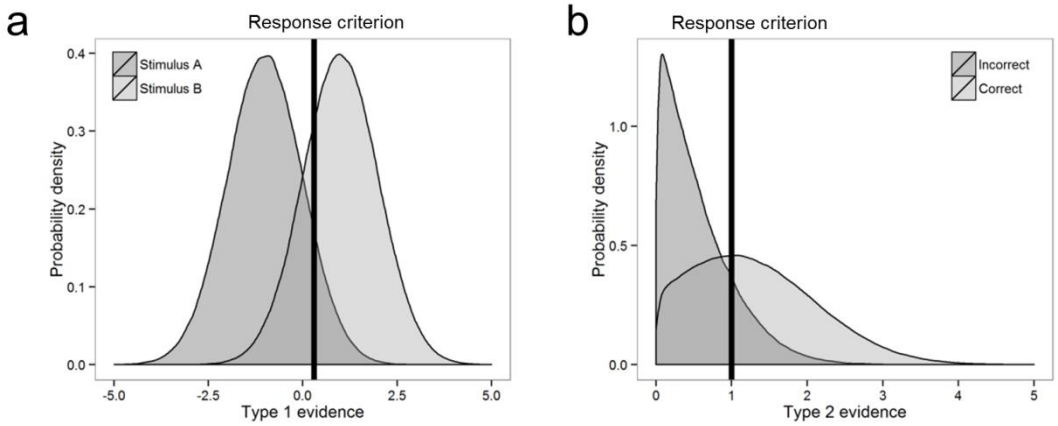

Figure 5-1. Signal detection theory. (a) Distributions of evidence created by the two stimuli A and B in a type 1 task, i.e. the observers' task is to decide which one of the two stimuli has been presented. When the type 1 evidence is greater than the response criterion, observers respond "B", and "A" 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.

5.2.3. Empirical studies on confidence and visual experience

Is there an effect of experience and confidence as content of subjective reports on type 2 sensitivity and bias? Concerning type 2 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 colour-blind, but is still able to make accurate colour discriminations and reports being confident about these colour 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 et al., 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 et al., 2012), a masked orientation discrimination task, a masked shape

discrimination task, and a random-dot motion discrimination task (see Chapter 2 and 3; Zehetleitner & Rausch, 2013).

For type 2 sensitivity, the evidence for a distinction between experience and confidence is less consistent. The only neuropsychological study informative of type 2 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 (see sections 2.3 and 2.7; Zehetleitner & Rausch, 2013); and no substantial differences were found in a masked discrimination task of affective face expressions (Sandberg et al., 2013; Szczepanowski et al., 2013) and a masked shape discrimination task (section 2.3; 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 type 2 sensitivity. One candidate factor may be the *method used to quantify type 2 sensitivity*: Those two studies that found type 2 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 type 2 sensitivity (Fleming et al., 2010), observed that type 2 sensitivity of confidence was substantially greater than type 2 sensitivity of experience or at least confidence tended to be associated with a greater type 2 sensitivity. Since the measure of type 2 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.

5.2.4. Meta-d_a as measure of type 2 sensitivity

The development of meta-d_a, a relatively new approach to quantifying type 2 sensitivity (Maniscalco & Lau, 2012), offers the possibility assess meta-d_a with improved control (Fleming & Lau, 2014). The conceptual idea of meta-d_a is to express type 2 sensitivity

in terms of sensitivity of a type 1 SDT model (see Fig. 5-2). 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_a indicates the distance between the two distributions of evidence available for subjective responses. If meta-d_a is smaller than d_a, 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, type 2 sensitivity is suboptimal. An introduction into the mathematics of meta-d_a is provided by (Barrett et al., 2013).

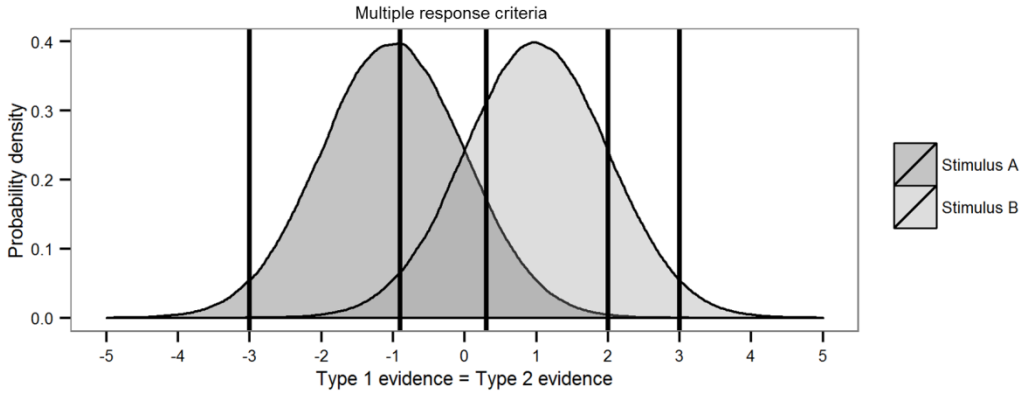

Figure 5-2. Signal detection model underlying meta-d_a. Meta-d_a is computed assuming metacognition is ideal, i.e. the same evidence is available for subjective reports than for discrimination judgments. The model is the same as a standard SDT model for a type 1 task, except that discrimination decisions 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.

Meta-d_a 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_a requires making assumptions about the

shape of these distributions, which may be incorrect (Fleming & Lau, 2014). On the other hand, meta- d_a provides two advantages over type 2 ROC analysis: First, meta-d_a accounts for bias regarding the two task alternatives. Second, meta-d_a can be used to easily compare type 2 sensitivity to objective task performance because meta-d_a is expressed in the same signal-to-noise units as the standard d_a 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_a.

5.2.5. Logistic regression as measure of type 2 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 (Sandberg et al., 2013, 2010; Wierzchoń et al., 2012, 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 type 2 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. 5-3). 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 type 2 sensitivity.

On the one hand, logistic regression provides several advantages over other methods to analyse 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., 2009; Pinheiro & Bates, 2000). Second, logistic mixed-model regression can be applied when the data is unbalanced (Bolker et al., 2009), 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.

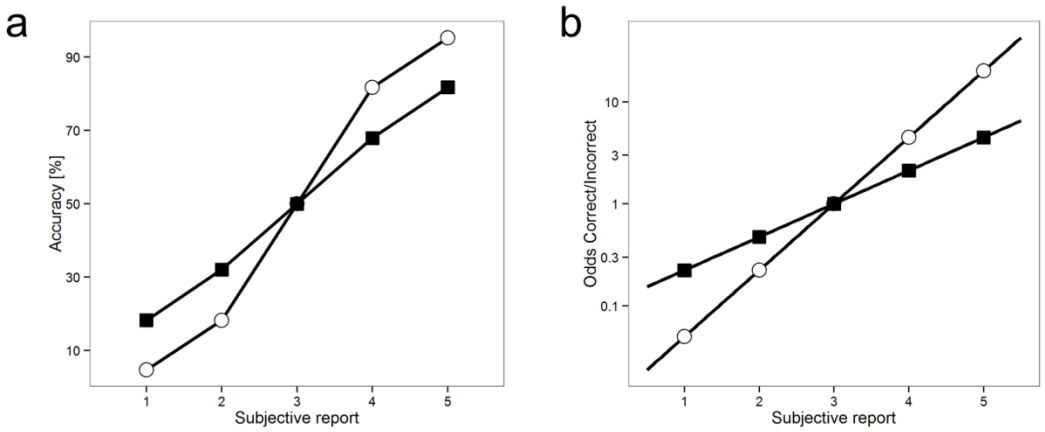

Figure 5-3. Quantifying the relationship between trial accuracy and subjective reports by logistic regression. (a) Data of a hypothetical experiment. Task accuracy in % correct is plotted as a function of subjective report. Lines indicate two separate conditions. (b) Same data with accuracy transformed 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.

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 & Brockhoff, 2013). In contrast, ratings on a visual analogue 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 -∞ 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 type 2 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 type 2 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.

5.2.6. Rationale of the present study

In present paper, we investigate two issues: First, we examined whether an analysis of meta-d_a and logistic regression would reveal the same effect of visual experience versus decisional confidence (as contents of subjective reports) on type 2 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 type 2 sensitivity, is justified.

Specifically, we predicted that if the method of assessing type 2 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 type 2 sensitivity of confidence should be greater no matter if quantified by meta-d_a 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 (three experiments of Chapter 2; 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 analysed type 2 sensitivity based on logistic regression analysis as well as meta-d_a.

5.3. Reanalysis

5.3.1. Material and Methods

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

5.3.1.1. 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. 5-4). 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 labelled 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. 5-4). 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.

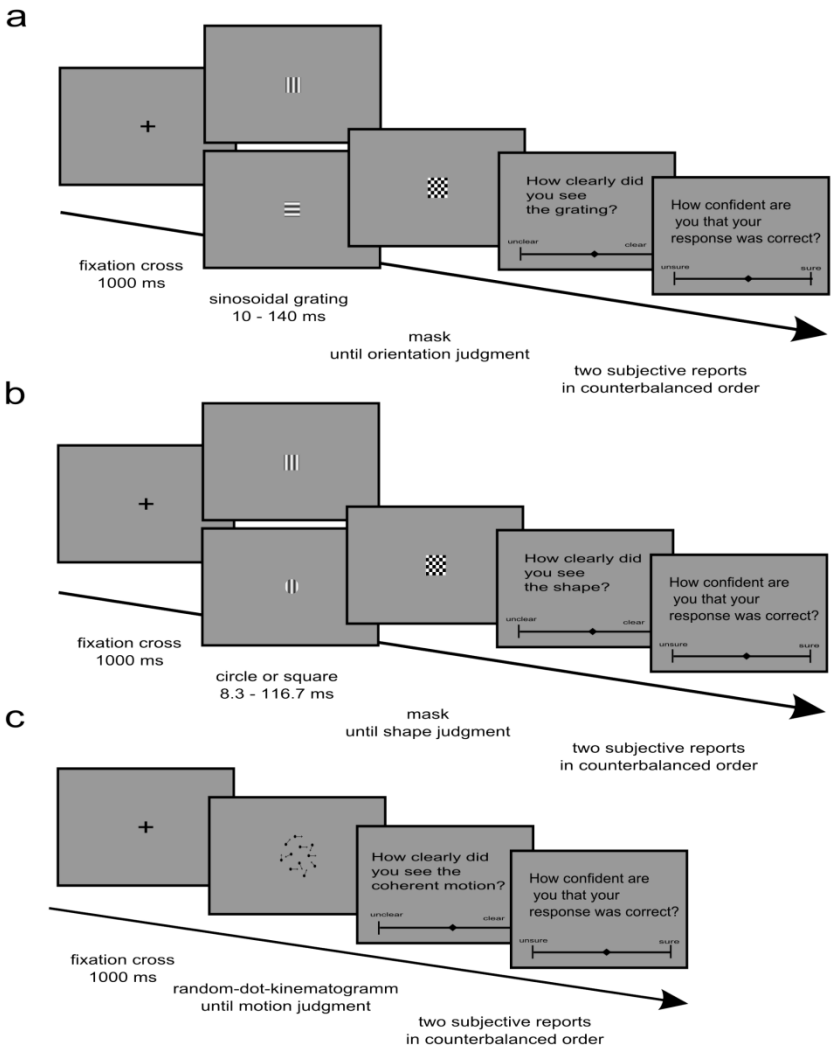

Figure 5-4. Trial sequence for (a) the masked orientation task, (b) the masked shape discrimination task, and (c) the random-dot motion discrimination task.

5.3.1.2. Analysis

All analysis were conducted in the free software R 3.0.2 (R Core Team, 2014). Trials of the masked orientation task on which participants did not report their subjective confidence in being correct were excluded from the analysis.

5.3.1.2.1. Logistic regression

Logistic mixed regression analysis was performed using the R library lme4 (Bates, Mächler, Bolker, & Walker, 2014; Bates, Maechler, et al., 2014), 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 centred 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.

5.3.1.2.2. Meta-d_a

Meta-d_a 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 section 5.7.), assuming normal distributions of evidence with non-equal variances. First, the continuous VAS rating data was divided into 13 equal bins. Then, meta-d_a 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 et al., 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.

5.3.1.2.3. 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.

5.3.2. Results

5.3.2.1. Logistic regression

The complete results of the mixed logistic regression models can be seen in Table 5-1. 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.

Table 5-1

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[13]	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

[13] 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.

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.

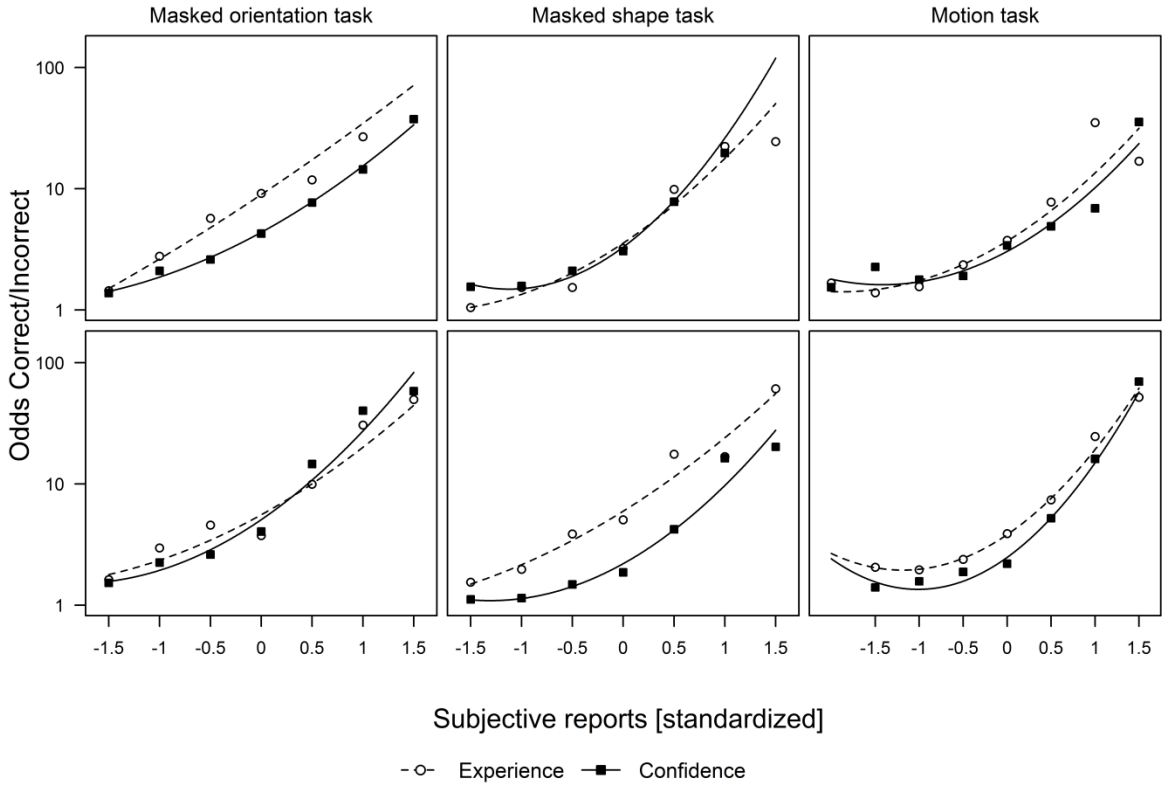

Figure 5-5. Relationship between subjective report and the odds of being correct, separately for scale, experiment, and time of the report. Upper row: First subjective report within one trial, Lower row: Second subjective report within one trial. Lines indicate the prediction from logistic regression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

The relationships between subjective report and transformed accuracy are depicted separately for scale, experiment, and time of the report in Fig. 5-5. 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.

5.3.2.2. Meta-d_a

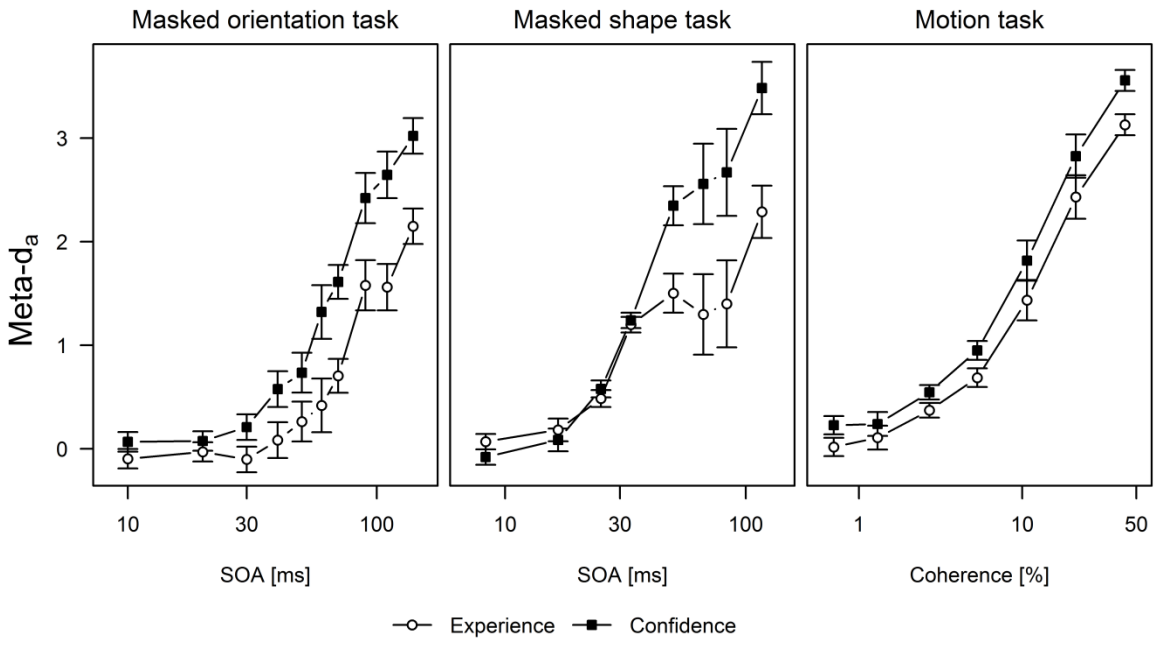

Figure 5-6. Meta-d_a as a function of stimulus quality, separately for each task in separate panels and scales as separate lines.

As can be seen from Fig. 5-6, meta-d_a scores estimated from confidence ratings were greater than meta-d_a 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_a of experience was still at chance level; in the masked shape task, by contrast, the effect became evident only at longer SOAs.

The results of the mixed linear regression models can be seen in Table 5-2. We found substantial negative effects of scale in all three experiments, indicating that meta-d_a scores computed from visual experience were indeed always smaller than meta-d_a 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.

Table 5-2

Results of a linear mixed regression model for meta-d_a 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.

Post-hoc tests comparing meta-d_a between experience and confidence separately at each SOA revealed that for the orientation task, meta-d_a 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_a of confidence to be above meta-d_a of experience at the SOA of 50 ms, $t(15) = 3.7$, $p_{cor} < .05$, as well as the SOA of 116.7 ms, $t(15) = 5.0$, $p_{cor} < .01$.

5.3.2.3. 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.

5.4. 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 type 2 sensitivity depend on the method used to quantify type 2 sensitivity?

(ii) Is logistic regression biased owing to a non-linear relationship between transformed accuracy and subjective reports?

Concerning the effect of content, meta-d_a indicated that type 2 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 type 2 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 type 2 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.

5.4.1. Why confidence outperforms experience in predicting 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[14]
(ii) distinct metacognitive mechanisms (Overgaard & Sandberg, 2012),
(iii) placement of different criteria (Wierzchoń et al., 2012, 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 et al., 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 type 2 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

[14] This argument can not only be framed in terms of conscious access, but also in terms of higher order thought theory, see section 6.1.3.

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 type 2 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, 2014). If the differences between scales were only due to type 2 bias, rather than type 2 sensitivity, there should be no effect of different scales if subjective criteria are controlled for. However, we find meta-d_a of confidence to be greater than meta-d_a of experience across all three experiments, indicating that the difference between experience and confidence is not due to type 2 bias alone.

5.4.2. 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 type 2 sensitivity: While type 2 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_a as measures of type 2 sensitivity revealed that the overall pattern of type 2 sensitivity of confidence compared to experience was largely independent of the method used to assess type 2 sensitivity. Consequently, the question what factors determine whether subjective reports of experience or confidence are associated with greater type 2 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 type 2 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 greater type 2 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 type 2 sensitivity of motion experience (Chapter 4, 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 type 2 sensitivity of visual experience could be driven entirely by experiences without content (see Chapter 2 and 4 for more detailed discussions; Rausch & Zehetleitner, 2014; Zehetleitner & Rausch, 2013). Finally, although logistic regression and meta-d_a 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 type 2 sensitivity of experience and confidence.

5.4.3. How should we quantify type 2 sensitivity?

Comparisons between previous studies on type 2 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_a (Maniscalco & Lau, 2012). Our reanalysis based on logistic regression and meta-d_a 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_a 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 type 2 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 type 2 sensitivity (Barrett et al., 2013), and given that meta-d_a also controls discrimination bias and may provide increased statistical power, we recommend meta-d_a for all future studies where it can be applied.

5.5. Conclusion

We report that logistic regression and meta-d_a 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 type 2 sensitivity, which is why we recommend meta-d_a for future studies.

5.6. Acknowledgements

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 Michael Zehetleitner).

5.7. Appendix: Code to compute meta-d_a in R

```
### Code to compute meta-d' in R ####
#################################

# computeMetaD computes meta-d' based method described by 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

# Arguments:
#   ratings: a factor with levels corresponding to the rating categories,
#          ordered from low to high
#   stimulus: a factor, levels corresponding to stimulus identities
#   correct: a vector with 1 indicating correct responses and 0 incorrect responses
#   distr: What distributions of the evidence should be assumed. Default "norm",
#        uses the normal distribution,
#        "logis" assumes the logistic distribution, and "gumbel" the distribution of smallest extremes
#   contraintsMetaD: a two-element vector with minimal and maximal value allowed for meta-d'
#   varEqual: a logical value indicating if equal variances should be assumed
#   addConstant:  a logical value indicating if a constant of .5 devided by the number of rating
#              categories should be added to each cell
#   nInnerIterations, nOuterIterations: number of inner and outer iterations
#                                passed to contrOptim
#
# Returns: a list with the elements
#   metaDPrime: estimated sensitivity from rating data
#   dPrime: estimated sensitivity from objective discrimination responses
#   logLikelihood: log of the likelihood of the best fit

computeMetaD <- function(ratings, stimulus, correct,
                  distr = "norm",
                  constrainMetaD = c(-20,20),
                  varEqual = FALSE, addConstant = TRUE,
                  nInnerIterations = 1000, nOuterIterations = 1000){

 if(!is.factor(ratings)) stop ("ratings should be a factor!")
 if(!is.factor(stimulus )|| length(levels(stimulus)) != 2) {
  stop("stimulus should be a factor with 2 levels")
 }
 if(!all(correct %in% c(0,1))) stop("correct should be 1 or 0")

 pfun <- switch(distr, norm = pnorm, logis = plogis,
            gumbel = function(x, location, scale) exp(-exp((location-x)/scale)))
 qfun <- switch(distr, norm = qnorm, logis = qlogis,
            gumbel = function(x) -log(-log(x)))
```

```
nRatings <-  length(levels(ratings))
 nCriteria <- nRatings * 2 - 1
 abs_corrects <-  table(ratings[correct == 1], stimulus[correct == 1])
 abs_errors   <-  table(ratings[correct == 0], stimulus[correct == 0])

 if (addConstant){
  abs_corrects <- abs_corrects + .5/nRatings
  abs_errors   <- abs_errors + .5/nRatings
 }

 nC_rS1 <- rev(as.vector(abs_corrects[, 1]))
 nI_rS1 <- rev(as.vector(abs_errors[, 2]))
 nC_rS2 <- as.vector(abs_corrects[, 2])
 nI_rS2 <- as.vector(abs_errors[, 1])

 abs_S1 <- c(rev(abs_errors[,2]),abs_corrects[,2])
 ratingHrs <- qfun(1 - cumsum(abs_S1)/sum(abs_S1))
 abs_S2 <-  c(rev(abs_corrects[,1]), abs_errors[,1] )
 ratingFrs <-  qfun(1 - cumsum(abs_S2)/sum(abs_S2))
 finits <- is.finite(ratingHrs) & is.finite(ratingFrs)
 ratingHrs <- as.vector(ratingHrs[finits])
 ratingFrs <- as.vector(ratingFrs[finits])

 if (varEqual) { s <- 1
 } else s <- as.vector(lm(ratingHrs ~ ratingFrs)$coefficients[2])

 meta_d1 <- (1/s) * ratingHrs[nRatings] - ratingFrs[nRatings]
 cs_1 <- (-1/(1+s)) * (ratingHrs + ratingFrs)
 initials <- c(meta_d1, cs_1)

 A <- matrix(0, nrow=nCriteria+1, ncol = nCriteria+1)
 A[1, 1] <- 1
 A[2, 1] <- -1
 diag(A[3:(nCriteria+1),2:nCriteria]) <- -1
 diag(A[3:(nCriteria+1),3:(nCriteria+1)]) <- 1
 b <- c(min(constrainMetaD),-1* max( constrainMetaD),rep(0, nCriteria-1))

 fit <- constrOptim(theta = initials, f = negLogLik, grad = NULL, ui = A, ci = b,
                outer.iterations = nOuterIterations,
                control = list(maxit = nInnerIterations),
                nC_rS1 = nC_rS1, nI_rS1 = nI_rS1, nC_rS2 = nC_rS2, nI_rS2 = nI_rS2,
                nRatings = nRatings, s = s, pfun = pfun)

 result <- list(dPrime = meta_d1 * s * sqrt(2/(1 + s^2)),
            metaDPrime = fit$par[1] * s * sqrt(2/(1 + s^2)),
            logLikelihood = -fit$value)

 return(result)
}
```

```
negLogLik <- function(parameters, nC_rS1, nI_rS1, nC_rS2, nI_rS2, nRatings, s, pfun) {
  t1c <- parameters[nRatings+1]
  S1mu <- -parameters[1]/2
  S1sd <- 1
  S2mu <- parameters[1]/2
  S2sd <- 1/s
  t2c1x <-  c(-Inf, parameters[2:length(parameters)],Inf)
  prC_rS1 <- (pfun(t2c1x[2:(nRatings+1)],S1mu,S1sd) -
           pfun(t2c1x[1:nRatings],S1mu,S1sd)) / pfun(t1c,S1mu,S1sd)
  prI_rS1 <- (pfun(t2c1x[2:(nRatings+1)],S2mu,S2sd) -
           pfun(t2c1x[1:nRatings],S2mu,S2sd)) / pfun(t1c,S2mu,S2sd)
  prC_rS2 <- ((1- pfun(t2c1x[(nRatings+1):(nRatings*2)],S2mu,S2sd)) -
           (1- pfun(t2c1x[(nRatings+2):(nRatings*2+1)],S2mu,S2sd))) /
           (1 - pfun(t1c,S2mu,S2sd))
  prI_rS2 <- ((1- pfun(t2c1x[(nRatings+1):(nRatings*2)],S1mu,S1sd)) -
           (1- pfun(t2c1x[(nRatings+2):(nRatings*2+1)],S1mu,S1sd))) /
           (1 - pfun(t1c,S1mu,S1sd))

  logL <-  -sum(nC_rS1*log(prC_rS1),nI_rS1*log(prI_rS1),nC_rS2*log(prC_rS2),nI_rS2*log(prI_rS2))

  return(logL)
}


### Examples

stimulus <- factor(rep(c("A", "B"), 10))
ratings <- factor(c(rep(1:3,4), rep(3,times=8)))
correct <- c(rep(0,4), rep(1,6), rep(0,2), rep(1,8))

computeMetaD(stimulus = stimulus, ratings = ratings, correct = correct)
computeMetaD(stimulus = stimulus, ratings = ratings, correct = correct, distr = "logis")
computeMetaD(stimulus = stimulus, ratings = ratings, correct = correct, distr = "gumbel")


computeMetaD(stimulus = stimulus, ratings = ratings, correct = correct,
           constrainMetaD = c(0,5))
computeMetaD(stimulus = stimulus, ratings = ratings, correct = correct,
           varEqual = TRUE)
computeMetaD(stimulus = stimulus, ratings = ratings, correct = correct,
           addConstant = FALSE)
```

6. FINAL DISCUSSION

The present series of experiments addressed the research question how subjective measures of conscious awareness need to be designed to provide valid and reliable data for consciousness research.

Regarding the content of subjective measures, a series of psychophysical experiments suggested that subjective measures about the accuracy of a discrimination response are different from measures about visual experience: First, measures related to the discrimination response were found to impose more liberal criteria. Second, they were associated with greater type 2 sensitivity. Third, different subjective measures about the accuracy of a discrimination response correlated more strongly with each other than each of them correlated with a subjective measure of visual experience (Chapter 2 and 5; Zehetleitner & Rausch, 2013; Rausch, Müller, & Zehetleitner, 2015). Finally, the earliest sensory ERP correlates of verbal reports were predictive of the fact whether participants reported that they made a discrimination response based on knowledge rather than guessing, but were not yet predictive whether participants reported a clear experience over and above that knowledge. The strongest correlate of visual experience closely preceded participants' response to the discrimination task (Chapter 3).

With respect to granularity, subjective measures of the experience of motion contained more information when participants selected a position on a visual analogue scale compared to a scale with four discrete labelled categories. The greater amount of information rendered subjective measures more predictive of task accuracy and improved coefficients of internal consistency. In addition, there was no evidence that participants' type 2 sensitivity was impaired by the greater number of response options offered by a visual analogue scale (Chapter 4, Rausch & Zehetleitner, 2014)

Finally, regarding the statistical procedure to quantify the relation between subjective measures and task accuracy, logistic regression was found to be a suboptimal method due to non-linear relationships between subjective reports and the transformed task accuracy. However, meta-d_a, a measure derived from signal detection theory, provided the most consistent results across studies (Chapter 5, Rausch, Müller, & Zehetleitner, 2015).

6.1. The content of subjective measures

It is widely assumed in the field of consciousness research subjective reports of visual experience and decisional confidence are equally valid, and thus typically used and/or interpreted as interchangeable (Ko & Lau, 2012; Lau & Rosenthal, 2011; Seth et al., 2008). The present data challenges this wide-spread assumption: Participants first report confidence about being correct in a discrimination task; only at greater strength of stimulation, they would report to experience the stimulus visually.

Is a dissociation between stimulus-related and response-related subjective measures consistent with the existing theories of consciousness? Although the advocates of phenomenal consciousness, global workspace theory, and higher-order theories have not discussed potential disagreements between subjective measures, it should be considered if such a distinction can be easily integrated in each of these frameworks, and if not, which assumptions need to be adjusted.

6.1.1. Theoretical implications for phenomenal consciousness

In the framework of phenomenal consciousness, the distinction between stimulus-related and response-related measures may be explained as participants experiencing an intuition of being correct in the task without experiencing the stimulus visually. Phenomenal consciousness involves the experiential properties of sensations, feelings, perceptions, thoughts, wants, and emotions (Block, 2002). These experiences can be assorted into two categories (Bischof, 1965): One set of phenomenology appears to arise from the observer's own mind, e.g. thoughts and emotions, and may be called *ostensibly mental* experiences. *Ostensibly physical* experiences in contrast appear to stem from the physical world, for instance from an external stimulus, or the observer's body. One possibility is that the phenomenology of participants that report confidence in the task but no experience of the stimulus is the ostensibly mental feeling-of-knowing. Participants have the experience that they know what the stimulus feature is, but the stimulus does not create visual phenomenology. Feeling-of-knowing has been originally described in the context of meta-memory (Koriat, 2007; Nelson & Narens, 1990), but visual perception may be able to generate feelings-of-knowing as well (Mangan, 2001). A second possibility is that the experience is ostensibly physical. So-called blindsight patients sometimes report residual phenomenology characterized by the awareness of the event, but without the phenomenology

of normal seeing (Sahraie et al., 2002; Zeki & Ffytche, 1998). Normal observers may have a similar experience if the stimuli are just at the threshold of conscious perception (Ramsøy & Overgaard, 2004).

While on a principal level, the concept of phenomenal consciousness has sufficient degrees of freedom to describe the distinction between stimulus-related and response-related subjective measures, there is one specific hypothesis about phenomenal consciousness that appears at first glance to be at odds with the current data. According to the overflow hypothesis, the contents of short-term sensory buffers are associated with phenomenal experience (Block, 2011). This short-term sensory buffer stores all visual objects for a short period of time, until it is overridden by the next stimulation. However, participants are only able to make correct discrimination judgments about 4 ± 1 objects, as cognitive access to the contents of sensory buffers is limited by the capacity of working memory (Sligte, Scholte, & Lamme, 2008; Vandenbroucke et al., 2011). As the capacity of the conscious sensory buffer is much larger than the capacity of working memory, the overflow hypothesis explains why participants are only able to make correct task responses about a small number of display items, although they report an experience of a rich phenomenal world (Block, 2011).

The standard and widely debated case is that phenomenal consciousness exceeds cognitive access. However, in present data, the relation between visual consciousness and access appears to be the other way round. Participants report to be confident more often than they report visual experience. Conscious access is a requirement for all subjective measures, as they require that neural systems engaged in decision making and language need to receive inputs from perceptual processes. However, although participants had conscious access when they reported they felt confident about task response, they reported no conscious visual experience. A similar pattern was reported by a patient suffering from achromatopsia (Carota & Calabrese, 2013): After bilateral temporal-occipital lesions, that patient reports to be colour-blind although he performs accurately in a colour recognition task. A potential explanation why the relation between phenomenal consciousness and conscious access varies is the number of items in the display: Visual short-term memory always used arrays of multiple stimuli. In contrast, the present studies always presented one stimulus at the screen at fixation. Consequently, phenomenal conscious may overflow cognitive access only for stimuli outside of the focus of attention, while at the focus of attention, conscious access occurs more frequently than phenomenal experience.

6.1.2. Theoretical implications for global workspace theory

Can the global workspace theory account for the distinction between stimulus and response-related subjective measures? An important flavour of global workspace theory conceives consciousness as an all-or-nothing phenomenon (Dehaene et al., 2003): Conscious access depends on a cerebral "ignition" where neural activity spreads from sensory cortical areas to frontal and parietal areas, making perceptual contents available to multiple cognitive functions. If global ignition does not occur, perceptual contents are not available for report (Dehaene et al., 2006). As this most radical form of global workspace theory permits only two states - either the observer is fully conscious of the stimulus or the observer is unconscious - global workspace theory predicts that all measures depending on behaviour that requires global access are generally in good agreement (Dehaene & Changeux, 2011). A dichotomous model of conscious awareness appears unable to explain the intermediate state where participants report some confidence in being correct in the discrimination task about the stimulus, but they report no experience of the stimulus (Carota & Calabrese, 2013; Charles et al., 2013; Sahraie et al., 1998; Schlagbauer et al., 2012; Zehetleitner & Rausch, 2013).

However, more complex flavours of global workspace theory are better able to accommodate the experience/confidence distinction. According to the partial awareness hypothesis (Kouider et al., 2010), a stimulus is represented by a hierarchy of features, with low-level features at the bottom and increasingly complex features at the top. Separate features can be consciously accessed independently from the other features. Partial awareness is a state where some of the features of a stimulus are consciously accessible but others features are missing. If participants are in a state of partial awareness, conscious access to the task-relevant stimulus feature may be sufficient for reporting confidence about the discrimination response (Dienes, 2004, 2008). In contrast, a report about visual experience may require conscious access to a greater number of stimulus features. After all, a striking feature of consciousness is the so-called unity of experiences. Humans do not experience colour, shape, location, etc. of a stimulus as separate. Instead, the conscious experience of the separate features of a stimulus seems to be integrated to one visual object (Bayne & Chalmers, 2003). The partial awareness hypothesis elegantly explains further details of the present results: If task-irrelevant stimulus features contribute exclusively or to a greater degree to subjective measures about the stimulus, they would not as efficient in predicting trial accuracy as subjective measures about task accuracy. However, as task-relevant and task-

irrelevant features depend on stimulus quality, the correlation between stimulus quality and subjective measures of both contents would be the same. Finally, involvement of different sets of features is consistent with different neural correlates during sensory processing.

A second possibility to incorporate the dissociation between stimulus and response-related subjective measures in the global workspace framework is in terms of unconscious evidence accumulation. According to this model, subliminal stimuli possess sufficient energy to evoke a feed-forward wave of activation in specialized processors, but insufficient energy to trigger global neural activity necessary for conscious access (Dehaene et al., 2006; Dehaene, 2010). This unconscious feedforward sweep may even reach higher areas, thereby leading to above chance performance as well as error detection in the absence of consciousness (Charles et al., 2014, 2013). If response-related subjective measures are in parts generated by unconscious evidence accumulation, it may account for dissociations between subjective measures as well. However, two aspects of the data fit better to the partial awareness hypothesis: First, the number of different judgments participants performed in absence of reported visual experience makes it plausible that certain stimulus characteristics were globally available and consciously accessed. Participants were able to discriminate the stimulus above chance, to estimate their confidence in having made a correct discrimination response, to wager imaginary money on the correctness of the discrimination response, and to attribute the reason for their choice on guessing or on knowledge. Moreover, subjective measures always imply some involvement of language areas, since participants need to align the labels of the scale steps with their degree of confidence in being correct. This variety of different judgments in absence of reported visual experience implies that the neural activation is not encapsulated in a single specialized processing module. Nevertheless, it cannot be ruled out that the unconscious evidence is encapsulated in multiple specialized modules without triggering a global workspace. However, this model implies that a great number of judgments can be performed without involvement of the global workspace, raising the question if there are any judgments at all for which the global workspace is necessary. Second, the timing of the ERP correlates of subjective and response-related subjective measures are not consistent with the hypothesis that confidence without reported experience is driven by feedforward processing with lower energy: Lower stimulus energy typically delays the most prominent ERP correlate, i.e. the mid-range negativity (Railo et al., 2011). However, the experiment of Chapter 3 suggested that the correlates specific to decisional confidence preceded the correlates of visual experience in the present study. Moreover, an ERP index of global

availability, the late positivity, was associated with both visual experience and confidence (although the correlation of visual experience and the late positivity was admittedly particularly strong).

6.1.3. Theoretical implications for higher-order theories of consciousness

In the framework of higher-order theories of consciousness (Carruthers, 2011; Lau & Rosenthal, 2011; Timmermans, Schilbach, Pasquali, & Cleeremans, 2012), stimulus-related subjective measures indicate whether participants possess a higher order mental state about the stimulus, while response-related subjective measures demonstrate higher order mental states about task accuracy. Consequently, the dissociation between stimulus-related and response-related subjective measures indicates that there can be higher-order states about the response without higher-order states about perception. Participants know that they respond correctly, but they do not know that they have seen the stimulus.

Are higher-order states about task responses without higher-order states about the stimulus consistent with higher order theories of consciousness? At first glance, it may seem they are not: Higher order theories strongly emphasize the link between consciousness and metacognition (Rosenthal, 2000). Under the assumption that stimulus-related measures do not depend on metacognition, confidence in decision making without experience of the stimulus undermine the strong link between consciousness and metacognition and thus one of the core tenets of higher-order theories (Charles et al., 2014, 2013). However, the assumption that stimulus-related measures do not depend on metacognition is controversial. First, it can be argued that both stimulus- and response-related measures both require metacognitive processes with the only difference that subjective measures about the stimulus require metacognition of stimulus perception, not task performance (see 2.2.3; Zehetleitner & Rausch, 2013). Consistent with this, stimulus-related measures are associated with neural activity in dorsolateral prefrontal cortex (dlPFC) – a brain region closely related to metacognition (Fleming & Dolan, 2012; Fleming et al., 2010) – as suggested by functional magnetic resonance imaging (Lau & Passingham, 2006) and theta-burst transcranial magnetic stimulation (Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010). Second, there is evidence that the decision processes that selects the response to the discrimination task is involved making a report on a stimulus-related measure, too: Both stimulus-related and response-related subjective measures can be modulated by the discrimination decision irrespective of the time the judgment is made (Wierzchoń et al., 2014). In addition, the

present EEG study suggested that ERP correlates of stimulus-related measures are most pronounced around the time the decision is made. Overall, the position that confidence in absence of experience indicates consciousness without metacognition seems hard to defend.

What mechanism can account for the occurrence of higher-order states about the accuracy of task responses without higher-order states about perception? The first two proposals both assume that response-related subjective measures require additional metacognitive processes. First, the metacognitive process specific to response-related subjective measures may be an unconscious error monitoring system (Charles et al., 2014, 2013). This error monitoring process could be informed by perceptual processes too weak to trigger higher-order thoughts about the stimulus, thus explaining why participants report confidence in being correct but do not report a visual experience of the stimulus. The present data is fully compatible with error monitoring exclusively involved in response-related measures. However, the timing of the neural correlates of stimulus- and response-related measures begin to diverge already during sensory processing, suggesting that at least some differential features of stimulus- and response-related measures arise earlier than metacognitive processes occur.

The second proposal suggests that stimulus-related measures are generated by a simple metacognitive process of monitoring one's experience. Response-related subjective measures are thought to stem from a more complex metacognitive process that relates the output of the first metacognitive process to one's accuracy in the task (Overgaard & Sandberg, 2012). This theory predicts that higher order thoughts about the task response are conditioned on higher-order thoughts about the stimulus. While this view was developed to explain the data of a previous study (Sandberg et al., 2010), the present data is not consistent with this view. In the present data, response-related measures are not conditioned on stimulus-related subjective measures; in all four experiments, response-related measures were associated with more liberal thresholds for report than stimulus-related measures. In addition, the predictive power of response-related subjective measures was more efficient in terms of discrimination task correctness as well as early ERP correlates, suggesting that response-related measures do not require a more complex judgment than stimulus-related measures do.

The final possibility is a variant of the partial awareness hypothesis framed within a higher order framework (see above): Partial awareness is a state where some features of the stimulus are globally accessible while others remain inaccessible (Kouider et al., 2010).

Global access allows a great variety of cognitive systems to make use of the perceptual information (Baars, 2002; Dehaene & Naccache, 2001). Metacognition could be one of the cognitive functions that depend on global access. Consequently, in a state of partial awareness where the task-relevant feature is accessible, participants may be able to form a higher order thought about task accuracy and thus report being confident. However, other features of the stimulus may be inaccessible and so participants lack a higher order thought about the stimulus and report no experience accordingly. Such a higher-order framing of the partial awareness hypothesis has the same explanatory power as the original partial awareness hypothesis with respect to psychophysical thresholds, type 2 sensitivity, correlations between subjective measures and stimulus quality, and timing of ERP correlates (see 6.1.2).

Overall, higher-order theories can be reconciled with dissociations between stimulus- and response-related measures if one accepts (i) that stimulus-related measures are dependent on metacognition, and (ii) that higher-order mental states about the task response are not conditioned on higher-order mental states about the stimulus.

6.1.4. Methodological implications

The methodological implications of the distinction between stimulus-related and response-related subjective measures are straight-forward: As stimulus-related and response-related measures are associated with different behaviours in visual psychophysics and distinct ERP correlates, both categories of subjective measures should no longer be treated as interchangeable (as previously argued by Charles et al., 2013; Rausch, Müller, & Zehetleitner, 2015; Sahraie et al., 1998; Schlagbauer et al., 2012; Zehetleitner & Rausch, 2013). First, as a consequence for future studies, researchers need to consider more carefully which conscious contents are relevant for their specific research question, and choose the content of their subjective measure accordingly. Some studies investigate the visual experience of a specific stimulus feature, e.g. studies measuring the neural correlate of experiencing “red” when seeing a red apple. In this case, participants should report their conscious experience of this particular feature. If participants were asked about their confidence in a task instead, there would be a risk that participants had just an intuition of being correct without visual experience, resulting in false positives. In contrast, if a study is about all conscious contents underlying performance in a specific task, participants should make a report that refers to the task response as the use of stimulus-related measures may lead to misses in this case (Dienes, 2004, 2008). If all conscious contents are relevant to a specific study, researchers may want to

consider if it is feasible to use both a stimulus-related and a response-related subjective measure. Second, for the interpretation of earlier studies of the NCC, it should be carefully considered if their results constitute correlates of stimulus-related measures, of response-related measures, or shared correlates of the two. Unfortunately, studies on the NCC are often not explicit about the precise content of the rating scale used as measure of conscious awareness. Based on the results of the present work, subjective measures about accuracy of task responses can be expected to favour correlates in earlier time ranges, while subjective measures about visual experience may reinforce comparably late neural correlates.

A series of previous research has compared different subjective measures with the objective of empirically identifying the "best" scale to measure conscious experience (Dienes & Seth, 2010; Rausch & Zehetleitner, 2014; Sandberg et al., 2010; Szczepanowski et al., 2013; Wierzchoń et al., 2012, 2014). This research program rests on the assumption that subjective measures under comparison are equally valid from a conceptual point of view, but some scales correlate more strongly with task performance and thus should be used as measure of conscious awareness (see 1.5.1.; Rausch, Müller, & Zehetleitner, 2015; Rausch & Zehetleitner, 2014). In the light of the present work, empirical studies identifying the best scale should come only after conceptual considerations what the relevant conscious contents are for a specific research question. When the conscious contents have been determined based on the research question, empirical studies are important to optimize other properties of the scale, e.g. the number of response options.

6.2. The granularity of subjective measures

6.2.1. Theoretical implications

The granularity of conscious awareness is of great theoretical interest because some theories make specific predictions whether conscious awareness is gradual or dichotomous. On the one hand, neural global workspace theory predicts that consciousness is an all-or-nothing phenomenon because it depends on a global ignition of spreading neural activity over widely distributed brain areas (Dehaene et al., 2003, 2006). Consequently, U-shaped distributions of subjective reports were considered as evidence for the global workspace theory. Such a sharp transition between unconscious and conscious perception was observed in word detection tasks during the attentional blink (Nieuwenhuis & de Kleijn, 2011; Sergent & Dehaene, 2004) as well as masked number discrimination tasks (Del Cul et al., 2007;

Windey, Gevers, & Cleeremans, 2013). On the other hand, according to the radical plasticity thesis, a variant of higher-order theories, conscious awareness depends on metarepresentations of gradually varying signal strength (Cleeremans, 2008, 2011). Consistent with this view, others have reported gradual transition between unconscious and conscious perception in masked shape discrimination tasks (Sandberg et al., 2011, 2010), masked colour discrimination task (Windey et al., 2013), or even in character identification during the attentional blink (Nieuwenhuis & de Kleijn, 2011). The present study adds random dot kinematograms to the list of stimuli where subjective reports of visual experience follow a more gradual trend (see Chapter 4, Rausch & Zehetleitner, 2014).

How can global workspace theory account for reports of gradually varying conscious experience? An explanation is again provided by the partial awareness hypothesis (Kouider et al., 2010): A stimulus is represented by an assemblage of different features, which can be consciously accessed independently of each other. As proposed by the global workspace model, conscious access to each single feature is all-or-none. The representation of the whole stimulus can be more or less complete, thus creating stronger and weaker experiences of the stimulus. However, the present data is not consistent with the all-or-none predictions of this model. First, the distribution analysis of discrimination responses suggested that among the trials where performance was not at chance, the precision of discrimination judgments still varied as a function of stimulus quality, suggesting that performance in the discrimination task was not binary. Second, subjective reports of motion experience recorded on a visual analogue scale were more predictive of the discrimination error than those reports of motion experience recorded on a four discrete category scale. If conscious awareness of the task-relevant stimulus feature varied in a dichotomous manner, measuring conscious experience of this feature with increased granularity should not increase the correlation with task performance since two distinct states can be represented by both scales without loss of information. As the correlation with the discrimination error increases with the granularity of the scale, it means that the scale is able to pick up more than two levels of the quality of experience of the task-relevant stimulus feature. Overall, the present studies are not consistent with all theories that predict that consciousness of all stimulus features is necessarily dichotomous.

As a consequence of the evidence for gradual conscious perception, a modification of the partial awareness hypothesis seems appropriate. A recent suggestion was that the

gradualness of conscious perception depends on the level of processing of the stimulus (Windey et al., 2013; Windey, Vermeiren, Atas, & Cleeremans, 2014). According to standard assumptions about the visual system, processing of a stimulus forms a hierarchy, where processes engaged with more basis features of the stimulus provide input to processes extracting more complex features (Riesenhuber & Poggio, 1999). Low-level processing of a stimulus, which endows the observer with single features of the stimulus such as colour and orientation, may vary gradually. In contrast, high-level processing, required for instance to determine the meaning of words, could be competed in an all-or-none fashion (Windey & Cleeremans, 2015).

6.2.2. Methodological implications for research on conscious awareness

The present analysis suggested that it can be advantageous to record subjective measures of visual experience using scales with high resolution. The advantage of subjective measures with higher resolution is that fine-grained measurements maximize the amount of information obtained by a fixed number of trials. However, increasing the resolution of subjective measures will only be feasible if a more fine-grained scale does not distort the measurements. The existing literature is critical about the possibility of obtaining high resolution measurements by visual analogue scales (VAS). The numerous concerns raised against VAS can be summarized into two main issues (Overgaard et al., 2006):

(i) Participants may be unable to distinguish between the numerous response alternatives offered by a VAS.

(ii) Participants' reports could be attracted by the scale ends, creating a more sharp transition between conscious and unconscious perception than there actually is.

Concerning the first issue, the present study suggested that using more precision in the scale increases the correlation with discrimination task performance and improves internal consistency, which is only possible if participants are able to use at least more than four positions on the VAS. Concerning the second issue, the distribution of reports collected by the VAS was not U-shaped; instead, for the intermediate levels of motion coherence, the distribution was centred at the middle parts of the scale. As can be seen in Fig. 4-6, the major difference between the distributions of VAS and the discrete scale was that participants preferred "weak experience" on the discrete scale and "clearer than the central position" on the VAS. Admittedly, there was also a non-significant tendency that participants reported no

experience at all slightly more often on a VAS than on a discrete scale. Should this effect be corroborated in the future, it could reflect misses of conscious experiences by the VAS, but it could also indicate false positives by the discrete scale. Independently of the explanation of this effect, it is certainly too small to mask a gradual transition of unconscious to conscious perception. Consequently, the benefits of using visual analogue scales outweigh the potential risks of using a discrete scale for future studies.

6.2.3. Methodological implications for research on subjective measures

Empirical studies investigating the number of categories of subjective measures need a method to validate the observed gradualness of the transition from unconscious to conscious perception. While many existing studies considered the frequent use of central scale steps as a beneficial feature of subjective measures (Overgaard et al., 2006; Sandberg et al., 2010, 2011; Wierzchoń et al., 2012, 2014), the present study indicates that central scale steps could also be *overused:* The second scale step of the discrete scale was the dominant response at four out of six levels of coherence, including the most difficult stimuli where performance was at chance. In this case, reports of weak experiences mainly reflect noise in the system, not signal. A more uniform distribution of reported visual experience may well indicate a smooth transition from unconscious to conscious perception, but a second possibility is that the gradual distribution reflects a greater amount of noise. After all, unsystematic noise always increases the variance of the distribution: A subjective measure that differs from another subjective measure only in the number of unsystematic noise picked up by the measure will create a more uniform distribution; if the scale picks up noise with infinite variance, the result would be a perfectly uniform distribution. The same is true for the second method to establish the gradualness scale, psychometric function analysis. A less steep slope of the psychometric function was interpreted as a smooth transition between unconscious and conscious perception (Sandberg et al., 2011; Windey et al., 2013). However, a flatter curve can also indicate a great amount of noise. After all, the slope of a psychometric function can also be interpreted as the relative sensitivity of the scale to changes of the quality of stimulation, with greater slopes being more sensitive (Gescheider, 1997).

In summary, research on subjective measures lacks a method to differentiate gradual conscious perception from noise. A potential solution to this problem is provided by type 2 sensitivity (Fleming & Lau, 2014; Galvin et al., 2003): If a more fine-grained subjective measure correlates more closely with task accuracy than a subjective measure with identical

content but a smaller number of categories, it suggests that participants were able to make use of the additional categories. In contrast, if the correlation is the same, it shows that the additional information in the fine-grained scale does not share variance with objective task performance, and thus the task-relevant conscious experience cannot be measured on the more fine-grained level. Overall, the gradualness of subjective measures is a topic where using the association with task performance as a reference frame appears to be beneficial (see section 1.5.3; Fleming & Lau, 2014).

A potential drawback of using type 2 sensitivity as reference frame is that signal detection theory derived measures are not available for every paradigm. Signal detection theory provides mathematical tools to differentiate between the observers' sensitivity to distinguish between signal and noise and participants' response criteria (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). While type 2 ROC curves and meta-d_a are two convenient measures of type 2 sensitivity in all tasks where participants are required to select one out of two options (Fleming & Lau, 2014; Fleming et al., 2010; Maniscalco & Lau, 2012), there is no method to distinguish between sensitivity and criteria when the number of task options is greater than two. Nevertheless, continuous tasks are appealing for research on the gradualness of conscious awareness because they ensure that a low number of conscious states cannot be caused by a binary task (see 4.2.3., Rausch & Zehetleitner, 2014). Cumulative logistic regression was used in Chapter 4, but this approach is likely to be subject to the same problems as standard logistic regression (cf. 5.4.3., Rausch et al., 2015). As the bias can be expected to be roughly the same between the full resolution VAS and the discretized VAS, the effects found in the present study cannot be explained by bias alone. Nevertheless, future studies will need to carefully outweigh the prospects of using a continuous task against the drawbacks of being unable to properly control discrimination error.

6.3. Quantifying the relation between subjective measures and task accuracy

6.3.1. Logistic regression as measure of type 2 sensitivity

Previous studies quantifying the relation between subjective measures and task performance seem to follow two different approaches: some studies explicitly applied a measure of SDT (Charles et al., 2013; Fleming et al., 2010; Scott, Dienes, Barrett, Bor, &

Seth, 2014; Szczepanowski et al., 2013; Vandenbroucke et al., 2014; Zehetleitner & Rausch, 2013), and the others used flavours of generalized linear regression models (Rausch & Zehetleitner, 2014; Sandberg et al., 2013, 2010; Wierzchoń et al., 2012, 2014). However, the distinction between the two approaches is only superficial: It can be shown that each generalized linear regression model is identical to always one specific SDT model (Brockhoff & Christensen, 2010; DeCarlo, 1998). Consequently, the core issue about generalized linear regression is whether the underlying SDT model is appropriate.

The standard regression model to quantify type 2 sensitivity involves the subjective measure as predictor, task accuracy as dependent variable, and the logit function $f(x) = \log(\frac{x}{1-x})$ to convert the probability of being correct bounded by 0 and 1 to a variable that is free vary between -∞ and ∞. This model corresponds to an SDT model assuming logistic distributions with equal variances for correct and erroneous trials. Specifically the assumption of equal variances in correct and incorrect trials is not consistent with standard models of perceptual decision making: Assuming that participants choose a response to the task and make a report based on the same sensory evidence, the distributions for correct and erroneous trials are expected to be unequal and heavily skewed (Galvin et al., 2003). The consequence would be that logistic regression models confound participants' type 2 sensitivity and participants' criteria. For another measure that relies on the assumption of equal variances, type 2 d', there is empirical evidence that it depends on participants' criteria (Evans & Azzopardi, 2007). Under the assumption of the equal Gaussian SDT model, type 2 d' it is maximized by conservative criteria (Barrett et al., 2013). The only difference between logistic regression and type 2 d' is that the former assumes logistic distributions and the latter Gaussian distributions. Although the tails of the logistic distribution are heavier than the tails of the Gaussian, the two distributions typically converge to very similar results SDT applications (DeCarlo, 1998). Consequently, it is reasonable to expect that logistic regression underestimates type 2 sensitivity of liberal subjective measures, too.

The reanalysis described in Chapter 5 revealed that logistic regression analysis and meta-d_a revealed by-and-large converging results with respect to the content of subjective measures and type 2 sensitivity (Rausch et al., 2015). As stimulus-related measures are typically associated with more conservative criteria but also smaller type 2 sensitivities when measured by meta-d_a or type 2 ROC curves, it might be expected that the use of logistic regression conceals the effect of contents on type 2 sensitivities. However, if there had been a

bias against liberal subjective measures, it would not have been sufficiently strong to mask the effect of content in the present experiments. Although there was no devastating bias in the present experiment, it would be a misinterpretation that the influence of criteria on logistic regression can be neglected at the interpretation of other studies: If logistic regression slopes are indeed maximized by a conservative reporting strategy, there are always two possibilities why an empirical effect on logistic regression slopes is observed: it could be a difference in type 2 sensitivity, but could also be a difference in criteria. Whenever greater logistic regression slopes are observed in a condition with more liberal criteria, it seems adequate to infer that the effect is due to type 2 sensitivity. As logistic regression slopes entail the risk to underestimate type 2 sensitivity of liberal criteria, a criterion confound cannot explain why there are greater slopes in the condition where the criteria are more liberal. For instance, in the experiment of Chapter 4, cumulative logistic regression slopes were greater for the visual analogue scale than for the discrete scale. This effect cannot be explained by a criterion shift, because the discrete scale was used in a slightly more conservative manner than the visual analogue scale (Rausch & Zehetleitner, 2014). In contrast, when greater logistic regression slopes occur in conjunction with more conservative criteria, the effect should not be interpreted as an effect of type 2 sensitivity: The effect of bias would go in the same direction and could account for the effect.

A second downside of logistic regression is that the logit link function frequently fails to linearize the relationship between subjective measures and task accuracy. Whenever non-linear trends occur, the logistic regression slope is hard to interpret as measure of type 2 sensitivity because there is no single slope for each condition; instead, the slope varies over the course of the scale. In addition, non-linear trends may aggravate the confound with criteria in n-AFC tasks because more liberal subjective scales feature a greater number of scale steps where the transformed accuracy cannot decrease any further due to a lower bound of accuracy imposed by the guessing probability.

Overall, those measures derived from SDT designed to control for criteria are clearly more promising options to quantify type 2 sensitivity than logistic regression. When interpreting the results of previous studies based on logistic regression, it should be critically considered whether an effect on logistic regression slopes can be explained by different criteria as well.

6.3.2. Alternative logistic regression models

While the standard logistic regression model seems not optimal to quantify type 2 sensitivity, there is still the question whether alternative regression models are more valid measures of type 2 sensitivity. Specifically, two potential modifications of the logistic regression model seem promising as solutions to some of the problems raised before:

(i) a model with adjusted link function to account for the probability of guessing in n-alternative forced-choice tasks (Brockhoff & Müller, 1997)
(ii) cumulative logistic regression with subjective reports as ordinal dependent variable to account for criteria (Ramsøy & Skov, 2014)

The first method addresses the bias of logistic regression by the lower bound of the logit transform of accuracy caused by the probability of guessing the correct response. If the participant has a chance of correctly guessing of *p*, the logit transform of the accuracy does not vary between -∞ and ∞; instead, the transformed accuracy is bounded at $\log(\frac{p}{1-p})$. To account for the guessing probability, it was proposed to use a link function that ensures that the transformed accuracy is free to vary in the full range between -∞ and ∞. This can be achieved by the adjusted link function $f(x) = \log\left(\frac{x-p}{1-x}\right)$ (Brockhoff & Müller, 1997; Knoblauch, 2014; Williams, Ramaswamy, & Oulhaj, 2006). As can be seen from Fig. 6-1, the adjusted link function creates fairly linear relationships between subjective measures and accuracy in the masked orientation task and the masked shape task, but linearization appears to fail in the global motion task.

The major problem is that the adjusted link function still does not provide any means to differentiate between sensitivity and criteria. Although the adjusted logit regression model does not consistently favour conservative or liberal subjective measures, it favours subjective measures with criteria spread over a wide range of performance levels. Consequently, different regression slopes between two conditions are not necessarily due to type 2 sensitivity; greater slopes can also be caused by one condition spreading participants criteria over a greater range of performance levels (Wierzchoń et al., 2012). In fact, the adjusted logit method was at the first place introduced as a method to quantify thresholds, i.e. criteria, not sensitivity (Brockhoff & Müller, 1997). As the adjusted logit link model does not control for criteria and does not even guarantee linear trends, measures such as meta-d_a and type 2 ROC

curves seem still more favourable option to quantify the relation between subjective measures and task accuracy.

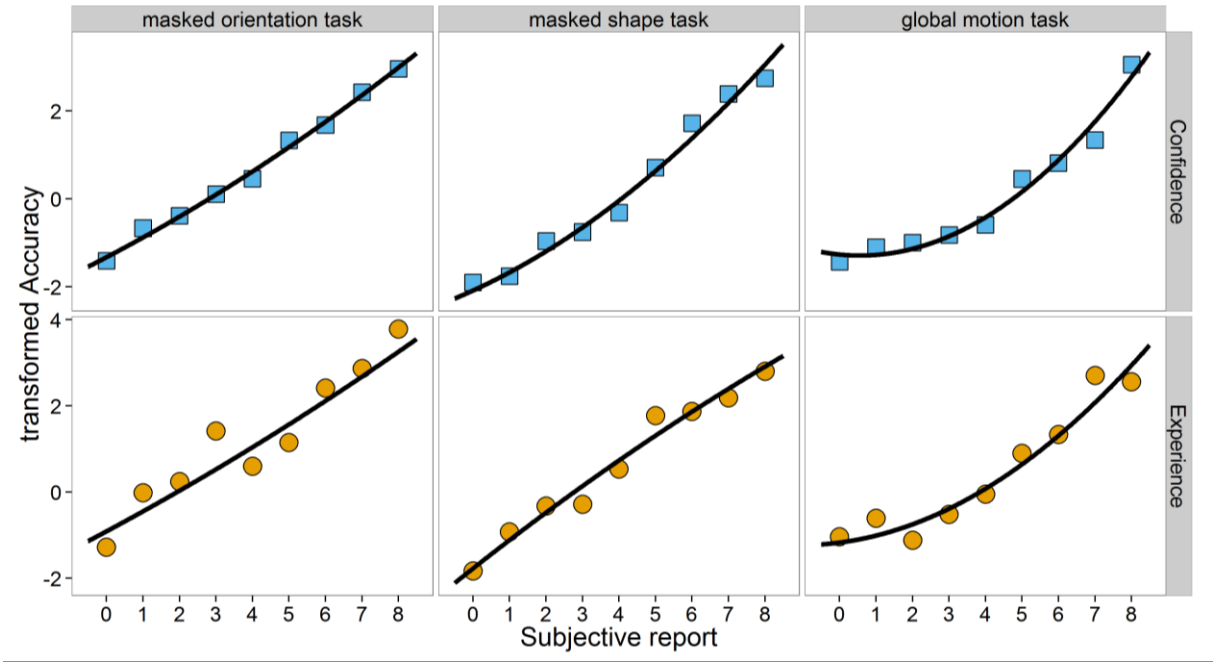

Figure 6-1: The relationship between subjective reports and accuracy transformed according to the adjusted logit link, separately for decisional confidence (upper row, blue) and visual experience (lower row, orange) in the masked orientation task, masked shape task, and the random-dot-motion discrimination task.

The second alternative to the standard logistic regression model is the cumulative logistic regression model (Ramsøy & Skov, 2014). In such a model, the cumulative probability of each rating category is predicted by trial accuracy (for a more thorough introduction into cumulative link models, see Christensen, 2015a). Again, cumulate logistic regression depends on a link function that relates predictors and probabilities. However, the model explicitly fits criteria that delineate between each two adjacent rating categories, and there is no assumption about a continuous linear relationship with subjective reports (Christensen & Brockhoff, 2013). This model is equivalent to the yes-no rating SDT model with logistic distributions (Christensen, Cleaver, & Brockhoff, 2011; DeCarlo, 1998). The model can include a scale parameter that allows distributions of different variances of evidence for correct and incorrect trials (Christensen, Cleaver, & Brockhoff, 2011). Moreover, slope and scale parameters can be converted into the area under the type 2 receiver operating characteristic (Christensen et al., 2011), a standard measure of type 2 sensitivity (Fleming et al., 2010). Overall, cumulative logistic regression has several promising features: explicit control over subjective criteria, unequal variances between correct and incorrect trials,

as well as easy conversion to the area under the ROC curve. It should be noted that many standard models of decision making predict quite asymmetric distributions of evidence (Barrett et al., 2013; Galvin et al., 2003; see also Fig. 5-1). The cumulative logistic regression model assumes logistic and thus symmetric distributions of evidence for both correct and incorrect trials, while the scale of the distributions can be different. Although cumulative regression can also be used with asymmetric distributions such as the log-log link (Christensen, 2015b), it is unlikely that these functions are able to emulate the precise distributions of evidence in correct and incorrect trials. However, models that describe subjective reports based standard models of decision making have come under pressure because several studies suggested that evidence for subjective reports can generated in parallel to discrimination performance (Fleming et al., 2015; Scott et al., 2014). The distributions of evidence generated by parallel or dual-route models have not been mathematically formulated yet; consequently, it would be premature to reject cumulative logistic regression based on theoretical intuitions concerning the shape of distributions while these intuitions have yet to be substantiated. Among the different options to quantify type 2 sensitivity based on regression, cumulative logistic regression is clearly the most promising option.

6.4. Subjective measures: useful data for consciousness research?

Subjective measures of conscious awareness have not been considered appropriate for an objective science for several decades (Boring, 1953; Danziger, 1980; Eriksen, 1960; Hannula et al., 2005; Irvine, 2012; Schmidt & Vorberg, 2006). This widely held belief within the scientific community sharply contrasts the philosophical view that mature sciences of psychology and neuroscience should strive to explain participants' subjective reports about their conscious experience just as they should explain any other behaviour of human beings (Dennett, 2003, 2007). The present series of studies suggests that subjective measures are eligible for non-trivial research about conscious awareness.

On the one hand, subjective measures were able to contribute data of great relevance to widely debated topics in consciousness, specifically to the debates about the NCC, the partial awareness hypothesis, and the gradual transition between unconscious and conscious perception. In addition, subjective measures proved to be heuristically fertile: The behavioural difference between stimulus-related subjective measures and response-related subjective

measures can be seen as a new constraint to existing theories of conscious awareness, which could not have been identified based on objective measures alone.

On the other hand, the present work also highlights that researchers need to invest time into careful considerations about their experimental designs when they use subjective measures: First, it is necessary to consider which conscious content are appropriate to measure in the context of a specific research question. Second, given the "subjective" reputation of subjective measures, it is always useful and for some research questions necessary to supplement subjective measures by a frame of reference based on more "objective" data: either by a measurement of neural events, or by computing the association between subjective measures and performance in a discrimination task. However, both requirements are relatively easy to implement. Consequently, consciousness research would benefit from the more wide-spread usage of subjective measures of conscious awareness.

7. REFERENCES

Baars, B. J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences*, *6*(1), 47–52.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research*, *150*, 45–53. doi:10.1016/S0079-6123(05)50004-9

Baranski, J. V, & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–28.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–52. doi:10.1037/a0033268

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *arXiv Preprint arXiv:1406.5823*. Retrieved from http://arxiv.org/abs/1406.5823

Bates, D., Maechler, M., Bolker, B., & Walker, A. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. Retrieved from http://cran.r-project.org/package=lme4

Bayne, T. ., & Chalmers, D. J. (2003). What is the Unity of Consciousness? In A. Cleeremans (Ed.), *The unity of consciousness: Binding, integration, dissociation*. Oxford: Oxford University Press.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7.1–11. doi:10.1167/9.10.7

Bays, P. M., & Husain, M. (2008). Dynamic Shifts of Limited Working Memory resources in human vision. *Science*, *321*, 851–854. doi:10.1126/science.1158023

Bischof, N. (1965). Erkenntnistheoretische Grundlagenprobleme der Wahrnehmungspsychologie. In H. Metzger & W. Erke (Eds.), *Handbuch der Psychologie in 12 Bdn. Bd. 1/I: Wahrnehmung und Bewusstsein*.

Block, N. (2002). Some Concepts of Consciousness. Retrieved from http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/Abridged BBS.htm

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, *9*(2), 46–52. doi:10.1016/j.tics.2004.12.006

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, *15*(12), 567–75. doi:10.1016/j.tics.2011.11.001

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–35. doi:10.1016/j.tree.2008.10.008

Boring, E. G. (1953). A history of introspection. *Psychological Bulletion*, *50*(3), 169–189.

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 of the United States of America*, *102*(46), 16875–9. doi:10.1073/pnas.0505332102

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brockhoff, P. B., & Christensen, R. H. B. (2010). Thurstonian models for sensory discrimination tests as generalized linear models. *Food Quality and Preference*, *21*(3), 330–338. doi:10.1016/j.foodqual.2009.04.003

Brockhoff, P. M., & Müller, H. G. (1997). Random effect threshold models for dose-response relationships with repeated measurements. *Journal of the Royal Statistical Society Series B-Methodological*, *59*(2), 431–446.

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. doi:10.1159/000351027

Carruthers, P. (2011). Higher-Order Theories of Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2011.). Retrieved from http://plato.stanford.edu/archives/fall2011/entries/consciousness-higher/

Chalmers, D. J. (1994). Facing Up to the Problem of Consciousness. *Journal of Consicousness Studies*, *2*(3), 200–219.

Chalmers, D. J. (1998). On the Search for the Neural Correlate of Consciousness. In S. R. Hameroff, A. W. Kaszniak, & A. Scott (Eds.), *Toward a science of consicousness II: The second tucson discussions and debates*. Cambridge, MA: MIT Press.

Charles, L., King, J.-R., & Dehaene, S. (2014). Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli. *The Journal of Neuroscience*, *34*(4), 1158–70. doi:10.1523/JNEUROSCI.2465-13.2014

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94. doi:10.1016/j.neuroimage.2013.01.054

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics*, *36*(4), 387–95.

Christensen, R. H. B. (2013). Ordinal—Regression models for ordinal data. R package version 2010.06-12. Retrieved from http://www.cran.r-project.org/package=ordinal/

Christensen, R. H. B. (2015a). A Tutorial on fitting Cumulative Link Models with the ordinal Package. Retrieved from https://cran.r-project.org/web/packages/ordinal/vignettes/clm_tutorial.pdf

Christensen, R. H. B. (2015b). Analysis of ordinal data with cumulative link models: estimation with the R-package ordinal. Retrieved from https://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pd

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

Christensen, R. H. B., Cleaver, G., & Brockhoff, P. B. (2011). Statistical and Thurstonian models for the A-not A protocol with and without sureness. *Food Quality and Preference*, *22*(6), 542–549. doi:10.1016/j.foodqual.2011.03.003

Cleeremans, A. (2008). Consciousness: the radical plasticity thesis. *Progress in Brain Research*, *168*(07), 19–33. doi:10.1016/S0079-6123(07)68003-0

Cleeremans, A. (2011). The radical plasticity thesis : how the brain learns to be conscious. *Frontiers in Psychology*, *2*, 1–12. doi:10.3389/fpsyg.2011.00086

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007). Consciousness and

metarepresentation: a computational sketch. *Neural Networks*, *20*(9), 1032–9. doi:10.1016/j.neunet.2007.09.011

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

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience*, *6*(2), 119–26. doi:10.1038/nn0203-119

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Dannenbaum, E., Chilingaryan, G., & Fung, J. (2011). Visual vertigo analogue scale: an assessment questionnaire for visual vertigo. *Journal of Vestibular Research : Equilibrium & Orientation*, *21*(3), 153–9. doi:10.3233/VES-2011-0412

Danziger, K. (1980). The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences*, *16*, 241–262.

Davey, H. M., Barratt, A. L., Butow, P. N., & Deeks, J. J. (2007). A one-item question with a Likert or Visual Analog Scale adequately measured current anxiety. *Journal of Clinical Epidemiology*, *60*(4), 356–60. doi:10.1016/j.jclinepi.2006.07.015

de Boer, A. G. E. M., van Lanschot, J. J. B., Stalmeier, P. F. M., van Sandick, J. W., Hulscher, J. B. F., de Haes, J. C. J. M., & Sprangers, M. A. G. (2004). Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Quality of Life Research*, *13*(2), 311–20.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186–205. doi:10.1037//1082-989X.3.2.186

Dehaene, S. (2010). Conscious and Nonconscious Processes: Distinct Forms of Evidence Accumulation? In B. Duplantier & V. Rivasseau (Eds.), *Biological Physicss: Poincaré Seminar 2009* (pp. 141–168). Basel: Springer. doi:10.1007/978-3-0346-0428-4_7

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–27. doi:10.1016/j.neuron.2011.03.018

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–11. doi:10.1016/j.tics.2006.03.007

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8520–8525. doi:10.1073/pnas.1332574100

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology*, *5*(10), 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 the Cognitive Sciences*, *6*(1-2), 247–270. doi:10.1007/s11097-006-9044-9

Dienes, Z. (2004). Assumptions of Subjective Measures of Unconscious Mental States.

Journal of Consciousness Studies, *11*(9), 25–45.

Dienes, Z. (2008). Subjective measures of unconscious knowledge. *Progress in Brain Research*, *168*, 49–64. doi:10.1016/S0079-6123(07)68005-4

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274–290. doi:10.1177/1745691611406920

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research* (pp. 1–32). Oxford, England: Oxford University Press.

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*, *21*(5), 1322–1338.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Psychological Research*, *69*(5-6), 338–51. doi:10.1007/s00426-004-0208-3

Dienes, Z., & Seth, A. (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*(2), 674–81. doi:10.1016/j.concog.2009.09.009

Eimer, M., & Mazza, V. (2005). Electrophysiological correlates of change detection. *Psychophysiology*, *42*(3), 328–342.

Eriksen, C. W. (1960). Discrimination and learning without awareness: A methodological survey and evaluation. *The Psychological Review*, *67*(5), 279–299.

Evans, S., & Azzopardi, P. (2007). Evaluation of a "bias-free" measure of awareness. *Spatial Vision*, *20*(1), 61–77. doi:10.1163/156856807779369742

Fleming, S. M., & Dolan, R. J. (2010). Effects of loss aversion on post-decision wagering: implications for measures of awareness. *Consciousness and Cognition*, *19*(1), 352–63. doi:10.1016/j.concog.2009.11.002

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1338–49. doi:10.1098/rstb.2011.0417

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., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-Specific Disruption of Perceptual Confidence. *Psychological Science*, *26*(1), 89–98. doi:10.1177/0956797614557697

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*(5998), 1541–1543. doi:10.1126/science.1191883

Frässle, S., Sommer, J., Jansen, A., Naber, M., & Einhäuser, W. (2014). Binocular rivalry: frontal activity relates to introspection and action but not to perception. *The Journal of Neuroscience*, *34*(5), 1738–47. doi:10.1523/JNEUROSCI.4403-13.2014

Frith, C. (2011). Consicousness is for sharing. *Cognitive Neuroscience*, *2*(2), 117–118.

Funke, F., & Reips, U. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*. doi:10.1177/1525822X12444061

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*(4), 843–76.

Genetti, M., Britz, J., Michel, C. M., & Pegna, A. J. (2010). An electrophysiological study of conscious visual perception using progressively degraded stimuli. *Journal of Vision*, *10*(14), 10. doi:10.1167/10.14.10

Gescheider, G. A. (1997). *Psychophysics: The fundamentals*. Mahwah, NY: Lawrence Erlenbaum Publishers.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–74. doi:10.1146/annurev.neuro.29.051605.113038

Graziano, M. S. A., & Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neuroscience*, *2*(2), 98–113. doi:10.1080/17588928.2011.565121

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

Hake, B. Y. H. W., & Garner, W. R. (1951). The effect of presenting various numbers of discrete steps on scale reading accuracy. *Journal of Experimental Psychology*, *42*, 358–366.

Hannula, D. E., Simons, D. J., & Cohen, N. J. (2005). Imaging implicit perception: promise and pitfalls. *Nature Reviews Neuroscience*, *6*(3), 247–55. doi:10.1038/nrn1630

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. doi:10.3758/BF03203619

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*(5), 791–804. doi:10.1016/S0896-6273(02)01091-7

Hornsey, M. J., Olsen, S., Barlow, F. K., & Oei, T. P. S. (2012). Testing a single-item visual analogue scale as a proxy for cohesiveness in group psychotherapy. *Group Dynamics: Theory, Research, and Practice*, *16*(1), 80–90. doi:10.1037/a0024545

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*, 485–488. doi:10.1038/nature10836

Irvine, E. (2012). Old Problems with New Measures in the Science of Consciousness. *The British Journal for the Philosophy of Science*, *63*(3), 627–648. doi:10.1093/bjps/axs019

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quaterly*, *32*, 127–136.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186–200. doi:10.1037/a0025960

Kepecs, A., Uchida, N., Zariwala, H. a, & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–31. doi:10.1038/nature07200

Kiesel, A., Miller, J., Jolicoeur, P., & Brisson, B. (2008). Measurement of ERP latency differences: a comparison of single-participant and jackknife-based scoring methods. *Psychophysiology*, *45*(2), 250–74. doi:10.1111/j.1469-8986.2007.00618.x

Knoblauch, K. (2014). psyphy: Functions for analyzing psychophysical data in R. Retrieved from http://cran.r-project.org/package=psyphy

Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1401–11. doi:10.1098/rstb.2011.0380

Koivisto, M., Lähteenmäki, M., Sørensen, T. A., Vangkilde, S., Overgaard, M., & Revonsuo, A. (2008). The earliest electrophysiological correlate of visual awareness? *Brain and Cognition*, *66*(1), 91–103. doi:10.1016/j.bandc.2007.05.010

Koivisto, M., & Revonsuo, A. (2003). An ERP study of change detection, change blindness, and visual awareness. *Psychophysiology*, *40*(3), 423–9.

Koivisto, M., & Revonsuo, A. (2010). Event-related brain potential correlates of visual awareness. *Neuroscience and Biobehavioral Reviews*, *34*(6), 922–34. doi:10.1016/j.neubiorev.2009.12.002

Kontou, E., Thomas, S. A., & Lincoln, N. B. (2012). Psychometric properties of a revised version of the Visual Analog Mood Scales. *Clinical Rehabilitation*, *26*(12), 1133–40. doi:10.1177/0269215512442670

Koriat, A. (2007). Metacognition and Consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge Handbook of Consciousness*. New York: Cambridge University Press.

Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: model-based and distribution-free measures and evaluation. *Perception & Psychophysics*, *68*(3), 393–414.

Kornell, N., Son, L. K., & Terrace, H. S. (2007). Transfer of Metacognitive Skills and Hint Seeking in Monkeys, *18*(1), 64–71.

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, *14*(7), 301–7. doi:10.1016/j.tics.2010.04.006

Kuznetsova, A., Brockhoff, 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

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. doi:10.1016/j.tics.2006.09.001

Lamy, D., Salti, M., & Bar-Haim, Y. (2009). Neural correlates of subjective awareness and unconscious processing: an ERP study. *Journal of Cognitive Neuroscience*, *21*(7), 1435–46. doi:10.1162/jocn.2009.21064

Lau, H. (2008a). A higher order Bayesian decision theory of consciousness. *Progress in Brain Research*, *168*(07), 35–48. doi:10.1016/S0079-6123(07)68004-2

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

Lau, H. (2011). Theoretical motivations for investigating the neural correlates of consciousness. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 1–7. doi:10.1002/wcs.93

Lau, H., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(49), 18763–8. doi:10.1073/pnas.0607716103

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious

awareness. *Trends in Cognitive Sciences*, *15*(8), 365–73. doi:10.1016/j.tics.2011.05.009

Leon, G. R., Koscheyev, V. S., & Stone, E. A. (2008). Visual Analog Scales for Assessment of Thermal Perception in Different Environments. *Aviation, Space, and Environmental Medicine*, *79*(8), 784–786. doi:10.3357/ASEM.2204.2008

Lindsey, J. (2010). gnlm: Generalized Nonlinear Regression Models. R package version 1.0. Retrieved from http://www.commanster.eu/rcode.html

Luck, S. J. (2005). *An introduction to the event-related potential technique*. (M. Press, Ed.). Cambridge, MA.

Lycan, W. G. (2004). The superiority of HOP to HOT. *Advances in Consciousness Research*, *56*, 115–136.

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

Mangan, B. (2001). Sensation's Ghost: The Non-Sensory "Fringe" of Consciousness. *Psyche*, *7*(18).

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–30. doi:10.1016/j.concog.2011.09.021

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 509–27. doi:10.1037/a0014876

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

Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.9. Retrieved from http://cran.r-project.org/package=BayesFactor

Morin, A., Urban, J., Adams, P. D., Foster, I., Sali, A., Baker, D., & Sliz, P. (2012). Shining Light into Black Boxes. *Science*, *336*(6078), 159–160. doi:10.1126/science.1218263

Mormann, F., & Koch, C. (2007). Neural correlates of consciousness. *Scholarpedia*, *2*(12), 1740. doi:10.4249/scholarpedia.1740

Nagel, T. (1974). What is it like to be a bat ? *The Philosophical Review*, *83*(4), 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*(1), 83–93. doi:10.1007/s10071-010-0345-6

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125–173.

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: a critical review. *The Behavioral and Brain Sciences*, *37*(1), 1–19. doi:10.1017/S0140525X12003214

Nieuwenhuis, S., & de Kleijn, R. (2011). Consciousness of targets during the attentional blink: a gradual or all-or-none dimension? *Attention, Perception & Psychophysics*, *73*(2), 364–e73. doi:10.3758/s13414-010-0026-1

Overgaard, M., Fehl, K., Mouridsen, K., Bergholt, B., & Cleeremans, A. (2008). Seeing without Seeing? Degraded Conscious Vision in a Blindsight Patient. *PloS One*, *3*(8),

e3028. doi:10.1371/journal.pone.0003028

Overgaard, M., Rote, J., Mouridsen, K., & Ramsøy, T. Z. (2006). Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Consciousness and Cognition*, *15*(4), 700–8. doi:10.1016/j.concog.2006.04.002

Overgaard, M., & Sandberg, K. (2012). Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1287–96. doi:10.1098/rstb.2011.0425

Peirce, C. S., & Jastrow, J. (1885). On Small Differences in Sensation. *Memoirs of the National Academy of Sciences*, *3*, 73–83.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Persaud, N., Davidson, M., Maniscalco, 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. *NeuroImage*, *58*(2), 605–11. doi:10.1016/j.neuroimage.2011.06.081

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), 257–61. doi:10.1038/nn1840

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team. (2012). nlme: linear and nonlinear mixed effects models.

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

Pins, D., & Ffytche, D. (2003). The neural correlates of conscious vision. *Cerebral Cortex*, *13*(5), 461–74.

Pitts, M. A., Martínez, A., & Hillyard, S. A. (2012). Visual processing of contour patterns under conditions of inattentional blindness. *Journal of Cognitive Neuroscience*, *24*(2), 287–303. doi:10.1162/jocn_a_00111

Pitts, M. A., Metzler, S., & Hillyard, S. A. (2014). Isolating neural correlates of conscious perception from neural correlates of reporting one's perception. *Frontiers in Psychology*, *5*, 1–16. doi:10.3389/fpsyg.2014.01078

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. doi:10.1037/a0019737

Pourtois, G., De Pretto, M., Hauert, C.-A., & Vuilleumier, P. (2006). Time course of brain activity during change blindness and change awareness: performance is predicted by neural events before change onset. *Journal of Cognitive Neuroscience*, *18*(12), 2108–29. doi:10.1162/jocn.2006.18.12.2108

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

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the

precision and likelihood of successful maintenance of visual working memory. *Journal of Vision, 12*(13), 21. doi:10.1167/12.13.21

Railo, H., Koivisto, M., & Revonsuo, A. (2011). Tracking the processes behind conscious perception: a review of event-related potential correlates of visual consciousness. *Consciousness and Cognition, 20*(3), 972–83. doi:10.1016/j.concog.2011.03.019

Rampling, J., Mitchell, A. J., Von Oertzen, T., Docker, J., Jackson, J., Cock, H., & Agrawal, N. (2012). Screening for depression in epilepsy clinics. A comparison of conventional and visual-analog methods. *Epilepsia*, *53*(10), 1713–21. doi:10.1111/j.1528-1167.2012.03571.x

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

Ramsøy, T. Z., & Skov, M. (2014). Brand preference affects the threshold for perceptual awareness. *Journal of Consumer Behaviour, 13*(1), 1–8. doi:10.1002/cb.1451

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108.

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

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

Rees, G., Kreiman, G., & Koch, C. (2002). Neural correlates of consciousness in humans. *Nature Reviews Neuroscience, 3*, 1–10. doi:10.1038/nrn783

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

Riddoch, B. Y. G. (1917). Dissociation of visual perceptions due to occipital injuries, with especial reference to appreciation of movements. *Brain*.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–25. doi:10.1038/14819

Rizopoulos, D. (2006). ltm: An R-package for latent variable modeling and item response theory analysis. *Journal of Statistical Software, 17*(5), 1–25.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*(3), 329–359. doi:10.1007/BF00355521

Rosenthal, D. M. (2000). Consciousness, content, and metacognitive judgments. *Consciousness and Cognition, 9*, 203–14. doi:10.1006/ccog.2000.0437

Rosenthal, D. M. (2009). Concepts and Definitions of Consicousness. In William P. Banks (Ed.), *Encyclopedia of Consciousness* (pp. 157–169). Amsterdam, Netherlands: Elsevier.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research, 47*(6), 877–903. doi:10.1080/00273171.2012.734737

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–37. doi:10.3758/PBR.16.2.225

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst

transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*(3), 165–75. doi:10.1080/17588921003632529

Sahraie, A., Weiskrantz, L., & Barbur, J. L. (1998). Awareness and confidence ratings in motion perception without geniculo-striate projection. *Behavioural Brain Research*, *96*(1-2), 71–7.

Sahraie, A., Weiskrantz, L., Trevethan, C. T., Cruce, R., & Murray, A. D. (2002). Psychophysical and pupillometric study of spatial channels of visual processing in blindsight. *Experimental Brain Research*, *143*(2), 249–56. doi:10.1007/s00221-001-0989-1

Salti, M., Bar-Haim, Y., & Lamy, D. (2012). The P3 component of the ERP reflects conscious perception, not confidence. *Consciousness and Cognition*, *21*, 961–968.

Sandberg, K., Bibby, B. M., & Overgaard, M. (2013). Measuring and testing awareness of emotional face expressions. *Consciousness and Cognition*, *22*(3), 806–9. doi:10.1016/j.concog.2013.04.015

Sandberg, K., Bibby, B. M., Timmermans, B., Cleeremans, A., & Overgaard, M. (2011). Measuring consciousness: task accuracy and awareness as sigmoid functions of stimulus duration. *Consciousness and Cognition*, *20*(4), 1659–75. 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*(4), 1069–1078. doi:10.1016/j.concog.2009.12.013

Schankin, A., & Wascher, E. (2007). Electrophysiological correlates of stimulus processing in change blindness. *Experimental Brain Research*, *183*(1), 95–105. doi:10.1007/s00221-007-1023-z

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), 25. doi:10.1167/12.11.25

Schmidt, T., & Vorberg, D. (2006). Criteria for unconscious cognition: three types of dissociation. *Perception & Psychophysics*, *68*(3), 489–504.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(5), 1264–88. doi:10.1037/a0012943

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind insight: Metacognitve discrimination depite chance task performance. *Psychological Science*, *25*, 1–20. doi:10.1177/0956797614553944

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391–400. doi:10.1038/nn1549

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

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(8), 314–321. doi:10.1016/j.tics.2008.04.008

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning

systems. *Behavioral and Brain Sciences*, *17*(3), 367–395. doi:10.1017/S0140525X00035032

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Simonsohn, U. (2013). Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875–88. doi:10.1177/0956797613480366

Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2008). Are there multiple visual short-term memory stores? *PloS One*, *3*(2), e1699. doi:10.1371/journal.pone.0001699

Spering, M., & Carrasco, M. (2015). Acting without seeing: eye movements reveal visual processing without awareness. *Trends in Neurosciences*, *38*(4), 247–258. doi:10.1016/j.tins.2015.02.002

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29.

Szczepanowski, R., Traczyk, J., Wierzchoń, M., & Cleeremans, A. (2013). The perception of visual emotion: comparing different measures of awareness. *Consciousness and Cognition*, *22*(1), 212–20. doi:10.1016/j.concog.2012.12.003

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1412–23. doi:10.1098/rstb.2011.0421

Ulrich, R., & Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology*, *38*(5), 816–27.

Vadillo, M. a., Konstantinidis, E., & Shanks, D. R. (2015). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-015-0892-6

Van Gulick, R. (2014). Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 201.). Retrieved from http://plato.stanford.edu/archives/spr2014/entries/consciousness/

Vandenbroucke, A. R. E., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. F. (2014). Accurate metacognition for visual sensory memory representations. *Psychological Science*, *25*(4), 861–873. doi:10.1177/0956797613516146

Vandenbroucke, A. R. E., Sligte, I. G., & Lamme, V. A. F. (2011). Manipulations of attention dissociate fragile visual short-term memory from visual working memory. *Neuropsychologia*, *49*(6), 1559–1568. doi:10.1016/j.neuropsychologia.2010.12.044

Velmans, M. (2000). *Understanding consciousness*. London: Routledge.

Velmans, M. (2007). An epistemology for the study of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell Companion to Consciousness* (pp. 1–17). New York: Blackwell.

Verleger, R. (2010). Markers of awareness? EEG potentials evoked by faint and masked events, with special reference to the "attentional blink ." In I. Czigler & I. Winkler (Eds.), *Unconscious memory representations in perception: Processes and mechanisms in the brain*. Amsterdam: John Benjamins Publishing Company.

Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.

Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.

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

Weiskrantz, L., Barbur, J. L., & Sahraie, A. (1995). Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (V1). *Proceedings of the National Academy of Sciences of the United States of America, 92*(13), 6122–6.

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*(3), 1141–53. 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

Williams, J., Ramaswamy, D., & Oulhaj, A. (2006). 10 Hz flicker improves recognition memory in older people. *BMC Neuroscience, 7*, 21. doi:10.1186/1471-2202-7-21

Windey, B., & Cleeremans, A. (2015). Consciousness as a graded and an all-or-none phenomenon: A conceptual analysis. *Consciousness and Cognition, 35*, 185–191. doi:10.1016/j.concog.2015.03.002

Windey, B., Gevers, W., & Cleeremans, A. (2013). Subjective visibility depends on level of processing. *Cognition, 129*(2), 404–409. doi:10.1016/j.cognition.2013.07.012

Windey, B., Vermeiren, A., Atas, A., & Cleeremans, A. (2014). The graded and dichotomous nature of visual awareness. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369*, 20130282.

World Medical Association. (2008). WMA declaration of Helsinki: Ethical principles for medical research involving human subjects. Retrieved from http://www.wma.net/en/30publications/10policies/b3/index.html

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

Zeki, S. (2003). The disunity of consciousness. *Trends in Cognitive Sciences, 7*(5), 214–218. doi:10.1016/S1364-6613(03)00081-0

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

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–5. doi:10.1038/nature06860

Zokaei, N., Gorgoraptis, N., Bahrami, B., Bays, P. M., & Husain, M. (2011). Precision of working memory for visual motion sequences and transparent motion surfaces. *Journal of Vision, 11*(2), 1–18. doi:10.1167/11.14.2

8. LIST OF PUBLICATIONS

Peer-reviewed journals

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 (Chapter 2 of this thesis)

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 (Chapter 4 of this thesis)

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 (Chapter 5 of this thesis)

* shared first authorship

Submitted manuscripts

Rausch, M., Wykowska, A., & Zehetleitner, M. (under review). Electrophysiological correlates of confidence and experience: Do you feel that you get it right before you see it? (Chapter 3 of this thesis)

9. AFFIDAVIT / EIDESSTATTLICHE ERKLÄRUNG

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation „Content, granularity, and type 2 sensitvity of subjective measures of visual consciousness“ selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation “Content, granularity, and type 2 sensitivity of subjective measures of visual consciousness” is the result of my own work that I have only used sources or materials listed and specified in the dissertation.

München, 1 October 2015 Manuel Rausch

10. DECLARATION OF AUTHOR CONTRIBUTIONS

Chapter 1 (Introduction)

Manuel Rausch wrote the Chapter.

Chapter 2 (Stimulus-related vs. response-related subjective measures)

Michael Zehetleitner conceived the research question, Michael Zehetleitner and Manuel Rausch designed the experiments, Manuel Rausch collected the data, Manuel Rausch analysed the data, Michael Zehetleitner and Manuel Rausch co-wrote the manuscript.

Chapter 3 (Electrophysiological correlates of confidence and experience

Manuel Rausch, Agnieszka Wykowska, and Michael Zehetleitner conceived the research question and designed the experiment, Manuel Rausch collected the data, Manuel Rausch and Agnieszka Wykowska performed the EEG analysis; Manuel Rausch and Michael Zehetleitner performed the statistical analysis; Manuel Rausch, Agnieszka Wykowska, and Michael Zehetleitner co-wrote the manuscript.

Chapter 4 (Visual analogue and discrete scales as measures of visual experience)

Manuel Rausch and Michael Zehetleitner conceived the research question and designed the experiment, Manuel Rausch collected and analysed the data, Manuel Rausch, and Michael Zehetleitner co-wrote the manuscript.

Chapter 5 (Type 2 sensitivity of decisional confidence and visual experience)

Manuel Rausch conceived the research questions and conducted the analysis; Manuel Rausch, Hermann J. Müller, and Michael Zehetleitner co-wrote the manuscript.

Chapter 6 (Final discussion)

Manuel Rausch wrote the Chapter.

Manuel Rausch	Michael Zehetleitner (Lab head)
Munich, 1 October 2015	Eichstätt, 1 October 2015