Consciousness and Cognition 22 (2013) 806–809

Contents lists available at SciVerse ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Commentary

Measuring and testing awareness of emotional face expressions

Kristian Sandberg [a,b,*], Bo Martin Bibby [c], Morten Overgaard [a,d]

[a] Cognitive Neuroscience Research Unit, Aarhus University Hospital, Nørrebrogade 44, Building 10G, 8000 Aarhus C, Denmark
[b] UCL Institute of Cognitive Neuroscience, University College London, 17 Queen Square, WC1N 3AR London, United Kingdom
[c] Department of Biostatistics, Aarhus University, Bartholins Allé 2, Building 1261, 8000 Aarhus C, Denmark
[d] Cognitive Neuroscience Research Unit, Dept. of Communication and Psychology, Aalborg University, Kroghstræde 3, 9220 Aalborg Ø, Denmark

ARTICLE INFO

Article history:
Received 26 February 2013

Keywords:
Consciousness
Perceptual awareness scale (PAS)
Confidence ratings (CR)
Post-decision wagering (PDW)
Emotion
Awareness

ABSTRACT

Comparison of behavioural measures of consciousness has attracted much attention recently. In a recent article, Szczepanowski et al. conclude that confidence ratings (CR) predict accuracy better than both the perceptual awareness scale (PAS) and post-decision wagering (PDW) when using stimuli with emotional content (fearful vs. neutral faces). Although we find the study interesting, we disagree with the conclusion that CR is superior to PAS because of two methodological issues. First, the conclusion is not based on a formal test. We performed this test and found no evidence that CR predicted accuracy better than PAS ($p = .4$). Second, Szczepanowski et al. used the present version of PAS in a manner somewhat different from how it was originally intended, and the participants may not have been adequately instructed. We end our commentary with a set of recommendations for future studies using PAS.

1. Introduction

Over the last few years, behavioural methods for assessing consciousness have become a topic of much scientific debate, and the main focus has been on comparing different measures (scales) in order to examine if one is superior to the others (Overgaard & Sandberg, 2012). A large part of this debate was motivated by the proposal of post-decision wagering (PDW) as a so-called "objective" measure of consciousness by Persaud, McLeod, and Cowey (2007). Here, the authors showed, among other findings, an imperfect correlation between task accuracy in a visual detection task and wagering behaviour of a blindsight patient, GY, thus indicating that GY was not fully aware of the targets that he was nevertheless able to report. However, PDW was quickly criticized for being an indirect measure of mental content rather than awareness (Seth, 2008), and it was pointed out that loss aversion (Schurger & Sher, 2008) and/or other factors could prevent participants from wagering high in the presence of sensory awareness (Clifford, Arabzadeh, & Harris, 2008). These claims were subsequently supported by empirical studies finding PDW to be inferior to simpler measures such as confidence ratings (CR) and ratings of the clarity of the visual experience (Dienes & Seth, 2010; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010; Wierzchoń, Asanowicz, Paulewicz, & Cleeremans, 2012).

Overlapping with, and slightly predating this debate, work in our group focused on examining which scale participants prefer to use for reporting conscious experiences if allowed to construct the scale themselves (Ramsøy & Overgaard, 2004). This resulted in the 4-point so-called perceptual awareness scale (PAS), ranging from "no experience" over "a brief glimpse" to "an almost clear experience" and ending at "a clear experience". PAS was subsequently compared to a dichotomous measure of awareness for normal participants (Overgaard, Rote, Mouridsen, & Ramsøy, 2006) and for a blindsight patient

DOI of original article: http://dx.doi.org/10.1016/j.concog.2012.12.003

* Corresponding author at: Cognitive Neuroscience Research Unit, Aarhus University Hospital, Nørrebrogade 44, Building 10G, 8000 Aarhus C, Denmark.
E-mail address: krissand@rm.dk (K. Sandberg).

1053-8100/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.concog.2013.04.015

(Overgaard, Fehl, Mouridsen, Bergholt, & Cleeremans, 2008), as well as to 4-point CR and PDW scales (Sandberg et al., 2010). In all cases, PAS was found to be superior to the competing scales, most frequently by showing better accuracy-awareness correlation and/or less subliminal perception when participants reported not having seen the target. In other words, these studies suggested that the best results were obtained by asking participants to report their experiences directly and allowing them to do this on a scale created either by themselves or other participants presented with similar stimuli.

In another recent article Szczepanowski, Traczyk, Wierzchoń, and Cleeremans (2013) compared participants' use of CR, PAS, PDW during a visual identification task, now using faces with fearful and neutral expressions instead of the simple visual shapes used in Sandberg et al. (2010). Szczepanowski et al. estimated the relationship between accuracy and awareness for each scale using Pearson chi-square tests of independence as well as type 2 receiver operating characteristics (ROC) analysis. Based on the resulting values, the authors argued that CR produced the best relationship between accuracy and awareness, PAS the second best, and PDW the worst, thus indicating that CR was the most exhaustive subjective measure when examining stimuli with emotional content. Although we find the article very interesting and a relevant contribution to the field, we disagree with the authors' conclusion that confidence ratings are more exhaustive than PAS in paradigms with emotional stimuli for two reasons. The first is the employed statistical method, and the second is the manner in which PAS was used. In the following, we will elaborate these points in detail.

2. Statistical model selection and tests of differences

As mentioned above, Szczepanowski et al. used type 2 ROC analysis and Pearson chi-square tests to examine if task accuracy varied across awareness ratings within each scale and compare the obtained chi^2 values numerically in order to rank the scales. The type 2 ROC analysis indicated that all scales were equally sensitive (mean sensitivity was between 0.58 and 0.59 for all scales) although no formal test was performed. The conclusions regarding the inter-scale differences were based on the chi-square test. We find that this method is inappropriate in two ways: First, it assumes that all observations are independent, which is not the case as each participant contributed with 80 observations for each scale. An optimal statistical model should thus take into account that the data originates from *N* clusters (where *N* is the number of participants). Second, even if it was reasonable to assume independence, then ranking of chi^2 values is not a formal comparison of scales and does thus not allow us to conclude whether a difference between scales is significant or not. An optimal model should allow for such formal testing of the null hypotheses that the relationship between task accuracy and awareness rating is equally good for scale 1 and 2, scale 1 and 3, and scale 2 and 3. Both these goals can be achieved by use of logistic regression.

Logistic regression furthermore allows for a second test of scale exhaustiveness – tests using the so-called "guessing criterion" (Dienes, Altmann, Kwan, & Goode, 1995). Using this, performance is compared between scales for those cases where participants claim to be guessing or to have no experience (that is, they use the lowest awareness rating). It has been argued that above-chance performance in these cases reflect unconscious processing. Yet if one scale finds unconscious processing but another does not, it is highly likely that the first scale is less exhaustive (unless the second scale is not exclusive, i.e. if it misclassifies unconscious processing as conscious processing) ((Sandberg et al., 2010), but see also Dienes and Seth (2010) and Timmermans, Sandberg, Cleeremans, and Overgaard (2010)). Szczepanowski et al. have kindly allowed us access to the original data to test for differences in: (a) how well each scale predicted task accuracy in general, and (b) how much subliminal perception was indicated at the lowest awareness ratings.

We created a logistic regression mixed model with accuracy as the dependent variable and with scale and awareness rating as independent variables, and with participant as a random effect. First, we conducted pair-wise comparisons between all three scales in order to test if one scale predicted accuracy significantly better than the others. We found that both CR and PAS predicted accuracy significantly better than PDW ($z = 4.07$, $p < .001$ for CR vs. PDW, and $z = 2.28$, $p = .023$ for PAS vs. PDW), but importantly, no significant difference was observed between CR and PAS ($z = 0.87$, $p = .39$). Overall, this analysis thus found both CR and PAS to be superior to PDW, but there was no evidence for a significant difference between CR and PAS.

Second, we calculated task accuracy at the lowest awareness rating for each scale and compared this to the chance-level of 50%. For CR, task accuracy was 42%, 95%-CI: (28;55)%, which was not significantly different from chance ($z = -1.21$, $p = .23$). For PAS, task accuracy was 46%, 95%-CI: (41;52)%, which was not significantly different from chance ($z = -1.19$, $p = .24$). For PDW, task accuracy was 56%, 95%-CI: (50;62)%. This was almost but not quite significantly different from chance ($z = 1.87$, $p = .06$). Furthermore, we found that accuracy for awareness rating 1 was significantly higher for PDW than for both CR ($z = -2.75$, $p = .006$) and PAS ($z = -2.15$, $p = .032$), but importantly, no significant difference was observed between CR and PAS ($z = -0.77$, $p = .44$). Overall, we thus failed to reject the null hypothesis that accuracy was at chance for all scales although particularly for PDW, this may have been a matter of statistical power. The accuracy was significantly higher for PDW than for CR and possibly also than for PAS. Again, it is important to note that no significant difference was found between CR and PAS.

In summary, we found no evidence of a significant difference in the exhaustiveness of CR and PAS, but both scales were significantly more exhaustive than PDW. Compared to Sandberg et al. (2010), the main impact of Szczepanowski et al.'s emotional stimuli were thus a lack of a statistically significant difference between CR and PAS. The absence of a statistically significant difference between PAS and CR could, in principle, be a matter of statistical power, yet as mentioned above, it could also be related to the manner in which PAS was used. Below, we will discuss this last point in detail.

3. The use of PAS

Szczepanowski et al. report that the PAS "is a 4-point verbal scale that attempts to measure the quality of conscious experience directly. It asks participants to evaluate the visibility of the percept as subjective certainty of its presence (the metacognitive judgment of a percept's accessibility" (p. 213). However, this interpretation is only partially true. In the publication originally introducing PAS, Ramsøy and Overgaard (2004) wrote: "In describing and reporting sensations in terms of clearness, it is important to make the distinction between degrees of clearness and degrees of certainty about one's answer" (p. 10). Thus, we would like to emphasize that PAS should not be equated with "subjective certainty". Essentially, "subjective certainty" is exactly what CR measures, and, accordingly, if PAS was introduced to participants by Szczepanowski et al. using a "subjective certainty" terminology, the non-significant differences between CR and PAS conditions are not surprising. We acknowledge, of course, that some judgement is needed when rating on the PAS (and one may be more or less confident in the accuracy of the judgment), but our main point here is that this is related to a comparison of the (remembered) visual experience and the scale step description and not an assessment of the probability of a stimulus being present based on the (remembered) visual experience.

Ramsøy and Overgaard (2004) report that PAS is constructed from the reasoning that methods for subjective reports of conscious experience should be developed in collaboration with the reporting subjects. Thus, they make no claim that 4-point scales should be preferred a priori – on the contrary, they argue that one cannot prefer *any* particular scale construction a priori. In the study, participants generated the scale while reporting their experience of the colour, shape, and position of simple figures. The validity of the PAS in later studies may reasonably be assumed to depend on the similarity of the stimulus material in the original and later study as well the how well the participants were instructed in the interpretation of scale step descriptions.

In most later studies performed by our group using PAS (e.g. Overgaard, Nielsen, & Fuglsang-Frederiksen, 2004; Overgaard et al., 2006; Overgaard et al., 2008), the exact meaning of the individual labels were discussed with the subjects and tested in several pilot trials where the experimenter made sure that there were no misunderstandings. One crucial aspect is the distinction between "brief glimpse" and an "almost clear experience". For brief glimpses, there is a conscious experience caused by the presentation of the relevant stimulus, yet the participant is nevertheless unable to report the content of this experience. Typically, this experience is a vague short-lived sensation that "something was there" or simply that "it was different from nothing at all" (see reviews in Overgaard and Sandberg (2012), and Overgaard (2012)). Few participants understand this distinction immediately if the procedure does not include thorough instruction, discussion and pilot trials. Indeed, recent (not yet published) findings indicate that in-depth instructions improve the accuracy-awareness correlation for PAS. The absence of in-depth instruction may thus have been a contributing factor to the results of Szczepanowski et al. Yet it should be noted that only written instructions were given on scale use in Sandberg et al. (2010) to ensure that participants were not given more in-depth instructions than participants using CR and PDW. For this reason, it is unlikely that verbal instructions alone are the cause of the difference in the findings.

The other possibility is the nature of the stimuli used in the experiment. Most studies using PAS, have used geometric figures very similar to those used by Ramsøy and Overgaard (2004) and asked participants to report the shape (Overgaard et al., 2008; Sandberg, Bibby, Timmermans, Cleeremans, & Overgaard, 2011; Sandberg et al., 2010; Schwiedrzik, Singer, & Melloni, 2011) while one study used different, but still very simple stimuli, and asked participants to report the position (Overgaard et al., 2006). Common to these studies is that perceiving only part of a stimulus would typically allow the participant to answer correctly (i.e. seeing part of a circle allows the participant to infer that it cannot have been a triangle). The task used by Szczepanowski et al. is markedly different from the tasks used in these above-mentioned previous studies and the application of PAS is not straightforward. For instance, it is possible to perceive several features of a face (e.g. the nose, part of the hair and eye-brows) without improving classification accuracy of emotional content. Similarly, the perception of just a few features (e.g. the eyes or the mouth) may improve classification accuracy significantly. The stimuli also differ simply in terms of their overall complexity. Accordingly, it may be too hasty to apply PAS in its "original form" to their study, and having applied it, it is difficult whether the results are caused by the emotional content of the stimuli, the overall stimulus complexity, or participants reporting on overall visual clarity rather than clarity of the relevant features.

Of course, we welcome very much the integration of subjective and objective measures and the use of PAS when relevant. We acknowledge that previous studies have not been entirely consistent in the use of PAS regarding stimulus material and instructions, and for this reason, we would like to make some recommendations for future experiments:

1. If necessary and when in doubt to re-do the entire calibration procedure and create a new scale as explained in Ramsøy and Overgaard (2004) rather than to just "import" the 4-point scale.
2. If one decides not to re-do the original procedure, always include (1) a full instruction explaining the meaning of each scale point in detail, (2) a pilot test with a good amount of trials (e.g. 30–50) in which the experimenter interrupts the subject frequently to ask about the use of the individual scale points (e.g. "I noticed you just reported "brief glimpse" – why did you do that/what did you mean with that/how would you define brief glimpse?").

If one does not do 1 or 2, we would be very reluctant to use PAS to test the correlation between subjective experience and any other measure.

4. Conclusions

Overall, we disagree with the claim of Szczepanowski et al. that CR is more exhaustive than PAS when using emotional stimuli. From a statistical perspective, we argue that the data needs to be appropriately modelled and formal tests are needed to support the conclusions. Using logistic regression, we fail to reject the null hypothesis that PAS and CR predict accuracy ratings to the same extent, and neither are we able to reject the null hypothesis that the two scales indicate the same amount of subliminal perception when participants report to be guessing or have no visual experience of the stimulus – in fact, we found no evidence for subliminal perception for either scale. In contrast, we find evidence to support that both scales are more exhaustive than PDW. We therefore see the study mainly as evidence that PDW performs poorly when using stimuli with emotional content. Based on this finding, we judge that there is now convincing evidence that PDW should only be used when participants are unable to use other, more direct measures such as CR or PAS (e.g. in studies using non-human animals), and when doing this, analyses should take loss aversion into account.

Turning to the procedures of the study, we are concerned that PAS was used in a manner that is somewhat different from how it was originally intended. The idea behind PAS was that it should be a scale that reflects the way that participants prefer to report. Therefore, the current PAS should not be seen as a single scale to be used in all tasks and for all stimuli, but rather as something to be used in visual identification tasks using simple stimuli. In the present study, PAS was used with somewhat more complex stimuli (faces) with several features of which only a small part were relevant (those separating fearful and neutral expressions). We are also concerned that the instructions to participants may not have clearly distinguished subjective visibility and certainty and that the scale points may not have been adequately explained. We recommend that ideally (though not always practically possible), it should be tested how participants prefer to report their visual experience for the particular stimulus type used. If this is not possible, we recommend that the experimenter thoroughly informs the participants of the distinction between certainty of accuracy and clarity of visual experience as well as how participants of previous experiments have defined and explained the meaning of the scale steps.

References

Clifford, C., Arabzadeh, E., & Harris, J. A. (2008). Getting technical about awareness. *Trends in Cognitive Sciences, 12*(2), 54–58. http://dx.doi.org/10.1016/j.tics.2007.11.009.

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. http://dx.doi.org/10.1037/0278-7393.21.5.1322.

Dienes, Z., & Seth, A. K. (2010). Measuring any conscious content versus measuring the relevant conscious content: Comment on Sandberg et al.. *Consciousness and Cognition*. http://dx.doi.org/10.1016/j.concog.2010.03.009.

Overgaard, M. (2012). Blindsight: Recent and historical controversies on the blindness of blindsight. *Wiley Interdisciplinary Reviews: Cognitive Science, 3*(6), 607–614. http://dx.doi.org/10.1002/wcs.1194.

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. http://dx.doi.org/10.1371/journal.pone.0003028.

Overgaard, M., Nielsen, J. F., & Fuglsang-Frederiksen, A. (2004). A TMS study of the ventral projections from V1 with implications for the finding of neural correlates of consciousness. *Brain and Cognition, 54*(1), 58–64.

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), 700–708. http://dx.doi.org/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), 1287–1296. http://dx.doi.org/10.1098/rstb.2011.0425.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10*(2), 257–261. http://dx.doi.org/10.1038/nn1840.

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences, 3*(1), 1–23. http://dx.doi.org/10.1023/B:PHEN.0000041900.30172.e8.

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*. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.concog.2009.12.013.

Schurger, A., & Sher, S. (2008). Awareness, loss aversion, and post-decision wagering. *Trends in Cognitive Sciences, 12*(6), 209–210 (author reply 210. doi: http://dx.doi.org/10.1016/j.tics.2008.02.012).

Schwiedrzik, C. M., Singer, W., & Melloni, L. (2011). Subjective and objective learning effects dissociate in space and in time. *Proceedings of the National Academy of Sciences, 108*(11), 4506–4511. http://dx.doi.org/10.1073/pnas.1009147108.

Seth, A. K. (2008). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and Cognition, 17*(3), 981–983. http://dx.doi.org/10.1016/j.concog.2007.05.008.

Szczepanowski, R., Traczyk, J., Wierzchoń, M., & Cleeremans, A. (2013). The perception of visual emotion: Comparing different measures of awareness. *Consciousness and Cognition, 22*(1), 212–220. http://dx.doi.org/10.1016/j.concog.2012.12.003.

Timmermans, B., Sandberg, K., Cleeremans, A., & Overgaard, M. (2010). Partial awareness distinguishes between measuring conscious perception and conscious content: Reply to Dienes and Seth. *Consciousness and Cognition, 19*(4), 1081–1083. http://dx.doi.org/10.1016/j.concog.2010.05.006.

Wierzchoń, M., Asanowicz, D., Paulewicz, B., & Cleeremans, A. (2012). Subjective measures of consciousness in artificial grammar learning task. *Consciousness and Cognition, 21*(3), 1141–1153. http://dx.doi.org/10.1016/j.concog.2012.05.012.