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

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.

© 2013 Elsevier Inc. All rights reserved.

## 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 blind-sight 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

---

* 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).

(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.