Subjective awareness scale length influences the prevalence, not the presence, of graded conscious states

Henk Pretorius a,⇑, Colin Tredoux a, Susan Malcolm-Smith a,b

a Department of Psychology, University of Cape Town, Rondebosch 7701, South Africa
b ACSENT Lab, Department of Psychology, University of Cape Town, South Africa

A R T I C L E   I N F O

Article history:
Received 11 December 2015
Revised 28 May 2016
Accepted 12 August 2016

Keywords:
Consciousness
Awareness
Graded
Dichotomous
Psychophysics
Measurement

A B S T R A C T

There is an ongoing debate about the graded or dichotomous form of visual consciousness. Studies involved in the disagreement have typically employed subjective awareness ratings in psychophysical experiments. Variations in scale length have made comparisons across studies difficult and have even been suspected of influencing conclusions about the form of consciousness. We tested the proposal that a 21-point awareness scale produces dichotomous awareness state reports. The experiment described in this article randomly assigned participants to use one of four scale lengths used in previous studies in a backward masking task. Our findings suggest that all scales indicate the presence of graded awareness states, but that the resulting proportion of degraded state reports differed across the scales. Consequently, we argue that the decision of whether the form of consciousness observed in a given study is dichotomous or graded is dependent on an interpretation of the relative degree of degradation.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Is our visual experience of the world characterised by events that appear suddenly or gradually in our awareness? This apparently simple question has proved difficult to answer despite a growing body of pertinent research. Inspired by theoretical predictions from the global neuronal workspace theory (Dehaene, Kerszberg, & Changeux, 1998), evidence for all-or-none visual consciousness1 has been found in some studies (e.g. Sergent, Baillet, & Dehaene, 2005; Sergent & Dehaene, 2004), while other studies have supported a graded view of visual consciousness (e.g. Overgaard, Rote, Mouridsen, & Ramsøy, 2006; Ramsøy & Overgaard, 2004). Recently there have also been attempts to integrate these disparate accounts through appeals to separable levels of visual consciousness, a view that holds that some aspects of visual consciousness are graded while others are dichotomous (Windey, Geyers, & Cleeremans, 2013; Kouider, De Gardelle, Sackur, & Dupoux, 2010). Despite these attempts at resolution a review of the literature reveals a debate that remains unresolved. As such, the question of the form of visual consciousness has been firmly established alongside other debates in the recently flourishing area of consciousness studies.

⇑ Corresponding author.
E-mail addresses: henkpret@gmail.com (H. Pretorius), colin.tredoux@uct.ac.za (C. Tredoux), Susan.Malcolm-Smith@uct.ac.za (S. Malcolm-Smith).

1 In this article “consciousness” “awareness”, and “subjective visibility” or “experience” are employed interchangeably. The use of multiple terms is largely unavoidable given that the debate is often referred to as a debate about “consciousness” while the names of the measures in the debate frequently employ the other terms used here.
Data in the debate about the graded or dichotomous nature of conscious visual perception are delivered predominantly through subjective awareness ratings. In a psychophysical study typical of previous investigations, a person views a briefly displayed image and is asked to identify the object and deliver a rating of their consciousness of it using a subjective awareness scale. Variations in the format of these subjective awareness scales pose a challenge for comparability across studies and there has been little agreement on which of these measures is the optimal index of conscious experience. In fact, this methodological issue is itself a matter of ongoing debate and study (Dienes & Seth, 2010; Koch & Preuschoff, 2007; Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008; Timmermans, Sandberg, Cleeremans, & Overgaard, 2010). One aspect of awareness scale variation that is unresolved is the matter of the number of response categories. Previous studies have employed anywhere from 3 to 21 response alternatives, as summarised in Table 1, and the use of a particular scale length is not always defended in these studies. A challenge for researchers interested in the debate is that these studies have delivered such different conclusions that some investigators (Nieuwenhuis & De Kleijn, 2011; Overgaard et al., 2006) have suggested that the inference that visual consciousness is graded or not may depend on scale length.

1.1. Variations in awareness scale length

In an early study that stimulated interest in the graded or dichotomous form of visual consciousness, Sergent and Dehaene (2004) employed a “continuous” subjective visibility scale with 21 categories where only the endpoints were labelled: ranging from “not seen” to “maximal visibility”. This specific scale was developed to be “…sensitive to continuous changes in perception” (Sergent & Dehaene, 2004, p. 723). In the attentional blink experiment participants’ awareness ratings revealed an all-or-none pattern, suggesting a dichotomous form of consciousness. However, Overgaard et al. (2006) argue that the 19 unlabelled intermediate categories were likely difficult for participants to interpret and that this may have caused a bias towards the extreme categories anchored with descriptions. Furthermore, the use of 21 categories also appears excessively fine-grained for someone to use in expressing their subjective experience, making the findings difficult to inter-

The idea that four points is optimal is also questioned in an fMRI study (Christensen, Ramsøy, Lund, Madsen, & Rowe, 2006) and Dehaene (2004) attentional blink study did not feel compelled to use all 21 points of that scale when they were available. Recently, the use of this scale length has been challenged, with Windey, Vermeiren, Atas, and Cleeremans (2014) suggesting that the 4-point PAS scale might cause a bias to finding degraded conscious states because participants may feel compelled to use all four categories when these are available. However, this proposal does not explain why respondents in the Sergent and Dehaene (2004) attentional blink study did not feel compelled to use all 21 points of that scale when they were available. The idea that four points is optimal is also questioned in an FMRI study (Christensen, Ramsøy, Lund, Madsen, & Rowe, 2006) where participants found it difficult to distinguish between the two intermediate categories of the scale. For this reason a 3-point alternative was employed to better corroborate participants’ subjective experiences. Moreover, another recent study found better performance for a 3-step version of the PAS than the typical 4-point scale when comparing information from curve estimations (Sandberg, Timmermans, Cleeremans, & Overgaard, 2011). The studies that have employed three categories have supported the idea that visual consciousness is graded, with participants using the middle category and not only the endpoints of the scale.

A further variation of awareness scale length was introduced by Nieuwenhuis and De Kleijn (2011), who employed a 7-point scale in four experiments. The scale descriptions were based on the 21-point scale by Sergent and Dehaene (2004) and had only endpoints labelled “not seen” at the minimum end and “maximal visibility” at the maximum end of the scale. While 7-point Likert-type rating scales are commonly employed in psychology, the authors recognise that the use of this scale length in consciousness research is not without uncertainties when they remark that “…7 is still a rather arbitrary choice of scale points” (Nieuwenhuis & De Kleijn, 2011, p. 367). The findings from their experiments suggest that the 7-point scale can elicit graded responses, at least in some circumstances. In that study participants selected the intermediate categories more frequently for some experimental conditions than others using the same scale length.

1.2. Inferring graded consciousness

It is important to point out that in prior studies the conclusion that consciousness is graded or not is an inference based on awareness scale data, since consciousness is not directly measured. These studies have used at least two strategies to conclude that a conscious mental process is graded or dichotomous. One method (e.g. Koch & Preuschoff, 2007; Sandberg, Bibby, Timmermans, Cleeremans, & Overgaard, 2011) is to infer graded consciousness through mean ratings plotted on psychophysical curves that compare average awareness ratings across a manipulated variable (e.g. stimulus duration). However, participants could produce intermediate mean ratings by using an awareness scale in a completely binary manner, implying
all-or-none consciousness on any given trial but graded consciousness across trials (Estes, 1956; Nieuwenhuis & De Kleijn, 2011). Another way of inferring graded consciousness is to consider whether participants claim to experience degraded consciousness on a trial-by-trial basis. This is the strategy advocated by Overgaard et al. (2006) and involves investigating whether graded ratings are present in the response distributions. These reports are then corroborated by exploring the association between the awareness data and other variables, such as accuracy and stimulus duration. In this paper our analyses relied on both strategies. In cases where the analyses compared mean ratings (e.g. analysis of variance techniques), the existence of degradation in the trial-by-trial responses was confirmed alongside the averaged data.

1.3. Summary and aims

In summary, the question of whether the findings from previous various studies have been influenced by scale length is a pertinent and unresolved question. Indeed, Table 1 reveals an interesting pattern where longer (7- and 21-point) scales were most likely to lead to inferences of a dichotomous form of consciousness while the shorter 3- and 4-step versions have generally led researchers to conclude that a graded pattern of responses was observed. However, these experiments also differed in other respects, such as use of attentional blink or masking as degradation techniques, variations in stimulus type, and differing forms of analyses to infer graded consciousness, so it is not clear that scale length was instrumental in underpinning these conclusions. As far as we know no previous studies in this area have compared the findings from different scale lengths in the same experimental set-up with randomised assignment to different scale length groups. Consequently, while the issue of the optimal range of a rating scale has been studied in other fields, such as marketing (e.g. Cox, 1980; Krosnick & Fabrigar, 1997), the unique nature of research in this area makes a direct investigation into the issue a vital requirement for progress in resolving the present debate about the dichotomous or graded form of conscious visual perception.

The central aim of the experiment reported here was to explore whether the variations in extant awareness scale lengths influence conclusions about the graded or dichotomous nature of visual awareness. To this end we employed four subjective awareness scales of varying length in a backward masking paradigm. The primary hypothesis, derived from previous findings, was that a 21-point awareness scale would result in less graded awareness ratings, and as such lead to an inference of consciousness as dichotomous, compared to the shorter 3-, 4- and 7-point scales. Aside from assessing this hypothesis, a further exploratory question was considered: Which of the four scale lengths tested in this set constitutes a better measure for use in studies about the graded or dichotomous nature of awareness? Specifically, we considered which of the scales delivered acceptable granularity for detecting graded responses, while presenting a clear, unambiguous task to participants and researchers.

2. Materials and methods

2.1. Participants and design

Forty participants were randomly assigned to one of four conditions in a between-participants design. The tasks were identical across conditions with only a variation in scale length. Participants were exposed to only one version of the scale.
in order to avoid potential contamination effects from other conditions. In particular, our concern was that the usage of one scale type would bias the participant to use a similar rating system with subsequent scales. In the data screening phase eight participants (two per condition) were removed due to anomalous awareness rating data. These participants used one category of the scale almost exclusively, neglecting other scale points. This resulted in a total of eight participants per condition in the final analyses. The 3-point scale condition had five female participants (M = 30.12, SD = 5.51), the 4-point scale condition seven females (M = 34.12, SD = 7.32), the 7-point scale condition had four females (M = 30.00, SD = 7.72) and the 21-point scale condition had 3 female participants (M = 29.75, SD = 6.04). All participants reported normal or corrected-to-normal vision, normal colour vision and no history of a psychiatric condition or neurological damage. The experiment was conducted in Johannesburg, South Africa in a quiet room with attenuated lighting, free from noise and other distractions.

2.2. Masking procedure

Participants took part in a masked visual identification task adapted from a previous study reported by Sandberg et al. (2010). A masking task was preferred because previous masking studies have generally been able to gather evidence in favour of degraded states of awareness, as opposed to attentional blink tasks (Sergent & Dehaene, 2004). It was important for the present study that any effect of scale length on form of consciousness was maximally detectable. As illustrated in Fig. 1, each trial started with a fixation cross presented on a computer screen for one of four durations (500, 1000, 1500 or 2000 ms). The fixation cross was replaced by one of four simple geometric shapes (circle, square, diamond or triangle) presented for one of 12 durations (ranging from 13 to 156 ms, determined by the 75 Hz refresh rate of the CRT monitor). The target stimulus was presented on a grey background in the centre of the computer screen at a viewing distance of ~60 cm, subtending a visual angle of 3.3° × 2.1°. A mask, consisting of all four of the shapes overlaid on each other, followed the target and covered up the entire area where the shape was presented. The mask was displayed for 3000 ms. Participants were first asked to identify the shape that was presented by pressing keys on the keyboard ("c", "s", "t" or "d"), and were then prompted to provide an awareness rating using one of the four scales. The awareness rating was made by using the

![Fig. 1. Experimental procedure for the four conditions.](image-url)
mouse cursor to select from the available response options, displayed as contiguous blocks that ran horizontally across the bottom of the screen.

Each session began with a practice block of 48 trials with only the longest stimulus durations. The formal trials consisted of 384 trials divided into two blocks of 192 trials. Participants could take a brief break between the two blocks.

2.3. Subjective measures of consciousness

After each shape identification attempt participants were asked to rate their level of awareness of the target stimulus using one of the four subjective measures of consciousness. Scale labels were adapted from the PAS (Ramsøy & Overgaard, 2004). The format of this scale asks for reports of the introspective experience of participants and not ratings of confidence. As Overgaard and Sandberg (2012) argue the difference between introspective and other metacognitive judgments is not trivial. As such the representativeness of these findings do not necessarily extend to reports other than those of an introspective nature. Similar labels were used for the minimum and maximum levels for all scales. The set of scale lengths was chosen to be consistent with the variations employed in previous studies. Only the longest, 21-step scale version was not fully labelled with verbal descriptors due to the difficulty in labelling this extended scale. The set employed here was formatted as follows:

- **3-point scale**: At least two studies have employed a 3-point awareness scale (Christensen et al., 2006; Sandberg et al., 2011). In this study the 3-point scale was formatted as follows: 1 – “No experience of stimulus”, 2 – “Brief glimpse or vague experience” and 3 – “Clear experience”.
- **4-point scale**: Studies that used the PAS (Overgaard et al., 2006; Ramsøy & Overgaard, 2004; Sandberg et al., 2010) typically employed four response options. In this condition the standard PAS descriptions were used: 1 – “No experience of stimulus”, 2 – “Brief glimpse”, 3 – “Almost clear experience” and 4 – “Clear experience”.
- **7-point scale**: The study by Nieuwenhuis and De Kleijn (2011) employed a 7-point confidence rating and post-decision wagering (PDW) scale. Only the PDW scale was fully labelled, with the amount of money that is wagered indicated at each level. To make this scale maximally comparable to the 3- and 4-point versions it seemed appropriate to fully label this scale: 1 – “No experience”, 2 – “Almost no experience”, 3 – “Slight or brief experience”, 4 – “Moderately clear experience”, 5 – “Somewhat clear experience”, 6 – “Almost clear experience” and 7 – “Clear experience”.
- **21-point scale**: Studies that employed this scale length (Del Cul et al., 2006, 2007; Sergent & Dehaene, 2004; Sergent et al., 2005) did not label every step of the scale due to its elaborate nature. This scale was the only version employed here that had just the endpoints labelled: “No experience” and “Clear experience”. Although this introduces a variation for this scale only, this is consistent with the way in which the scale has been employed previously, and participants were thoroughly briefed that all points on the scale should be used to match their experience as closely as possible.

A printed copy of the scale label descriptions was left with participants for reference during the sessions.

3. Results

3.1. Accuracy data

Before proceeding with the analyses of the awareness data it was important to rule out the possibility that accuracy differed across the experimental groups. The percentage of shapes accurately identified across the conditions was highly similar for all conditions: 86% for the 3-point scale, 86% for 4-point scale, 89% for 7-point scale, and 88% for the 21-point scale. Any differences observed in awareness ratings across conditions are thus unlikely to have been caused by differences in accuracy.

3.2. Awareness data

Our analyses of the awareness data started with an initial investigation of the combined data which we compared across conditions.

3.2.1. Combined response distributions of awareness ratings

The response distributions of the awareness ratings delivered by the different scales are displayed in Fig. 2(a). All of the scales show an expected pattern of increasing awareness as stimulus duration increases. In addition, all of the scale conditions appear to contain a significant number of degraded awareness reports, especially at certain intermediate durations in the 39–91 ms range. Participants seem to have used intermediate steps on all scales, not just the endpoints. Although it is not possible from the aggregated data only to say that all participants used these middle points.

3.2.2. Association with stimulus duration and accuracy

The association between stimulus duration (the independent variable) and mean awareness rating (the dependent variable) was tested through one-way analysis of variance. The mean ratings by duration are presented in Fig. 3(a). Aside from
the 3-point scale (which shows a range of mean ratings skewed to the higher end of the awareness scale range) all other scales show a very similar range and shape of the psychophysical curve. In all four conditions the effect of stimulus duration on awareness rating was statistically significant. Post hoc comparisons were employed to assess which intermediate ratings resulted in significantly different ratings compared to both the 13 ms (least visible) and 156 ms (most visible) conditions. These states would suggest intermediate awareness since they differ from the highest and lowest awareness states. The effect was significant for the 3-point scale, $F(11,2984) = 196.83, p < 0.001$, eta-squared = 0.420, 4-point scale, $F(11,3026) = 157.28, p < 0.001$, eta-squared = 0.364, 7-point scale, $F(11,3031) = 158.64, p < 0.001$, eta-squared = 0.365, and 21-point scale, $F(11,3052) = 273.57, p < 0.001$, eta-squared = 0.496. The post hoc Games Howell comparisons revealed that there are intermediate ranges of stimulus durations where all four of the scales show averaged ratings that are different from both the 13 and 156 ms conditions. This range starts from 39 ms for all scale lengths, and extends to 91 ms for the 3-point scale ($p < 0.05$), to 117 ms for the 4-point ($p < 0.05$) and 7-point scales ($p < 0.05$) and 104 ms for the 21-point scale ($p < 0.05$). The response distributions shown in Fig. 2(a) rule out the possibility that the averaged data resulted from combining purely dichotomous (aware and unaware) ratings at the intermediate ranges. The data shows a high prevalence of degraded awareness reports at these durations.

Next we explored the potential association between the awareness ratings and accuracy (Fig. 3(b)). For all scales there is a generally monotonically increasing relationship between the awareness rating and level of accuracy. The 21-point scale shows a larger range from lowest to highest accuracy levels with the lowest range evident for the 3-point scale. Importantly for our main question is the observation that there are rating ranges, around 20–60% for all scales where the resulting accuracy falls between the lowest and highest levels. This suggests a corroboratiom of the idea that these ratings represent meaningfully different states compared to lower and higher ratings.

Fig. 2. Awareness ratings as a function of stimulus duration in the four conditions for: the untransformed data (a), the transformed data (b). Shading represents awareness ratings from 1 (“no experience”) to max (3, 4, 7 or 21 depending on scale, and labelled as “clear experience” for all scales). See Fig. 1 for full format and labels.
Fig. 3. Association between untransformed awareness rating and timing duration and (a) between awareness rating and identification accuracy (b). Awareness is expressed as a fraction of the range for each scale, where 100 = scale max, 0 = scale minimum. Error bars represent one standard error of the mean.
In summary, the analyses of the combined data support the idea that all four scales deliver graded awareness data that can be corroborated through associations with other measures. As such these results are consistent with the proposal that detecting the presence of degraded states is possible with all of the scale lengths that we tested. However, as we argue in the next section, it is more accurate to consult the individual-level data before firmly making this inference.

3.3. Individual-level response distributions of awareness ratings

As we have noted earlier, aggregated response distributions and psychophysical curves may be misleading since the combination of data across participants may lead to a distribution that is different from any individual’s rating distributions. We investigated the participant-level responses (Fig. 4) in order to address this potential confound. In particular, we are interested in whether the individual response distributions suggest a graded or dichotomous form of awareness. This analysis also allowed us to begin to answer our secondary objective of which of the scale lengths tested here performs better as a measure of awareness.

For the 3-point scale six out of the eight participants used the middle and highest category almost exclusively (<6 reports out of 384 trials of “no experience”). The response distributions for this scale are predominantly unimodal with the highest number of ratings at the “clear experience” category. This explains our previous finding that the averaged data are skewed towards the upper end of the range of responses. The 4-point scale shows evidence of more distributed responses, with all participants using all scale categories. The rating distributions show more individual variation with half of participants favouring the two middle categories, three showing a unimodal pattern slanting towards the highest category and one participant (numbered 8 in Fig. 4) showing a bimodal pattern of ratings at the ends of the scale that appears more dichotomous.

The 7-point scale generally shows ratings across all categories of the scale, with only two participants using less than the maximum number of categories available (using 5 and 6 respectively). The rating pattern is more diverse than the two shorter scales, ranging from unimodal at the second highest step (e.g. participant 2) to bimodal (e.g. participant 5) along with several more distributed patterns. In the case of the 21-point scale only two participants used less than 21 categories (using 20 and 15). The response distribution patterns were most diverse for this scale. In fact, this distributed responding in the 21-point case is evident even when considering trials that have the same objective conditions. We analysed individual

![Fig. 4. Frequency distribution of awareness ratings on a participant level for four awareness scale length conditions. Numbers represent awareness ratings from 1 (“no experience”) to max (3, 4, 7 or 21 depending on scale, and labelled as “clear experience” for all scales). See Fig. 1 for full format and labels.](image-url)
responses across scales at 156 ms, the longest display duration, and hence the most visible condition. Each participant was exposed to 32 trials at this duration. The data reveal a trend towards more widely distributed responses for longer scale lengths. This contrast is clear when comparing the shortest and longest scale versions. For the 3-point scale all participants used the third step (clear experience) more often than any others, an expected result. In addition, the second most common rating (the middle category “2”) contained far fewer responses and was used by fewer respondents. However, for the 21-point scale there were four modal responses, depending on the participant, all of which were not contiguous. Furthermore, the range of responses at this stimulus duration for the longest scale was 5–21 for one person, a large dispersion of ratings. The variability under similar viewing conditions within and across participants was greatest for this scale. These effects on longer scales were evident for the intermediate 4- and 7-point scale lengths, but to a lesser extent.

Despite these differences in response styles across and within scale lengths there is little evidence to support unanimously dichotomous responding for the 4-, 7- and 21-point scales. The exception is the 3-point scale where unexpectedly there was a lot of binary responding. However, instead of using the lowest and highest categories, as would be predicted by a dichotomous view of consciousness, participants favoured the middle and highest categories. Overall, the individual-level data thus supports the conclusion from the combined analysis which is that the 21-point scale did not lead to purely dichotomous responding.

### 3.4. Transformation of awareness data to facilitate a direct comparison

In order to facilitate a more direct comparison of the scale data it was necessary to isolate three similar kinds of trials. It was considered best to transform the data on an individual level since our previous analysis showed participants used even the same scale length quite differently. The following review shows the challenges in comparing data from different scale lengths and presents our solution to this transformation challenge.

#### 3.4.1. Scale transformation strategies

Two simple mathematical solutions to the problem of comparing different scale lengths involve applying a simple proportional transformation and normalising scores (Colman, Norris, & Preston, 1997). For the simple proportional transformation, the longer scales (4-, 7- and 21-point) would be divided by a factor to achieve scores comparable to the 3-point scale. A similar strategy is to convert scores from the different scales into standardised scores (e.g. Z-scores) with fixed means and standard deviations. The problem with both of these solutions is that they assume that a mathematical relationship underlies the scale use, while the differences are more likely to be psychological (Colman et al., 1997). If the rating strategy is governed by a mathematical relationship, then someone should rate their experience in a way that will make their score on a longer scale a proportional match if they had used a shorter scale. For example, if someone uses the midpoint on the 21-point scale (a rating of 11), their rating on the 7-point scale should be at 4, which is the midpoint of that scale. However, the real equivalence between scales is more likely to conform to a psychological relationship that can only be empirically determined. An indication of this mismatch is evident in our data where the data from the 3-point scale delivered far less responses in the lower category than would be predicted by looking at data from a similarly low range for the other scales. A further problem with these strategies is that even though the range of all scales will be comparable, after being transformed to range from 1 to 3, a further transformation would be required to convert all scales to three levels in order for them to be closely comparable.

Another approach involves binning data from contiguous scale positions into the same category. One straightforward strategy would be to collapse all intermediate categories from the longer scales into a “degraded states” bin to be comparable to the 3-point scale. Initially this may seem justified based on the fact that for all scales the endpoints had the same descriptions (“no experience” and “clear experience”). However, comparing the data evident in the response distributions shown earlier suggest this approach may have shortcomings. The data (see for example Figs. 2(a) and 4) shows a marked increase in the proportion of degraded states as the scale length increases. This raises the question of whether, in the case of the longer (7- and 21-point) scales, this strategy might have confounded “clear experience” and “no experience” with degraded steps. It is possible for example that a rating of “2” on the 21-point scale is perceptually similar to a rating of “1” on the same scale, and that the binning of all intermediate steps has missed this response error. A further simple binning strategy was employed in a previous study investigation (Nieuwenhuis & De Kleijn, 2011). In that study data from a 21-point scale was converted by simply binning all three contiguous categories in series to get to a 7-point version. However, the authors admit that it is questionable whether the data would have looked the same had the original data been collected with a 7-step scale.

Empirical approaches to transformation have also been employed in previous studies. In these strategies the data contributes to how the transformation is done. Sergent et al. (2005) used individual-level median responses as cut-off points to define “seen” and “not seen” states. However, with this binary transformation of the graded data it is impossible to detect degraded states. In another study (De Gardelle, Kouider, & Sackur, 2010) subjective visibility ratings and objective stimulus durations were used to define three different states. Low visibility trials were defined as those involving the shortest durations and lowest visibility ratings. Intermediate visibility trials included intermediate durations and ratings, while high visibility trials included the highest durations and ratings. Another empirical approach was employed by Sergent et al. (2013) who explained the subjective visibility data through a model of two distributions on an individual level. Empirical approaches have the benefit of relying not on a priori criteria but rather on the structure of the data to guide the transformation.
3.4.2. Transformation strategy based on awareness and accuracy data

We employed an empirical transformation strategy to define three similar categories of trials across the scale conditions. We defined these categories on a subject-by-subject basis by using both the awareness ratings and the accuracy data. We grouped together adjacent awareness ratings for each participant so that the corresponding identification accuracy was in a similar range for each new category across participants. For each participant contiguous awareness ratings were first collapsed so that there was a minimum of 10 trials in each of the rating categories. This was done to get a more accurate estimate of the accuracy level associated with the awareness categories. For the longest scale the 21 categories were first collapsed into 7 categories by grouping ratings of three together in series. After that the same approach was applied as with the other scales. The new rating categories were then binned based on their relationship with accuracy. These accuracy ranges were <75% for the low awareness trials, between 75% and 97.5% for the intermediate awareness trials and >97.5% for the high awareness trials. An example of how this was applied to data from one participant in the 7-point scale condition is shown in Fig. 5(a). While no data are excluded in our transformation for some participants certain states are absent in the newly defined categories since no contiguous grouping of ratings corresponded to the level of accuracy defined in our new ranges. For all participants the new awareness categories thus have a comparable effect on accuracy, providing a reasonable similarity between the categories for the following analyses.

3.4.3. Overall transformed awareness categories by scale length

The percentage of total trials per scale that delivered intermediate awareness states after this transformation is displayed in Fig. 5(b). All scales produced intermediate state trials. However, the frequency of these degraded state trials differs markedly between scales. The chi-square test revealed that the effect of scale length on the proportion of intermediate trials is significant, \( \chi^2(3, N = 32 \text{ participants and 12,288 trials}) = 728.78, p < 0.001 \), Cramer’s \( V = 0.244 \). The 4-point scale resulted in more intermediate states than any of the other scales (48%) while the 21-point scale resulted in the lowest incidence of intermediate trials (16%). The 7-point (34%) and 3-point scale (28%) resulted in a similar number of these trials.

Even though all the scales resulted intermediate awareness trials, the magnitude of degradation suggested by the scales is significantly different.

3.4.4. Transformed awareness categories by stimulus duration

We next considered how the new awareness categories develop across the stimulus display duration. First, we looked at the mean awareness data, see Fig. 5(c). For all scales the association between stimulus duration and mean awareness category was significant as assessed through multiple one-way analysis of variance tests. The effect was significant for the 3-point scale, \( F(11,2984) = 145.63, p < 0.001 \), eta-squared = 0.349, 4-point scale, \( F(11,3026) = 179.73, p < 0.001 \), eta-squared = 0.395, 7-point scale, \( F(11,3031) = 198.59, p < 0.001 \), eta-squared = 0.419, and 21-point scale, \( F(11,3052) = 261.89, p < 0.001 \), eta-squared = 0.486. The post hoc Games Howell tests revealed that there are similar timing ranges that produce intermediate averaged awareness levels that are significantly different from the mean awareness at both the lowest (13 ms) and highest (156 ms) durations. For the 3-point scale the range was 39–91 ms (all \( p < 0.01 \)), for the 4-point scale extended from 39–91 ms (all \( p < 0.05 \)), for the 7-point scale was 39–78 ms (all \( p < 0.001 \)), and for the 21-point scale was 26–91 ms (all \( p < 0.05 \)).

As described earlier there is the possibility that intermediate averaged awareness can result from completely dichotomous awareness categories. For this reason, we considered the response distributions of the intermediate states across the stimulus durations in the four conditions (Fig. 2(b)). It is evident that the range of 39–78 ms resulted in more intermediate awareness trials than shorter or longer durations across all scales. This becomes even clearer when we isolated the intermediate states (Fig. 5(d)). This is consistent with the range observed in the averaged data and points to a window where intermediate awareness states are most likely to occur.

Regarding the central question and hypothesis that this data aimed to answer it is clear that the hypothesis regarding an absolute bias of the 21-point scale to find only dichotomous states is rejected. However, it should be added that the magnitude of degraded trials this scale produced, under similar awareness definitions, was significantly less than for other scales.

3.5. Best scale length in this set

A secondary aim was to assess which of these commonly used scale lengths performed better in our experiment. Firstly, the investigation of the participant-level data leads us to prefer some scales over others when investigating the form of consciousness. The binary fashion in which participants used the 3-point scale in our study makes it problematic when the goal is distinguishing between at least three classes of experience. Of course it is possible that respondents really only experience two types of awareness states, but the fact that all three other scales lead to more granular responses makes us question this possibility. With this scale there is the risk that a participant neglects one scale point, which would make the scale binary. On the other end of the range the 21-point scale leads to such varied responses that it is hardest to interpret and meaningfully aggregate the data. This was clear when our transformation strategy was considered. A further challenge with the 21-point scale was that, while we did not find complete dichotomous responding it did lead to less graded states in our transformed

---

2 This is similar to the effect noted by De Gardelle et al. (2010) who had to exclude 19% of trials in that study due to their combined definition of the three awareness states.
data, even when using the same definition for the new trial classes across all scales. This leaves the 4- and 7-point scales. Both these scale lengths strike a balance between finer granularity (compared to 3 points) and clearer interpretability (compared to 21 points). We consider these scales to have the most favourable characteristics, at least with the current experimental set up. One speculative possibility worth investigating is the correspondence between the contents of consciousness and Miller’s (1956, p. 343) famous conclusion that the “magic number seven, plus or minus two” is a limit for human information processing and judgement.3

Fig. 5. Transformed awareness data. (a) Example of an actual participant’s awareness rating and accuracy data in the 7-point scale condition and how the transformation was done. Thresholds represents 75% and 97.5% accuracy. Error bars represent one standard error of the proportion. (b) The total distribution of transformed awareness data across scales. (c) Mean transformed awareness by stimulus duration. Error bars represent one standard error of the mean. (d) Response distribution of intermediate awareness categories by stimulus duration.

3 As pointed out by one reviewer of an earlier version of this paper Miller’s (1956) proposal was directed at short-term memory and not the contents of consciousness, so this remains a speculation until tested.
Of course the optimal number of response categories was not the main purpose of our experiment and a better answer to this question can only be achieved by assessing a larger range of scale points. The set we have tested, while consistent with those employed in the literature, jumps from 7 to 21 points, leaving a large range of potential scale lengths untested.

4. Discussion and conclusion

This study addressed the question of whether rating scale length influences conclusions about whether conscious experience is all-or-none or graded. With the exception of the 3-point scale our findings showed that the 4-, 7- and 21-point scales produced graded awareness reports. This was evident in the analyses of the aggregated and individual-level raw awareness data as well as the transformed data. This finding was also consistent in both tests using the frequency of intermediate response distributions and the averaged data plotted on psychophysical curves. Therefore, our data does not support the notion that a 21-point scale should necessarily lead to dichotomous awareness ratings, as has been proposed (Nieuwenhuis & De Kleijn, 2011; Overgaard et al., 2006). In fact, our finding is consistent with data reported by Sergent and Dehaene (2004), where this scale was introduced. Aside from the attentional blink experiment that study also contained a masking experiment that indicated graded use of this same scale length. Furthermore, another study by Nieuwenhuis and De Kleijn (2011) showed that a 7-point scale used in attentional blink experiments could be used either dichotomously or in a graded manner depending on other factors in the experiment such as the type of stimuli and the scale wording and instructions. As such scale length alone is unlikely to have been the reason why a dichotomous form of consciousness was inferred in the attentional blink study reported in Sergent and Dehaene (2004). Indeed, the use of masking vs. attentional blink manipulations are also key suspects in variations of findings of graded or dichotomous awareness, of which the Sergent and Dehaene (2004) is an example.

It is interesting to note that we did find that the magnitude of the degradation observed across trials with the different scales varied significantly. This was especially evident the transformed data was considered. The analysis of this data showed that the 21-point scale resulted in less frequent trials of intermediate awareness, according to the definition we developed here. So while this scale did not preclude the detection of graded awareness it did limit the prevalence of graded responding. This may explain partly the prior observations about the scale’s bias. It may have resulted from a relative, not absolute, bias to dichotomous responding. As for the other scales and their effect on graded responding, the transformed awareness data suggests the 4-point version delivered the most reports of intermediate awareness states. This is perhaps not surprising given the fact that the scale was developed to be sensitive to degraded awareness states (Ramsey & Overgaard, 2004). However, the possibility that the scale contains bias towards graded responding, as has been recently proposed (Windey et al., 2014), is also consistent with this result (but see our earlier point that this possibility should then extend to the longer scales too). The 7-point scale produced the second highest frequency of intermediate awareness trials in the transformed data analysis, followed by the 3-point scale. A challenge for the 3-point scale was that these intermediate awareness states were delivered by only some participants since many of those using this scale used only two scale points.

What is clear from these findings is that consciousness can be inferred to be graded with various scale lengths, albeit to different degrees. As such, the decision of whether the form of consciousness observed in a given study is dichotomous or graded comes down to an interpretation of the relative degree of degradation. When confronted with 20% degraded states across all trials in a study, this could interpreted as being a mostly dichotomous cognitive process. Similarly, strong non-linearity observed in the averaged data across a timing variable could be interpreted as being suggestive of a dichotomous form of consciousness. However, if there are robust degraded states evident in the response distributions at all, and if there is an intermediate mean value despite a sharply rising psychophysical curve, it seems more accurate to comment on the relative degradation of awareness. So from this perspective the 21-point scale data from our study produced relatively less intermediate awareness trials compared to a 4-point scale, but cannot be used to infer that consciousness is dichotomous. The ability to detect the window of intermediate awareness appears to vary with scale length, but does not disappear completely.

We propose several salient opportunities for further research arising from this study. Firstly, the present sample sizes were relatively small, especially after the removal of problematic responses. Future work with larger samples may deliver even firmer direction on these issues. Furthermore, our study tested the impact on the conclusions about the form of consciousness using only one aspect of variation in awareness scale format, namely the number of response options. There are other aspects of these scales, including the nature of instructions and option wording, the format of reporting on a computer (as opposed to verbally or via another medium) that are valuable areas for future investigations. As we mentioned earlier recent experiments (reviewed in Overgaard & Sandberg, 2012) show meaningful differences in introspective reports about consciousness, like those we employed here, and other metacognitive reports such as confidence ratings. We consider it worthwhile exploring differences in scale length for these other types of reports. Further, while our second area of focus was on testing which format of awareness scale performed best in the present set we leave unanswered questions around how other scale lengths and formats would have performed had they been tested. Finally, considering the large individual variation in scale use evident in this study future work could assess how to better make use of empirical approaches to participant-level analyses like the approach we employed here.
Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.concog.2016.08.007.

References


