Contents lists available at ScienceDirect

Consciousness and Cognition

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

# Measuring consciousness: Is one measure better than the other?

Kristian Sandberg [a,1], Bert Timmermans [b,1], Morten Overgaard [a,*], Axel Cleeremans [b]

[a] *Cognitive Neuroscience Research Unit, Hammel Neurorehabilitation and Research Center, Denmark*
[b] *Consciousness, Cognition, and Computation Group, Université Libre de Bruxelles, Belgium*

ARTICLE INFO

*Article history:*
Received 20 October 2009
Available online xxxx

*Keywords:*
Consciousness
Methodology
Perceptual Awareness Scale
Confidence ratings
Post-decision wagering
Subliminal perception

ABSTRACT

What is the best way of assessing the extent to which people are aware of a stimulus? Here, using a masked visual identification task, we compared three measures of subjective awareness: The Perceptual Awareness Scale (PAS), through which participants are asked to rate the clarity of their visual experience; confidence ratings (CR), through which participants express their confidence in their identification decisions, and Post-decision wagering (PDW), in which participants place a monetary wager on their decisions. We conducted detailed explorations of the relationships between awareness and identification performance, looking to determine (1) which scale best correlates with performance, and (2) whether we can detect performance in the absence of awareness and how the scales differ from each other in terms of revealing such unconscious processing. Based on these findings we discuss whether perceptual awareness should be considered graded or dichotomous. Results showed that PAS showed a much stronger performance-awareness correlation than either CR or PDW, particularly for low stimulus intensities. In general, all scales indicated above-chance performance when participants claimed not to have seen anything. However, such above-chance performance only showed when we also observed a correlation between awareness and performance. Thus (1) PAS seems to be the most exhaustive measure of awareness, and (2) we find support for above-chance performance in the absence of subjective awareness, but such unconscious knowledge only contributes to performance when we observe conscious knowledge as well. Similarities and differences between scales are discussed in the light of consciousness theories and response strategies.

© 2010 Published by Elsevier Inc.

## 1. Introduction

A systematic comparison of measures of subjective awareness is long overdue (see also Dienes and Seth (2009), Wierzchoń, Taraday, Hawrot, and Asanowicz (2009)) since such measures are currently widely used in consciousness research (for an overview, see Seth, Dienes, Cleeremans, Overgaard, and Pessoa (2008)). For instance, the search for the neural correlates of consciousness typically involves contrasting brain activation during task performance with and without awareness (Baars, 1988; see e.g. Christensen, Ramsøy, Lund, Madsen, and Rowe (2006), Lau and Passingham (2006); but see Lamme (2006), for a different view). In this paper, we compare three currently popular measures of subjective awareness and assess how well each correlates with performance in a masked identification task. The Perceptual Awareness Scale (PAS; (Ramsøy & Overgaard, 2004)) is a purely introspective measure that requires participants to indicate the clarity of their experience of a stimulus. Confidence ratings (CR; e.g. (Cheesman & Merikle, 1986; Dienes, Altmann, Kwan, &

[*] Corresponding author. Address: Cognitive Neuroscience Research Unit, Hammel Neurorehabilitation and Research Center, Voldbyvej 15, 8450 Hammel, Denmark.
*E-mail address:* mortover@rm.dk (M. Overgaard).
[1] Shared first-authorship.

1053-8100/$ - see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.concog.2009.12.013

Goode, 1995) require participants to indicate their confidence in their decisions. Finally, post-decision wagering (PDW; Persaud, McLeod, & Cowey, 2007) requires participants to place a monetary wager on the accuracy of their decisions (i.e., stimulus identification). All three measures potentially present substantial advantages over other methods aimed at assessing the relationships between awareness and task performance. In particular the measures can be collected almost concurrently with decisions and can thus be correlated with task performance on a trial-by-trial basis, hence addressing Shanks and St. John's "retrospective assessment" problem (1994). However, it is unclear which method is most sensitive, that is, which method shows the best relationship between task performance and self-reported awareness. Likewise, it is unclear which measure is most exhaustive, that is, which method reveals the most conscious processing (Reingold & Merikle, 1988).

Dienes et al. (1995) have proposed the "zero-correlation criterion" (see also Chan (1992)) and the "guessing criterion" as tests for such conscious and unconscious processing, i.e. how sensitive and exhaustive the measures are. When analyzing using the zero-correlation criterion one looks for correlations between performance (an objective measure) and self-reported awareness or confidence in being correct (a subjective measure) across different conditions of task difficulty (e.g. various stimulus durations). Any positive relationship between performance and awareness suggests the involvement of at least some conscious knowledge in determining performance. However, the involvement of conscious processes does not exclude the involvement of unconscious processes. To examine these, the "guessing criterion" is used. Using this, performance is assessed for those cases where participants claim to be guessing (that is, when they claim to be performing randomly). If participants' performance is at chance, there is no knowledge contributing to the task, unconscious or otherwise, and subjective and objective thresholds are identical. If, however, participants who claim to be guessing perform above chance, then one would conclude that their performance is based on knowledge they are not aware of possessing, that is, on unconscious knowledge. An important caveat to this reasoning is that above-chance performance can also be the consequence of the test failing to be exhaustive when subjects claim to be guessing, meaning that participants fail to be complete in their report about their conscious contents. Given this state of affairs, the best one can do is to consider that if one scale indicates less unconscious processing than another, then that scale should be taken to be more exhaustive than the others, all else being equal (that is, assuming that there are no differences in the extent to which each scale promotes awareness in and of itself, and in the extent to which the different scales erroneously labels some unconscious knowledge as conscious knowledge). One should thus look for the scale that is simultaneously most sensitive and most exhaustive. In other words, the most promising scale is the one that (a) shows better correlation than others between performance and awareness at different levels of difficulty (the zero-correlation criterion), and (b) shows the least above-chance performance for trials on which participants claim to be guessing (the guessing criterion).

The current study was thus motivated by two simple goals. First, we aimed at determining whether the three measures predict the same relative contribution of conscious and unconscious processing. To this end, we determined the relationship between performance and awareness at different levels of task difficulty. Additionally, we explored the extent to which each scale indicates the same level of above-chance performance in the absence of awareness, if any (for an overview of this debate, see Kouider and Dehaene (2007), or Overgaard and Timmermans, 2009). The second goal was to explore whether perceptual awareness should best be considered as graded or as dichotomous (e.g., Overgaard, Rote, Mouridsen, & Ramsøy, 2006; Sergent & Dehaene, 2004). Though there are theoretical complications, comparing the three scales in this light should be informative.

### 1.1. Three awareness scales

In the current experiment, we compare three scales, each of which measures awareness in a different way. Each of these scales has a number of claimed advantages and disadvantages. Even though some of these are difficult to validate empirically, they will be mentioned in the following as they may still influence the evaluation of the scales.

### 1.2. Perceptual awareness scale

When using PAS, participants report on the quality of their subjective experience directly. PAS was originally created by the participants in an experiment by Ramsøy and Overgaard (2004). In this experiment, participants were asked to describe the quality of their visual experience as they looked at briefly displayed stimuli, using a scale they had created themselves. It was suggested to participants that they start the scale with 'No experience' and ended it with 'A clear image', but they were free not to follow the suggestion and/or to use any number of categories. All five participants ended up using a 4-point scale with the elements (1) 'No experience', (2) 'Brief glimpse', (3) 'Almost clear image', and (4) 'Absolutely clear image'. Although the participants differed in their labeling of the categories, they agreed in their definitions of the categories.

PAS can be claimed to be intuitive in that the categories used are created not by an experimenter, but by other research participants evaluating their conscious experience (Ramsøy & Overgaard, 2004). In addition, as it is not related to a participant's evaluation of how good their answer is (as is post-decision wagering), PAS and other direct measures of conscious experience can easily be used in tasks in which there is no "correct" answer such as the perception of an ambiguous figure or binocular rivalry. Finally, Persaud and colleagues (2007) have argued that participants using numerical confidence ratings may withhold knowledge, as they have no motivation to reveal it. This criticism also applies to PAS.

Please cite this article in press as: Sandberg, K., et al. Measuring consciousness: Is one measure better than the other?. *Consciousness and Cognition* (2010), doi:10.1016/j.concog.2009.12.013

One claimed advantage of PAS is that the participants are asked directly to provide the information that experimenter is looking for, that is, their conscious experience. Paradoxically, this is also a claimed disadvantage as it depends on how good participants are at reporting their experience. If the participants are reasonably good introspectionists, then asking them to report directly minimizes the risk of confusion or errors that might arise if the participants are asked about something very different, and their conscious experience is inferred from their answer. However, if participants are poor introspectionists, then asking them to report on their experience is associated with a large risk. Thus, because it seems possible to argue both ways, the answer must be obtained empirically, in comparison with other scales.

### *1.3. Confidence ratings*

Confidence ratings (CRs) have been used either with respect to perception itself, in which case participants directly report their confidence in having perceived something (Bernstein & Eriksen, 1965; Cheesman & Merikle, 1984), or with respect to participants' performance, in which case they report their confidence in having provided a correct answer. In the former case, CRs closely resemble PAS. In the latter case, however, CRs are metacognitive judgments in which participants express the extent to which they are certain that their answer is correct in forced-choice tasks (Cheesman & Merikle, 1986) or in the discrimination tasks typical of implicit learning research (Dienes et al., 1995; Kuhn & Dienes, 2006). Although CRs may be expressed on very different scales, most variations include 'guessing' or 'no confidence' in the description of the lowest rating. Examples include dichotomous scales such as "guess/know" and "guess/anything else", as well as gradual scales.

In many ways, CR have the same advantages and disadvantages as scales that ask directly about conscious experience. Nevertheless, in light of the observation that participants may not be good introspectionists, CR may have an advantage, as participants are not asked directly to introspect. A potential challenge with having participants rate their own performance, however, is that while two participants may have a comparable clarity in their experience of a stimulus, they might use different criteria to decide themselves confident.

### *1.4. Post-decision wagering*

Post-decision wagering (PDW) is a recently suggested measure of conscious content (Persaud & McLeod, 2008; Persaud et al., 2007). After performing a task, the participants place a wager on having performed the task correctly. The rationale is that the wagers are based on the awareness of the participants, but the participants never need to introspectively report their awareness. According to Persaud and colleagues, they simply perform a task that requires awareness to be completed. For this reason, PDW has been put forward as an "objective" or direct measure. Persaud and colleagues used wagering dichotomously so that participants could place either a low or a high wager at even odds on one of two possibilities. The degree to which a participant maximizes his gains through advantageous wagering (betting high after a correct decision, or low after an incorrect decision) is assumed to be indicative of conscious experience. More importantly however, since the possibility to gain (real or imaginary) money provides participants with a strong incentive to reveal any conscious knowledge they may possess, failure to wager advantageously should reflect absence of awareness in a more exhaustive manner that other measures.

Despite being intuitive as well as potentially exhaustive, post-decision wagering as a method to assess awareness has been questioned from a theoretical point of view. The claim that PDW is a direct measure of awareness has been questioned by Seth (2008) who argues that it is in fact a second-order judgment of the reliability of a first-order experience. Such a "metacognitive comment" does not exhaustively describe the rich phenomenology of conscious experience and metacognitive competences are susceptible to biases (see Persaud, McLeod, and Cowey (2008), for a reply to the critique). In addition, PDW (as applied by Persaud and colleagues) seems to presuppose that conscious experience is dichotomous. If conscious experience is not dichotomous, however, a problem arises in that criterion setting about when to start wagering high may vary significantly between participants, as recently argued by Clifford, Arabzadeh, and Harris (2008). As a consequence, it is impossible to ascertain, on a trial-by-trial basis, whether a participant was conscious or not, as a low wager is not necessarily synonymous with absence of awareness – participants could be reluctant to take a risk even though they have a vague experience of the stimulus. The influence of risk aversion on wagering behavior was indeed recently confirmed empirically by Dienes and Seth (2009).

Additionally, Clifford and colleagues (2008) argue that when wagering is used as it is by Persaud and colleagues, the optimal wagering strategy is always to bet high, as this will give the same outcome if accuracy is at chance, but a higher outcome if accuracy is above chance, whether above-chance accuracy is subtended by unconscious knowledge or not. This behavior was not observed in the previously mentioned experiments, and Schurger and Sher (2008) report that only 2 out of more than 100 tested participants adopted an optimal strategy even when encouraged to wager high. These results seriously question the claim that wagering is intuitive, but it remains to be seen empirically if it is a more substantial issue for wagering scales than it is for the other scales.

It is often preferable to have a measure of subjective experience that does not alter task accuracy. The possibility of monetary profit, however, has been shown to improve task accuracy, whereas this is not the case for CR (cf. Persaud & McLeod, 2008). When used in neuroscientific experiments, this is a problem, particularly if improved accuracy is caused by emotional arousal. For instance, one might imagine that emotional arousal is different for betting high vs. betting low, and this would make it difficult to differentiate neural activity related to awareness from activity related to emotional arousal.

## 2. Experiment 1

The main goal of the experiment was to examine if the three measures allow us to draw the same conclusions about the relationship between performance and awareness. To this end, we had participants perform a masked stimulus identification task. After each response, participants had to perform one of three judgments: (1) rate stimulus visibility (PAS), (2) give a confidence rating in their answer, or (3) place a post-decision wager.

### *2.1. Method*

#### *2.1.1. Participants*

Thirty-six healthy participants (18 at Université Libre de Bruxelles, Belgium, 18 at Aarhus University, Denmark) participated in the experiment and were assigned randomly to one of three conditions defined by which measure of awareness is used. This design resulted in three groups of 12 participants: a PAS group, a CR group, and a PDW group. Mean age was 23.9 years (22.3–25.5). Age did not differ between groups, $F(32,2) = 0.6$, $p = .557$. All participants had normal or corrected-to-normal vision.

#### *2.1.2. Procedure*

Participants performed a visual identification task. At the onset of each trial, a fixation mark appeared on a computer screen for 500, 1000, 1500 or 2000 ms. The fixation mark was followed by one of four geometrical shapes (circle, square, diamond, and triangle), shown for one of 12 durations (range: 16–192 ms in steps of 16 ms on a 60 Hz CRT screen). The stimulus was then masked by a figure consisting of all four possible shapes (Fig. 1). The mask remained on the screen until participants had responded or until 3000 ms had elapsed (participants rarely responded later than 2000 ms from stimulus onset). The task was to identify the displayed shape by pressing one of four keys (‘c’, ‘v’, ‘b’, ‘n’) as fast and accurately as possible (using the index and middle finger of each hand), prioritizing accuracy over speed. After participants had responded, a graphical representation of one of the three scales (PAS, CR, and PDW) appeared on screen, and participants were asked to indicate their response using the left- and right-arrows on the keyboard. All scales were displayed as a bar divided into four equally large segments. A number and a description were displayed below each segment. For PAS, the descriptions were: (1) No experience, (2) A vague experience, (3) An almost clear experience, and (4) A clear experience. For CR, the descriptions were: (1) Not confident at all, (2) Slightly confident, (3) Quite confident, and (4) Very confident. For PDW, the descriptions were: (1) €5, (2) €10, (3) €15, and (4) €20 (or similar amounts in Danish kroner). The only difference between groups was thus the scale-specific instruction and the descriptions that appeared on screen when using the 4-point scale. Imaginary money was used for PDW participants (as has been done previously by Persaud et al. (2007)) so as not to introduce additional differences between conditions. Each participant used one scale throughout the experiment to avoid that the use of one scale would affect ratings on another scale.

Participants performed the experiment individually, and all instructions appeared on the screen. The experiment began with a practice block consisting of 48 trials, with each of the 12 stimulus durations used four times and the longest durations appearing first. Each shape was presented four times in a random order. Next followed five experimental blocks for a total of 336 trials. In all experimental blocks, stimulus duration and shape were randomized and did not coincide in a systematic way. Due to further experimental goals not discussed in the present article, these five blocks differed in the following manner. The first experimental block consisted of 96 trials, with each stimulus duration presented eight times. Blocks

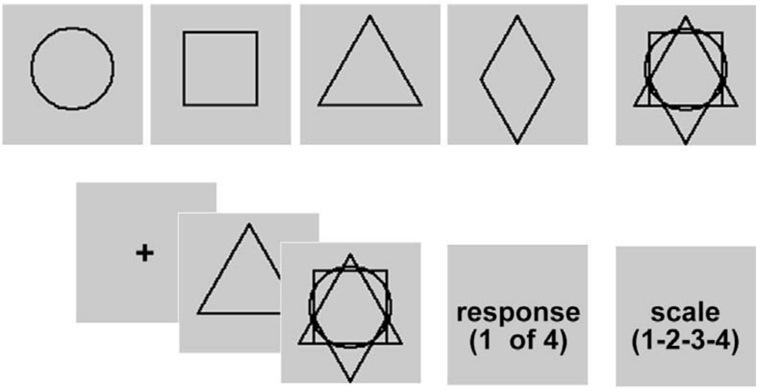

**Fig. 1.** Trial procedure: a fixation mark (500–2000 ms) was followed by a stimulus consisting of one of four possible geometrical shapes (possible durations 16–192 ms in steps of 16 ms). The figure was masked by a figure consisting of all four shapes (remained on screen until response or 3000 ms had passed). Participants’ task was to report the displayed shape as fast and accurately as possible, prioritizing accuracy over speed. After the response, participants indicated their awareness using either the Perceptual Awareness Scale (PAS), a confidence rating scale (CR), or a post-decision wagering scale (PDW).

Please cite this article in press as: Sandberg, K., et al. Measuring consciousness: Is one measure better than the other?. *Consciousness and Cognition* (2010), doi:10.1016/j.concog.2009.12.013

2–4 consisted of 48 trials each. One of these blocks included only "difficult" stimuli (stimulus duration: 16–96 ms); another only "average" stimuli (stimulus duration: 64–144 ms), and the third only "easy" stimuli (stimulus duration: 112–192 ms) stimuli. Half of the participants were exposed to one difficult, one average, and one easy block. The other half of participants was exposed to three blocks of the same difficulty. In the first case, each of six possible block orders appeared once within each of the three groups. In the second case, each of the three possible block orders appeared twice within each group. The final, fifth, block was identical to the first. Stimulus durations were randomized in each block. Every block order appeared the same number of times within each group, and any effect of block order can therefore be assumed to be neutralized between groups. Therefore, neither block type nor block order will be further discussed.

## *2.2. Results*

The data were analyzed using R version 2.8.1 (R Development Core Team). We assessed which scale was the most exhaustive, that is, we explored whether scales indicated similar subjective thresholds, comparing their sensitivity to participants' awareness by means of the zero-correlation criterion and the guessing criterion. The scale indicating awareness at the lowest stimulus duration can be considered to be the most exhaustive of the set. We also looked at how the scales' sensitivity to awareness was comparable across stimulus durations. Results are depicted in Figs. 2 and 3.

### *2.2.1. Response distribution*

Before analyzing the relationship between awareness and performance, we looked at how the participants used the scales. As can be seen in Fig. 2, response distributions are comparable overall, and all scale points are used on all scales. However, PAS appears to have been used in a more gradual manner than CR. Likewise, CR responses also appear to be more distributed than PDW responses. Crucially, for both CR and PDW the scale extremes "1" and "4" are always used more than the others regardless of stimulus duration (for PDW, scale points "2" and "3" are in fact only rarely used). For PAS, scale points "2" and "3" are used more frequently than any other scale point at 80 ms and 96 ms durations, respectively. This suggests that, using PAS, participants frequently report that they had a "vague experience" – nothing more, and nothing less – at certain stimulus durations.

### *2.2.2. Scale exhaustiveness*

We analyzed the *zero-correlation criterion* by means of logistic regression, which is similar to calculating Chan difference scores (Dienes et al., 1995). By basing the regression model on all data points, we avoided loss of statistical power associated with previously employed methods such as calculating a single gamma correlation score for each participant and comparing these scores across groups (e.g. Kuhn & Dienes, 2006).

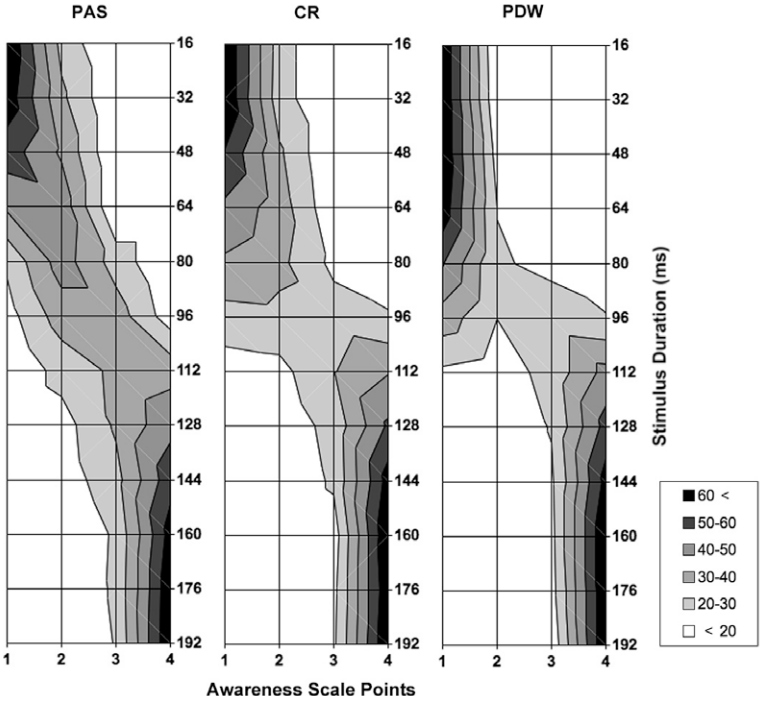

**Fig. 2.** Contour plots of awareness scale response distributions across stimulus durations, for each of the scales. Shadings indicate percentage of trials.

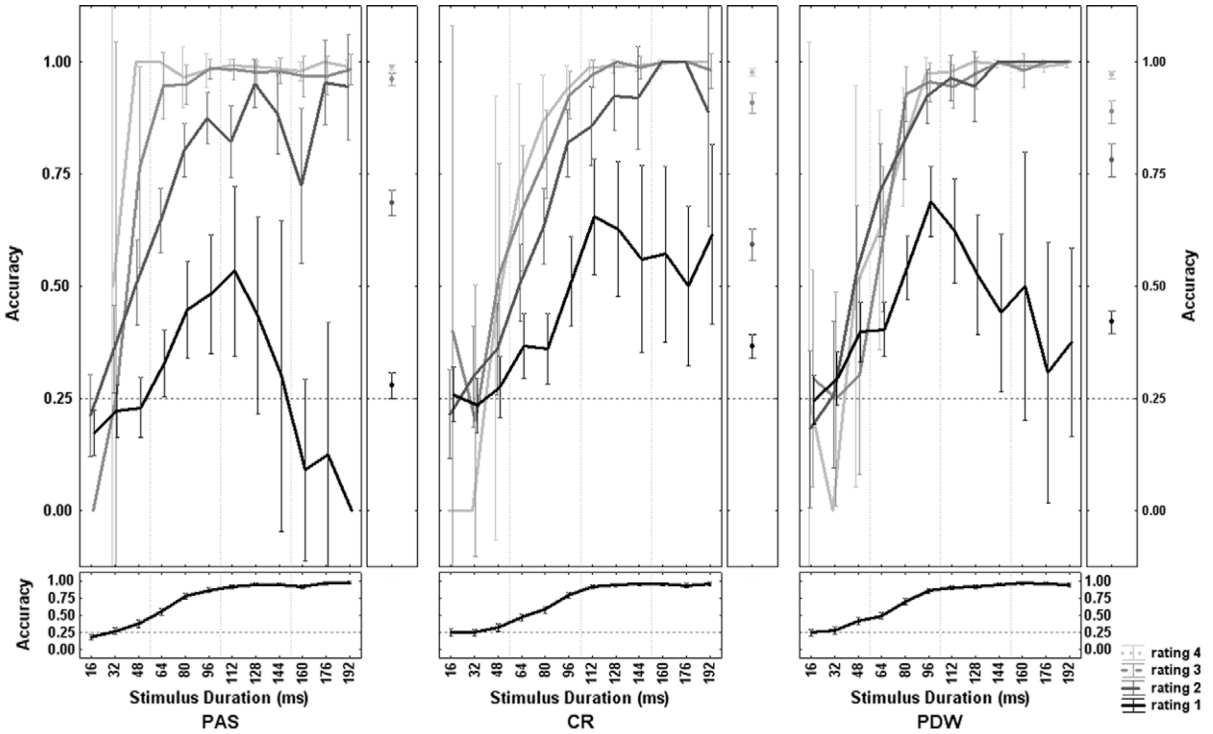

**Fig. 3.** Average accuracy (proportion of correct trials) as a function of stimulus duration, for each awareness scale point of each of the three scales. Small panels to the bottom and the right of the main panel depict marginal means across scale points and stimulus durations, respectively. All error bars represent 95% Confidence Intervals. The horizontal dotted line is chance level (25%).

**Table 1**
Regression coefficients for the logistic regression mixed model for accuracy.

| *N* = 36, # observations = 12029 | Coefficient | SE | *z* | Odds ratio |
|---|---|---|---|---|
| Random effect of subject (intercept): variance = 0.43 (SD = 0.66) | | | | |
| *Main effects (fixed)* | | | | |
| 1a. Scale: PAS vs. CR | −0.414 | 0.300 | −1.38 | 0.66 |
| 1b. Scale: PAS vs. PDW | −0.105 | 0.301 | −0.35 | 0.90 |
| 1c. Scale: CR vs. PDW | 0.309 | 0.303 | 1.02 | 1.36 |
| 2a. Stimulus duration for PAS | 0.020 | 0.002 | 9.43[a] | 1.39° |
| 2b. Stimulus duration for CR | 0.038 | 0.002 | 15.08[a] | 1.82° |
| 2c. Stimulus duration for PDW | 0.035 | 0.002 | 14.68[a] | 1.76° |
| 3a. Awareness rating for PAS | 1.838 | 0.098 | 18.79[a] | 6.29 |
| 3b. Awareness rating for CR | 1.340 | 0.079 | 16.90[a] | 3.82 |
| 3c. Awareness rating for PDW | 1.372 | 0.078 | 17.50[a] | 3.94 |
| *Two-way interaction effects (fixed)* | | | | |
| 4a. Stimulus duration ∗ awareness rating for PAS | 0.004 | 0.002 | 2.03[c] | 1.06° |
| 4b. Stimulus duration ∗ awareness rating for CR | 0.017 | 0.002 | 9.30[a] | 1.32° |
| 4c. Stimulus duration ∗ awareness rating for PDW | 0.013 | 0.002 | 7.33[a] | 1.22° |
| 5a. Stimulus duration ∗ scale: PAS vs. CR | 0.017 | 0.003 | 5.18[a] | 1.32° |
| 5b. Stimulus duration ∗ scale: PAS vs. PDW | 0.015 | 0.003 | 4.57[a] | 1.27° |
| 5c. Stimulus duration ∗ scale: CRvs. PDW | −0.002 | 0.003 | −0.68 | 0.96° |
| 6a. Awareness rating ∗ scale: PAS vs. CR | −0.498 | 0.126 | −3.96[a] | 0.61 |
| 6b. Awareness rating ∗ scale: PAS vs. PDW | −0.467 | 0.125 | −3.72[b] | 0.63 |
| 6c. Awareness rating ∗ scale: CR vs. PDW | 0.032 | 0.112 | 0.28 | 1.03 |
| *Three-way interaction effect (fixed)* | | | | |
| 7a. Stimulus duration ∗ awareness rating ∗ scale: PAS vs. CR | 0.013 | 0.003 | 5.12[a] | 1.24° |
| 7b. Stimulus duration ∗ awareness rating ∗ scale: PAS vs. PDW | 0.009 | 0.003 | 3.47[b] | 1.15° |
| 7c. Stimulus duration ∗ awareness rating ∗ scale: CR vs. PDW | −0.005 | 0.003 | −1.86 | 0.93° |

*Notes*: Stimulus duration and awareness rating were mean centered. Scale was a 3-level factor. °, change in odds for 16 ms change (=OR[1 ms]^16). Coefficient confidence intervals are not yet operationalized for mixed effects logistic regression.

[a] $p < .0001$.

[b] $p < .001$.

[c] $p < .05$.

Please cite this article in press as: Sandberg, K., et al. Measuring consciousness: Is one measure better than the other?. *Consciousness and Cognition* (2010), doi:10.1016/j.concog.2009.12.013

### 2.2.3. The model

We created a logistic regression mixed model (using the R lme4 package) for accuracy, with scale, awareness rating, stimulus duration, and all 2- and three-way interactions as predictors. Scale was coded as a 3-level factor, whereas the two other predictors were entered as continuous variables, centered on their mean. A potential excess variation between participants compared to within participants was taken into account by including a random subject effect in the model, on the intercept. The full interaction model explained 42% of the Accuracy variance, $R^2 = .42$, $F(8,12020) = 1089$, $p < .0001$), which was significantly better than both the null model containing only a constant and a random subject effect, $\chi^2(11) = 5806$, $p < .0001$, and than a model containing only main effects and the random subject effect, $\chi^2(7) = 248$, $p < .0001$. Nevertheless, as only 42% of the variance is explained, we should treat the results with caution, especially since between-subject variance is considerable (0.43; SD = 0.68). Regression results are detailed in Table 1 (Numbers preceded by the letter "E" in the text below refer to the main Effect with that number in the table – so E1 would refer to main effect 1: scale).

### 2.2.4. Awareness as a predictor of performance

As expected and required, accuracy did not differ between scales (E1). Also as expected, both stimulus duration and awareness ratings significantly predicted accuracy (E2 and E3). Crucially, they differed in the extent to which they did so for the three scales (see below). Awareness rating and stimulus duration also interacted significantly with each other (E4) showing that, for all scales, the degree to which awareness ratings predict accuracy is not the same across stimulus durations, as clearly shown in Fig. 3. This effect only just reached significance for PAS (E4a), suggesting that for this scale, the relationship between awareness rating and accuracy remains more consistent across stimulus durations than for CR and PDW (E4b,c), between which there is no difference. This is confirmed by the three-way interaction (E7). Furthermore, in terms of differences between scales, the two-way interactions showed that PAS differs significantly from both CR and PDW in how stimulus duration and awareness ratings predict accuracy (E5a,b; E6a,b), whereas CR and PDW do not differ (E5c, E6c). With respect to stimulus duration, for PAS, an increase of 16 ms in stimulus duration resulted in a 1.39 change in odds ratio for a correct response (E1c), whereas for CR and PDW such an increase resulted in a change in odds ratio that is 32% or 27% higher, respectively (E5a,b). However, with respect to the awareness rating, for PAS, an increase of 1 scale point resulted in a 6.29 change in odds ratio for a correct response (E3a), whereas for CR and PDW this increase produced a change in the odds ratio that was only 61% and 63% of PAS's, respectively (E6a,b). This suggests that for PAS, as compared to CR and PDW, awareness ratings are relatively more predictive than stimulus duration for accuracy. In other words, for PAS, a stimulus has to be shown for 90 ms longer in order to produce the same shift in odds ratio than 1 awareness scale point increase; for CR and PDW, a stimulus that is shown for 36 ms longer already produces a bigger shift in odds ratio than 1 awareness scale point increase.

To summarize, the extent to which awareness scores predict accuracy differs between PAS and the other scales, with PAS scores being more predictive. However despite the fact that these predictions are more consistent across stimulus durations for PAS than for the other scales, predictability of all scales varies across stimulus durations.

### 2.2.5. Predictability across stimulus durations

Visual observation of Fig. 3 makes it clear that while all scales share an overall pattern, there are pronounced differences depending on stimulus duration. First of all, for all scales, accuracy levels do not remain stable for each awareness scale point, but instead rise. Specifically, for longer stimulus durations (112 ms and longer), the prediction of all scales' awareness ratings dichotomizes, with lower accuracy corresponding to the lowest rating ("1", not everywhere at chance: see guessing criterion analysis), and near 100% accuracy corresponding to the other three ratings ("2" through "4"). Only for PAS does "2" predict a slightly below 100% accuracy level, $\chi^2(1) = 101.4$, $p < .0001$, which most likely indicates that subjects are able to distinguish some "not quite clear" experiences from the very clear experiences they have for most stimuli at these high durations. At no point on any scale do we distinguish an accuracy difference between rating "3" and rating "4", suggesting either that a 3-point scale is more appropriate in the context of the current task, or perhaps that only a certain degree of awareness is required to be able to respond correctly, but that the clarity of the subjective experience can still increase after this point is reached.

For shorter stimulus durations (96 ms and shorter) we see the scales displaying more divergent patterns. First, if we look at the point at which accuracy for rating "1" is clearly below accuracy for any of the other ratings, we see that for PAS accuracy for "1" and "2" differs from 32 ms onward, $\chi^2(1) = 5.69$, $p = .017$. For CR accuracy for "1" and "3" differs from 48 ms onward, $\chi^2(1) = 3.98$, $p = .046$, and for "1" and "2" from 64 ms onward, $\chi^2(1) = 5.62$, $p = .018$. For PDW, accuracy for "1" and "2" differs from 64 ms onward, $\chi^2(1) = 21.8$, $p < .0001$. Second, when we look at the stimulus duration for which accuracy for either "3" or "4" (taken together to minimize occurrence of low expected frequencies) differs from accuracy for "2", we see that for PAS, this is almost the case for 48 ms, $\chi^2(1) = 3.49$, $p = .062$, and clear-cut at 64 ms, $\chi^2(1) = 16.1$, $p < .0001$; the distinction disappears from 176 ms onwards (even though from 112 ms onwards the significant differences become unreliable due to accuracy ceiling effects). For CR, accuracy for "2" differs from "3" or "4" at 64 ms, $\chi^2(1) = 4.97$, $p = .026$, but this distinction disappears from 160 ms onward (unreliable differences from 112 ms onward). For PDW, this distinction never occurs, as awareness remains dichotomous with respect to accuracy levels for all stimulus durations. Furthermore, from 64 ms onward, PAS's "3" and "4" ratings correspond to a >95% accuracy, whereas "3" or "4" confidence ratings or wagers correspond to >95% accuracy from 112 ms and 96 ms onward, respectively.

To summarize, participants using PAS were better able to distinguish between different conscious experiences than subjects using CR or PDW; PDW participants performing particularly poorly. These results were most clear for low stimulus durations, but continued to be present for higher stimulus durations.

To assess the data through the *guessing criterion*, we performed a chi-square test of goodness-of-fit to determine whether the lowest awareness ratings ("1") of the three scales yielded comparable chance accuracy levels (25% correct discriminations expected). Accuracy corresponding to "1" differed between scales, $\chi^2(2) = 50.4$, $p < .0001$, with the difference between PAS and both others being most pronounced, $\chi^2(1) > 15$, $p < .0001$, and CR and PDW differing slightly less between each other, $\chi^2(1) = 8.17$, $p = .0043$. Subsequent chi-square tests revealed that overall, accuracy differed from chance for all scales, albeit for PAS (27.9%) only just, $\chi^2(1) = 4.21$, $p = .040$; CR (36.6%) $\chi^2(1) = 86.3$, and PDW (42.0%) $\chi^2(1) = 229.4$, both $p < .0001$. If we examine short stimulus durations only, we see that for PAS, performance corresponding to "1" reliably rises above chance from 80 ms onward, $\chi^2(1) = 6.65$, $p = .0099$; for CR, "1" performance surpasses chance from 64 ms onward, $\chi^2(1) = 5.21$, $p = .022$; for PDW, "1" performance surpasses chance from 48 ms onward, $\chi^2(1) = 9.73$, $p = .0018$ (though note that this difference partly arises from that fact that, due to more distributed use of all scale points, PAS has significantly less observations in the "1" category than CR, and CR much less than PDW).

To summarize, none of the scales indicate consistent chance performance when people give the lowest awareness rating, but PAS seems to fare better than the others, indicating less unconscious processing overall, and indicating it with a later onset than the other scales. Again, the worst results were obtained for PDW.

## 2.3. Discussion

In the above reported experiment, a number of differences between the scales could be identified, as well as some general similarities. As scales are compared between groups (to avoid inter-scale contamination) part of the differences could be claimed to be caused by subtle differences between the groups. However, due to the size of the groups (12 participants in each) and the fact that neither age nor task accuracy differed significantly between groups, any impact of group differences seems modest.

### 2.3.1. Scale comparison

In terms of differences, the data suggest that PAS is the most exhaustive scale, as it indicated the presence of more conscious processing than the other two scales by the zero-correlation criterion, and less unconscious processing by the guessing criterion. PAS also appears to be more sensitive to different levels of awareness, each corresponding to different accuracy levels. Looking at response distributions, this seems to be the consequence of participants using the PAS scale in a more gradual manner than either CR or PDW. PDW, which has been claimed to provide an extra incentive for people to reveal their knowledge, fares worst in the reported experiments. Not only is it the least sensitive of all three scales to small variations of experience, but it also appears to promote binary decisions with respect to accuracy in general: participants either wagered very low or very high. CR falls somewhere between PAS and PDW, more closely resembling the latter.

On no scale did a single scale point relate to a specific level of accuracy across stimulus durations (meaning that "a vague experience" or "slightly confident" did not predict the same task performance in different contexts). It is worth noting, however, that the accuracy level corresponding to a specific PAS rating varied less as a function of stimulus duration than for the other scales. The relationship between accuracy and awareness ratings were thus better generalized across conditions for PAS than CR and PDW.

In the current experiment, all differences were found using similar 4-point scales that were displayed in a similar manner to the participants, and all were presented with identical stimulus sets. Thus, it must be the instructions given to participants about how to report that are responsible for the observed differences. Crucially, participants using PAS were asked to rate the clarity of their visual experience, not their beliefs about being correct. For difficult stimuli, participants reporting only the clarity of their visual experience may be able to differentiate different degrees of clarity before they are willing to commit to a statement of certainty in their answer, as is required for participants using a confidence or wagering scale. Participants using the confidence or wagering scales thus seem to withhold reports of certainty when they were fairly uncertain. On the other hand, overall changes in the proportion of low vs. high confidence ratings or wagers changed in relation to their accuracy in discriminating the stimuli. This suggests that participants were actually reporting confidence in being correct and not the clarity of their visual experience. For easier stimuli, participants keep differentiating between degrees of clarity even though they had reached maximum task accuracy, and this differentiation has predictive power with respect to accuracy, as shown in the slightly lower accuracy for "2" ratings than for "3" or "4" ratings. Thus, the data suggest that a certain degree of clarity is needed to be able to discriminate one figure from another, but this does not mean that the visual experience cannot become more vivid after this point is reached.

### 2.3.2. Performance without awareness and unconscious contributions

An important result of this study is that all scales indicate the presence of task performance without awareness. As mentioned earlier however, this is not conclusive evidence for the existence of unconscious processing in our experimental paradigm. Our calculations of relative exhaustiveness of the scales seem valid, but that does not mean that the most exhaustive scale is fully exhaustive. Nevertheless, examining where the scales indicate the presence of unconscious processing seems relevant. Inspecting Fig. 3, we see that when average accuracy is at chance (before around 48 ms), awareness ratings are not very meaningful in the sense that all ratings were related to chance or near-chance accuracy. This is not surprising. However as average accuracy increases for higher stimulus durations, awareness ratings become related to increasingly different levels of accuracy – for instance, accuracy associated with a rating of "1" becomes increasingly different from the accuracy associated with a rating of "4". According to the zero-correlation criterion, this would indicate an increase of awareness, which is indeed what the above analysis shows. However, along with the increase in average accuracy comes an increase in accuracy for ratings of "1" – though perhaps delayed by some 15–30 ms of stimulus duration. According to the guessing criterion, this indicates the emergence of unconscious processes as well. When accuracy asymptotes at stimulus durations of about 128 ms, the influence of unconscious processes peaks as well (as inferred from the guessing criterion) and then seems to stabilize (for CR) or decrease to a lesser (for PDW) or greater (for PAS) extent. Assuming full exhaustiveness of the most exhaustive scale (which of course may not be the case), this observation indicates that unconscious and conscious processes influencing task accuracy appear at around the same time, though perhaps slightly later for the unconscious processes. It also indicates that the influence of unconscious processes on task performance increases until the task becomes very easy, at which point their influence gradually decreases.

### *2.3.3. Graded vs. dichotomous consciousness*

The experiment also contributes to the debate of whether perceptual consciousness is graded or dichotomous. Our results suggest the former (at least within the present experimental setup). First, participants use all scale points on all of the three scales, and not just scale extremes. Especially for PAS, and to a lesser degree for CR, participants do use mostly intermediate awareness ratings to rate stimuli of specific intermediate durations. These intermediate awareness ratings clearly reflect a different level of processing, as they are correlated with different levels of performance. Therefore, depending on their duration, stimuli seem to be processed differently, as reflected not only in a gradual increase in average awareness rating, but also in a gradual shift across scale point use.

Second, the correlation between awareness ratings and accuracy does not arise abruptly. Typically, first the lower two ratings start to yield different performance predictions, after which the third and fourth rating start to reflect performance differing from the other ratings. This would also seem to indicate that consciousness is gradual to some extent.

## 3. Conclusion

Going with the tacit assumption that objective measures should be preferred over subjective (i.e. introspective) ones when studying consciousness, one would expect results from the present experiment to advocate the use of PDW over CR, which again should be preferred over PAS, given that it is the most "subjective" of the three. However, overall, PDW proved to be less sensitive than PAS for stimuli that were hard to identify, with CR taking up an intermediate position. Thus, the present experiment did not find any support for the claimed advantages of using wagering as a measure of conscious experience. In fact, on the basis of the present results, PDW can be argued to perform worse than either of the other two competing methods. Specifically, the claim that participants are more likely to reveal information when they are in a position to profit from it has not been confirmed in previous experiments – rather, participants seem more reluctant to express their awareness, as they risk losing their wager. In the present experiment, this pattern was confirmed despite our using imaginary money. Furthermore, comparing PAS and CR, PAS seemed to constitute a better predictor of task accuracy when the stimuli are difficult to identify. Finally, PAS ratings were more consistently related to particular levels of accuracy than ratings on the other scales.

One parsimonious way of interpreting our results is the following: Participants tend to do exactly as instructed. Thus, when participants are asked to report their confidence, they do exactly that – they do not give a report about their experience of the stimulus. It just so happens that cases of feeling confident of, say, the contents of a perceptual event empirically correlate well with an experience of that content. This correlation, however, is not perfect, especially not in the case of very difficult stimuli. Likewise, when participants are asked to place a wager on their decision, this is what they will do. They will, basically, perceive the task as a gambling situation rather than issue a veridical report about their experience of the stimulus. In such a situation, factors such as emotional arousal (Persaud & McLeod, 2008), risk aversion (Dienes & Seth, 2009) and gambling strategy may influence both task performance and awareness ratings. Asking participants directly about the content of their experience thus seems to offer a much more direct way of getting information about conscious content. This, obviously, leaves direct introspective methods as the most promising ones. It should be noted, however, that our results about PDW were obtained with a particular version of PDW closely resembling the proposed version of Persaud and colleagues (2007). The present experiment would of course not allow us to conclude about the effectiveness of any improvements to the method made later to decrease the impact of, for instance, loss aversion. On the other hand, any such improvement would need to demonstrate improvements in exhaustiveness compared not only to traditional PDW, but also to other measures of awareness.

To conclude, our results highlight the simple, but important fact that the specific manner in which one measures awareness matters: not only is it the case that different measures tap into slightly different aspects of what it means to be aware of a particular state of affairs, but the measures also interact in subtle ways with stimulus difficulty and performance. These considerations further have theoretical import in that our concept of consciousness is very much determined by our measures of it. In a sense thus, what you get is what you measure.

## References

Baars, B. (1988). *A cognitive theory of consciousness*. New York: Cambridge University Press.

Bernstein, I. H., & Eriksen, C. W. (1965). Effects of "subliminal" prompting on paired-associate learning. *Journal of Experimental Research in Personality, 1*(1), 33–38.

Chan, C. (1992). Implicit cognitive processes: Theoretical issues and applications in computer systems design. Unpublished doctoral dissertation, University of Oxford, Oxford, England.

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception and Psychophysics, 36*(4), 387–395.

Cheesman, J., & Merikle, P. M. (1986). Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology Revue Canadienne de Psychologie, 40*(4), 343–367. doi:10.1037/h0080103.

Christensen, M. S., Ramsøy, T. Z., Lund, T. E., Madsen, K. H., & Rowe, J. B. (2006). An fMRI study of the neural correlates of graded visual perception. *NeuroImage, 31*(4), 1711–1725. doi:10.1016/j.neuroimage.2006.02.023.

Clifford, C., Arabzadeh, E., & Harris, J. (2008). Getting technical about awareness. *Trends in Cognitive Sciences, 12*(2), 54–58. doi: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. doi:10.1037/0278-7393.21.5.1322.

Dienes, Z., & Seth, A. (2009). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*. doi:10.1016/j.concog.2009.09.009.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362*(1481), 857–875. doi:10.1098/rstb.2007.2093.

Kuhn, G., & Dienes, Z. (2006). Differences in the types of musical regularity learnt in incidental- and intentional-learning conditions. *Quarterly Journal of Experimental Psychology, 59*(10), 1725–1744. doi:10.1080/17470210500438361.

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.

Lau, H. C., & 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), 18763–18768. doi:10.1073/pnas.0607716103.

Overgaard, M., & Timmermans, B. (2009). How unconscious is subliminal perception? In S. Gallagher, & D. Schmicking (Eds.), *Handbook of phenomenology and cognitive science*. London: Springer.

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. doi:10.1016/j.concog.2006.04.002.

Persaud, N., & McLeod, P. (2008). Wagering demonstrates subconscious processing in a binary exclusion task. *Consciousness and Cognition, 17*(3), 565–575. doi:10.1016/j.concog.2007.05.003.

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

Persaud, N., McLeod, P., & Cowey, A. (2008). Commentary to note by Seth: Experiments show what post-decision wagering measures. *Consciousness and Cognition, 17*(3), 984–985. doi:10.1016/j.concog.2007.06.002.

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

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception and Psychophysics, 44*(6), 563–575.

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

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science: A Journal of the American Psychological Society/APS, 15*(11), 720–728. doi:10.1111/j.0956-7976.2004.00748.x.

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

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*, 367–447.

Wierzchoń, M., Taraday, M., Hawrot, A., & Asanowicz, D. (submitted for publication). Measuring consciousness of implicit learning using decision wagering, confidence ratings and feeling of warmth.

**Axel Cleeremans** is a Research Director with the F.R.S.–FNRS (Belgium). This work was supported by The MindBridge project, funded by the European Commission under the Sixth Framework Programme, Contract No. 043457 and by Concerted Research Action 06/11-342 titled "Culturally modified organisms: What it means to be human in the age of culture", financed by the Ministère de la Communauté Française – Direction Générale l'Enseignement non obligatoire et de la Recherche scientifique (Belgium).