RESEARCH ARTICLE

REVISED What can we experience and report on a rapidly presented image? Intersubjective measures of specificity of freely reported contents of consciousness [version 2; peer review: 2 approved]

Zhang Chuyin[1*], Zhao Hui Koh[1*], Regan Gallagher[1], Shinji Nishimoto[2,3], Naotsugu Tsuchiya[1,2,4]

[1]Turner Institute for Brain and Mental Health & School of Psychological Sciences, Faculty of Medicine, Nursing, and Health Sciences, Monash University, Melbourne, Victoria, 3168, Australia
[2]Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology, Suita, Osaka, Japan
[3]Osaka University, Suita, Osaka, Japan
[4]Advanced Telecommunications Research Computational Neuroscience Laboratories, Hikaridai, Kyoto, Japan

[*] Equal contributors

v2 **First published:** 20 Jan 2022, **11**:69
https://doi.org/10.12688/f1000research.75364.1
Latest published: 24 Aug 2022, **11**:69
https://doi.org/10.12688/f1000research.75364.2

Abstract

Background: A majority of previous studies appear to support a view that human observers can only perceive coarse information from a natural scene image when it is presented rapidly (<100ms, masked). In these studies, participants were often forced to choose an answer from options that experimenters preselected. These options can underestimate what participants experience and can report on it. The current study aims to introduce a novel methodology to investigate how detailed information participants can report after briefly seeing a natural scene image.
Methods: We used a novel free-report paradigm to examine what people can freely report following a rapidly presented natural scene image (67/133/267ms, masked). N = 600 online participants typed up to five words to report what they saw in the image together with confidence of the respective responses. We developed a novel index, Intersubjective Agreement (IA). IA quantifies how specifically the response words were used to describe the target image, with a high value meaning the word is not often reported for other images. Importantly, IA eliminates the need for experimenters to preselect response options.
Results: The words with high IA values are often something detailed (e.g., a small object) in a particular image. With IA, unlike commonly believed, we demonstrated that participants reported highly specific

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 24 Aug 2022		✓ view
version 1 20 Jan 2022	✓ view	? view

1. **Marius Usher**, Tel Aviv University, Tel Aviv, Israel
2. **Qiufang Fu**, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

Any reports and responses or comments on the article can be found at the end of the article.

and detailed aspects of the briefly (even at 67ms, masked) shown image. Further, IA is positively correlated with confidence, indicating metacognitive conscious access to the reported aspects of the image.
Conclusion: These new findings challenge the dominant view that the content of rapid scene experience is limited to global and coarse gist. Our novel paradigm opens a door to investigate various contents of consciousness with a free-report paradigm.

Keywords
Consciousness, free report, rapid scene, metacognition, confidence, gist, intersubjective agreement

Corresponding authors: Zhang Chuyin (chuyin.zhang@monash.edu), Naotsugu Tsuchiya (naotsugu.tsuchiya@monash.edu)

Author roles: Chuyin Z: Conceptualization, Data Curation, Formal Analysis, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Koh ZH**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Gallagher R**: Conceptualization, Methodology; **Nishimoto S**: Resources, Writing – Review & Editing; **Tsuchiya N**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Australian Research Council (N.T., DP180104128, DP180100396); National Health Medical Research Council (N.T., APP1183280); Foundational Question Institute (N.T., FQXi-RFP-CPW-2017); and Japan Promotion Science, Grant-in-Aid for Transformative Research Areas (N.T. 20H05709, 20H05710).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Chuyin Z, Koh ZH, Gallagher R *et al.* **What can we experience and report on a rapidly presented image? Intersubjective measures of specificity of freely reported contents of consciousness [version 2; peer review: 2 approved]** F1000Research 2022, **11**:69 https://doi.org/10.12688/f1000research.75364.2

First published: 20 Jan 2022, **11**:69 https://doi.org/10.12688/f1000research.75364.1

REVISED **Amendments from Version 1**

Based on the reviewers' comments, we made the following changes in Version 2:

1. We modified Figures 2 and 3, and updated the text to explain them.

2. We added Figure 6(d) into Figure 6, and a paragraph in the main text to describe and explain the data shown in Figure 6(d).

3. We added a few footnotes to explain the reason we use "Word IA" instead of "Word-Image IA", our motivation behind Word IA, the reason we asked participants to report 5 words, and to say that the results won't change after removing artificial images.

4. We added a paragraph under "Presentation duration" to discuss further studies using shorter SOAs.

5. We edited the last paragraph under "Future directions" to discuss the implications to the overflow in consciousness research.

6. We clarified the number of participants we analysed throughout the paper.

7. The analysis code on Github has been updated to include additional analysis we ran, in order to respond to reviewers' comments.

Any further responses from the reviewers can be found at the end of the article

Introduction

Intuitively, we have an impression that our visual experience is immensely rich. When we look at a complex natural scene, we can rapidly extract meaningful information and categorize it accurately in as short as 150 ms (Li *et al.*, 2002; Biederman, 1981; Fabre-Thorpe *et al.*, 2001). In certain situations, the performance accuracy is correlated with confidence rating (Fu *et al.*, 2016; Thunell & Thorpe, 2019), indicating some level of metacognitive monitoring is possible. Rapidly extracted information in this context is often called "scene-gist understanding" (Oliva & Torralba, 2006) or "gist" in short. Previous studies demonstrate excellent human capability to rapidly categorise a natural scene based on the scene gist in a global and coarse manner. However, it is unclear whether we have conscious access to more detailed information upon seeing a complex natural image briefly (Bayne & McClelland, 2018; McClelland & Bayne, 2016). The current study will focus on the information beyond "gist", so that we will use a more general term: rapid scene experience.

In fact, the current dominant view on contents of rapid scene experience is that it is about high-level descriptions of the scene, which is coarse and lacks the fine details (Campana & Tallon-Baudry, 2013; Fei-Fei *et al.*, 2007; Greene *et al.*, 2015; Kimchi, 1992; Oliva & Torralba, 2006; Greene & Fei-Fei, 2014).

This dominant view inherits the idea from the Gestalt theory of perception (Koffka, 1922). One of the Gestalt principles states our tendency to perceive a scene as a whole rather than individual parts. This idea was further developed by Navon (1977) in his global precedent hypothesis. This hypothesis posits that visual perception first processes global features of an image, then proceeds to analyze local features. Supplemented by further experimental evidence, Campana and Tallon-Baudry (2013) recently extended this idea into conscious perception, where they claim conscious perception starts at the global and coarse level before the local and detailed level.

Taxonomy-based levels of categorization ("semantic categorization"; Rosch, 1998) was frequently used to study the content of rapid scene experience. The categories varied in hierarchies based on their degree of general-ness or specific-ness. In this categorization, the content of rapid scene experience can be reported either at the superordinate level or the basic level (Fei-Fei *et al.*, 2007; Larson *et al.*, 2014; Walther *et al.*, 2009), but less likely at the subordinate level (Malcolm *et al.*, 2012), with the subordinate level being the most specific category. Further evidence shows that the superordinate categorization is made prior to the basic-level categorization (Loschky & Larson, 2010; Sun *et al.*, 2016). These findings provide evidence for the dominant view.

Overall, studies in this field often used forced-choice discrimination tasks, such as identifying the same image through a series of rapidly presented images (e.g., Thunell & Thorpe, 2019), ascertaining the superordinate category (natural-ness, urban-ness) of a scene (e.g., Loschky & Larson, 2010), and determining if the briefly viewed images contain animals (e.g., Thorpe *et al.*, 1996) or certain objects (e.g., Hollingworth, 1998). However, there are two methodological limitations in the forced-choice paradigm.

First, the forced-choice paradigm restricts the response options. The studies that employ the forced-choice paradigm often implicitly assume that participants can only process the images at a certain level (usually coarse). Due to the limited response options, participants were unable to report detailed information about their experience even if they perceived it (Haun *et al.*, 2017). Researchers may employ the forced-choice paradigm just for the sake of convenience. However, other paradigms, such as free-report paradigms, are potentially more difficult to verify from a third-person view (Dennett, 2007).

Second, the forced-choice paradigms create expectations of the content of images that will be presented for participants. For example, asking participants to choose between "animal" and "non-animal" will tell them they might see an animal. This could help participants perceive more detailed information (McLean *et al.*, 2021; Sun *et al.*, 2017), while it can suppress seeing other aspects of rapidly presented scenes.

To overcome these problems, here we introduce a novel free-report paradigm. First, it addresses the restrictions on what participants can report. They freely reported what they saw in a briefly presented image and rated their confidence in the report. Second, to reduce any expectation that participants might have about the upcoming image category, we used a wide array of 412 natural and 8 artificial images and did not tell participants any information about the types of the images that they will see or types of the responses that we expect from them.

Of course, these modifications generate new challenges: how can we scientifically investigate such unconstrained data? What can we do to eliminate bias from us, the experimenters, on what participants should report? Rather than manually going through each image to define the correct answers, we utilize the majority vote of the participants. In other words, based on a large baseline of what participants report, we quantify the degree of specificity of reported words, which we call "intersubjective agreement (IA)". For this purpose, we recruited a large pool of 600 participants. With our novel index applied to the free-report paradigm, we reexamine the dominant view that the content of rapid scene experience is limited to coarse information lacking in detail. The current study will answer if people can consciously see and report the details of an unexpected image at a brief glance.

Results

To summarise, our novel index quantifies how "specific" a response is, with respect to a certain image. IA ranges from 0 to 1. A high IA means the word is specifically reported under the target image and not reported under other images, signifying detailed information. A low IA (~0.5) means the word is reported nonspecifically for all the images, signifying coarse information.

Figure 1(a) explains the time course of one trial in our experimental paradigm (see also Methods). Upon brief exposure to an image, participants typed five words in the response box with confidence indicated in each word (See (b) in Figure 1). We tested 670 participants online, but analysed the first 30 participants within each of 20 groups. Thus, our results reflect those from 600 participants.

Intersubjective agreement (IA)

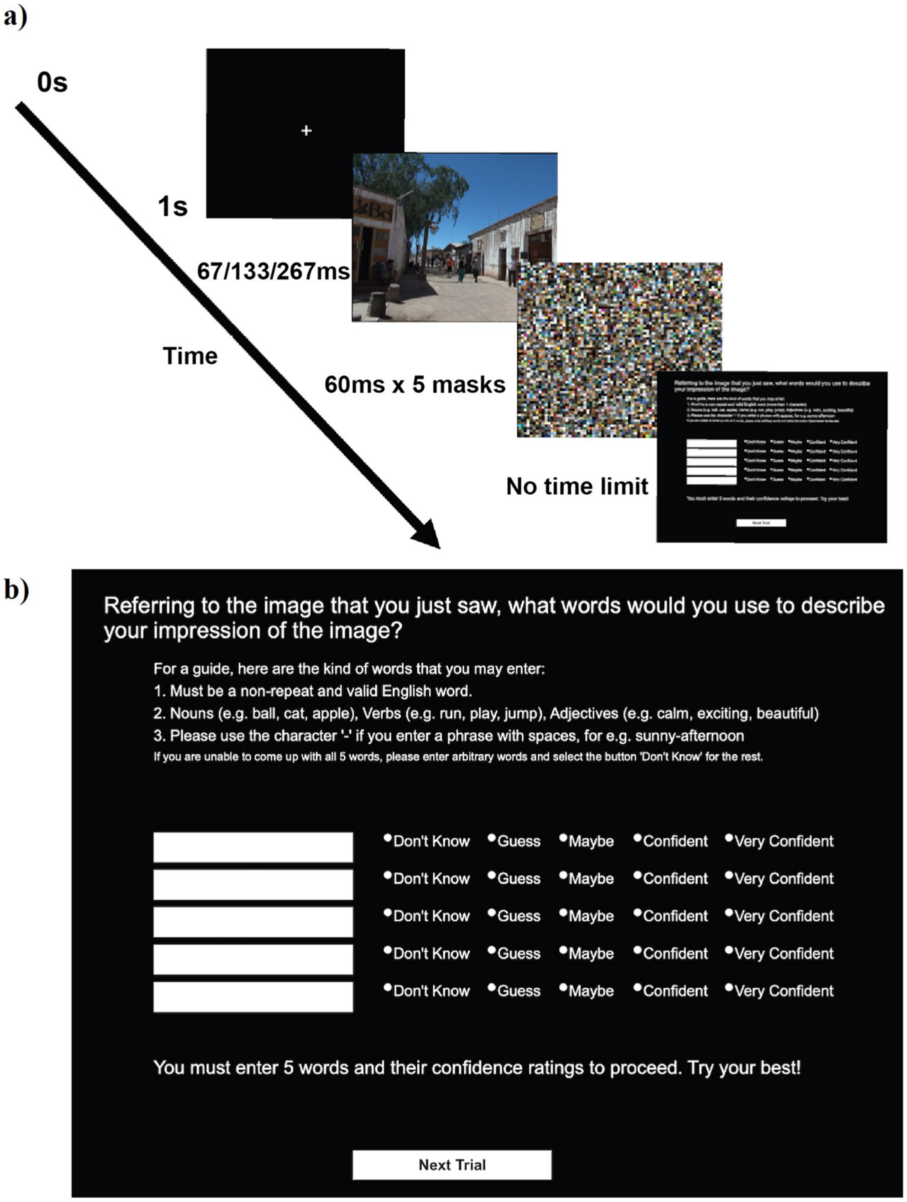
To quantify the degree of "specificity" of a reported word on a given image, we compare the report frequency of the word between the target image and the rest of the images. We call the index "intersubjective agreement (IA)" because it describes how well the reported word agrees with what other participants reported on that image but not on the other images. We refer "Word IA"[1] to the IA associated with a particular word for a given image.

With our IA index, we do not need to assume any *a-priori* correct answers. A word response that has high Word IA with respect to an image indicates that the word is highly specific to the image, compared to the rest of the images.

Calculation of Word IA

Figure 2 explains the steps involved in calculating Word IA under a particular stimulus onset asynchrony (SOA). Word IA is defined on a particular image and a particular word. Let us explain the case using a target image in Figure 2(a) and a word response "eiffel-tower" from participant 1 in Figure 2(b) (red solid box; For details, see Methods: Calculation of Word IA for each SOA). Figure 2(b) represents words reported by ten participants after viewing the target image for 67 ms.

[1] To be precise, we should perhaps use "Word-Image IA" instead of "Word IA" as it links to an image-word pair. Just for the simplicity and readability, we use "Word IA".



Referring to the image that you just saw, what words would you use to describe your impression of the image?

For a guide, here are the kind of words that you may enter:
1. Must be a non-repeat and valid English word.
2. Nouns (e.g. ball, cat, apple), Verbs (e.g. run, play, jump), Adjectives (e.g. calm, exciting, beautiful)
3. Please use the character '-' if you enter a phrase with spaces, for e.g. sunny-afternoon
If you are unable to come up with all 5 words, please enter arbitrary words and select the button 'Don't Know' for the rest.

	●Don't Know	●Guess	●Maybe	●Confident	●Very Confident
	●Don't Know	●Guess	●Maybe	●Confident	●Very Confident
	●Don't Know	●Guess	●Maybe	●Confident	●Very Confident
	●Don't Know	●Guess	●Maybe	●Confident	●Very Confident
	●Don't Know	●Guess	●Maybe	●Confident	●Very Confident

You must enter 5 words and their confidence ratings to proceed. Try your best!

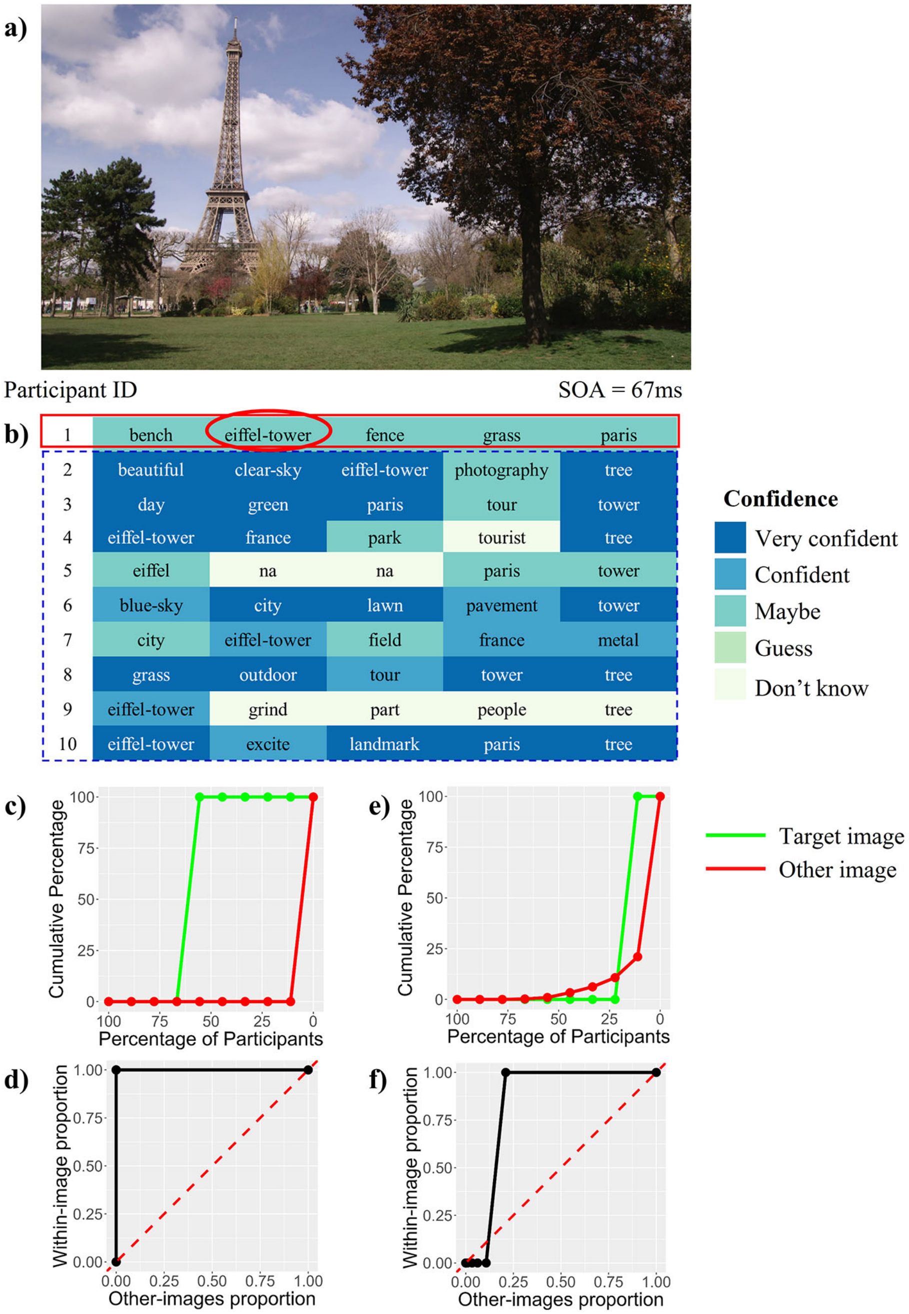
Figure 1. A single trial. a) Time course of a single trial. b) The response screen. The instruction said "For a guide, here are the kind of words that you may enter: 1, must be a non-repeat and valid English word; 2, nouns (e.g., ball, cat, apple), verbs (e.g., run, play, jump), adjectives (e.g., calm, exciting, beautiful); 3, please use the character '-' if you enter a phrase with spaces, e.g., sunny-afternoon; if you are unable to come up with all 5 words, please enter arbitrary words and select the button 'Don't know' for the rest".

Here, 'eiffel-tower' was reported by five out of nine "other" participants (= 56%).

Next, we estimate the baseline report frequency of the word "eiffel-tower" after seeing any one of the other images. Under SOA = 67 ms, no one reported the word "eiffel-tower" under the other 419 images.

We convert the count into percentages and calculate the cumulative percentages for within-image and other-image counts to obtain Figure 2(c) (green for the target image and red for the other images).

From this cumulative percentage, we construct the Receiver Operating Characteristic (ROC) curve (d) in Figure 2 and calculate the Area Under the ROC Curve (AUC). Supplementary Table 1 shows the cumulative percentage used to



1	bench	eiffel-tower	fence	grass	paris
2	beautiful	clear-sky	eiffel-tower	photography	tree
3	day	green	paris	tour	tower
4	eiffel-tower	france	park	tourist	tree
5	eiffel	na	na	paris	tower
6	blue-sky	city	lawn	pavement	tower
7	city	eiffel-tower	field	france	metal
8	grass	outdoor	tour	tower	tree
9	eiffel-tower	grind	part	people	tree
10	eiffel-tower	excite	landmark	paris	tree

Figure 2. The example Calculation of Word IA. a) The image used in this example. b) The 50 words reported by 10 participants, color-coded by confidence ratings. SOA = 67 ms. c) Red line: The cumulative percentage (y-axis) of the other images (i.e., all other 419 images we tested) as a function of the percentage of participants who reported the target word (x-axis, note in the descending order from 100% from left to 0% to the right). Here the target word is “eiffel-tower”. This red curve serves as the baseline report frequency of the target word. The value on this curve doesn’t have to be 0 or 1, it jumps from 0 to 1 because no one reported “eiffel-tower” under the other images. Green line: it shows the same for the target image, which jumps from 0 to 1 at X = 56%, reflecting the fact that 5 out of 9 “other” participants who saw this image for 67 ms reported “eiffel-tower”. d) The Receiver Operating Characteristic (ROC) curve constructed based on two cumulative curves in c. To construct the ROC, we shifted the criterion. We plotted the cumulative percentage for a pair of the within-image (green) as y-coordinate and other-images (red) as x-coordinate across all criteria. e) & f) Another example for the target word “grass” after viewing the same image (a), whose Word IA = 0.84. They are constructed in the same way as (c) and (d). This example shows values on the red line are not binary.

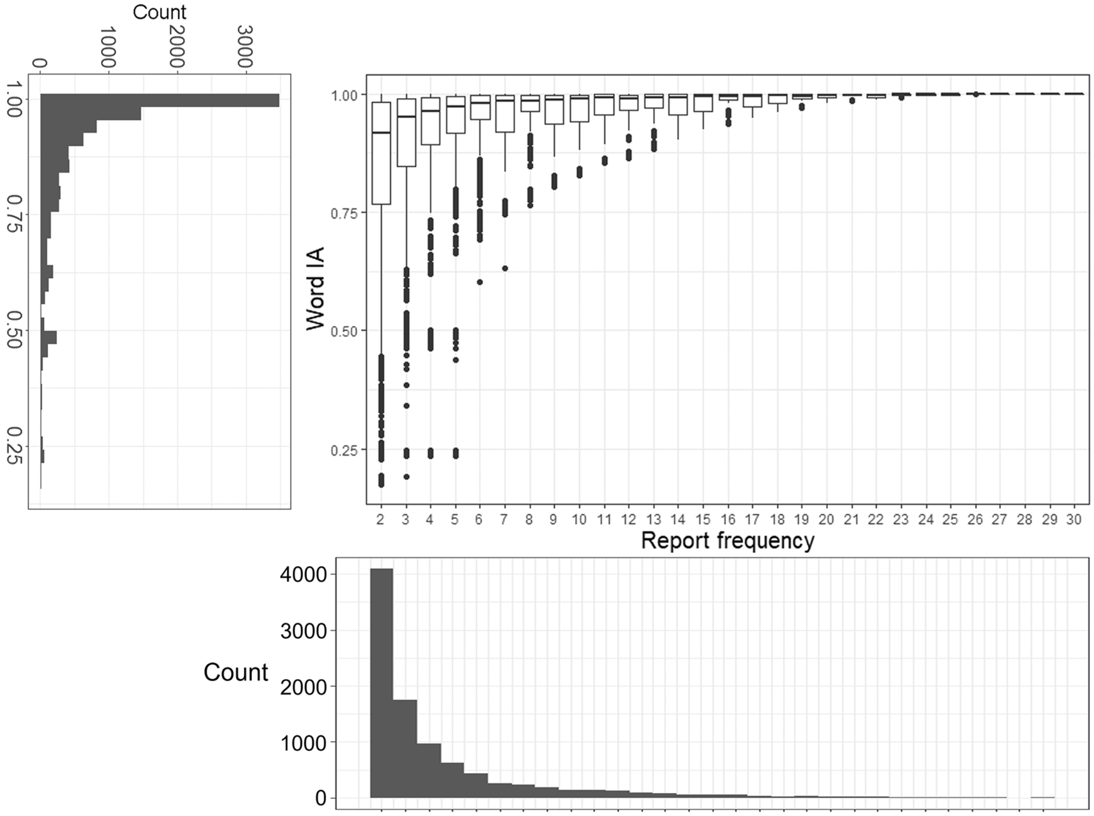


Figure 3. The distribution of Word IA and report frequency of each word. X axis is the report frequency of a word under a certain image (maximum = 30). Y axis is the Word IA values. The overall distribution of report frequency and Word IA is shown next to the axis.

compute Figure 2(c) and (d). Figure 2(e) and (f) shows the cumulative curves and ROC curve of another response word "grass" after viewing Figure 2(a).

Finally, we repeat the steps above for each "eiffel-tower" reported by different participants under SOA = 67 ms and calculate the final Word IA value (SOA = 67 ms) of "eiffel-tower" under (a) in Figure 2 by averaging the AUC values.

Note that a word must be reported by at least one "other participant" (i.e., reported by two or more people in total) under the target image to have a valid Word IA value. The words that were reported by only one person are called "rarely reported words". We will explain more later.

Similar to Word IA for each SOA, we can also calculate Word IA across SOA (see Methods: Calculation of Word IA across SOA for details).[2] In the following sections, we will first introduce the results from Word IA across SOA. This measures the specificity of response words robustly based on as many observations for each image. After that, we will compare Word IA between SOAs.

Word IA are very high across images

As we mentioned above, Word IA ranges from 0 to 1. Most words that were reported had very high Word IA (close to 1) across SOAs. Over 9,463 words and 420 images, Word IA was distributed with the mean (± std) as 0.89 (± 0.16) and median (25%, 75%) as 0.96 (25%-tile 0.85 to 75%-tile 0.99) (Figure 3). Thus, upon seeing a totally unexpectable image, people freely report five unique words that are rarely used to describe any other images in our set.

[2] We described our motivations and rationale behind our definition of IA over other simpler methods, such as ratio of reported words for the target vs all images in Supplementary Material (see https://osf.io/nvfhs).

A traditional account of rapid scene experience predicts that people perceive a global and coarse description of the image upon brief seeing. If participants' vocabulary is strongly confined to those global and coarse descriptions, Word IA should be low because those descriptions would be shared across images.

However, upon examination of each reported word in Figure 2 as well as Figure 3's distribution, this is unlikely the case. First, Figure 2(b) contains the words that would be normally regarded as global and coarse gist descriptions (e.g., clear-sky) but also local and specific words (e.g., eiffel-tower, paris). While the former types of descriptors are expected to be reported for the rest of images, the latter are unlikely to be reported. Thus, Word IA should be lower for the former and higher for the latter. Very high Word IA in Figure 3 implies that participants' vocabulary is not confined to global and coarse descriptions.

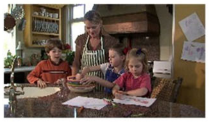
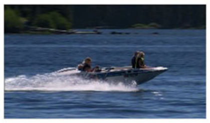
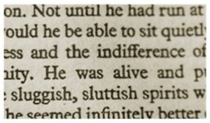
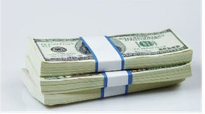
To understand how the nature of the target image affected Word IA, we calculated the mean of Word IA (across SOA) for all words reported on a particular image. We call it Image IA. Image IA estimates the consistency and selectivity of the reported words for a particular image.

Figure 4(a) and (b) list the five images that had the highest (and lowest) Image IA (across SOAs) among 420 images, respectively.

The five images in Figure 4(a) had the mean Word IAs across all reported words to be 0.98 to 1. This means that a set of reported words for each of these images were extremely specific and almost never used to describe any of the rest of the images.

For example, the top image with Image IA = 1 was the letter array, typically used in the so-called Sperling paradigm in psychology (Sperling, 1960). This type of image is often used to demonstrate the limit of what people can report upon seeing an image in a brief moment. Here, the reported words are almost always a single letter (e.g., U, P, L, etc) that was actually contained in the array. Note that these words were reported spontaneously without warning participants that they will be tested with these types of arrays. In fact, if anything, this image was included among 20 other natural scene images, thus participants would have had very low expectation in seeing anything like this (see Discussion: Elimination of expectation).

a)

Image IA	Rarely reported	Image	5 words with highest IA (All word IA = 1)		
			Word	Freq.	Conf.
1	64%	U G J X P D M B F C A L	U P L G J	10 5 4 4 3	4.1 4 3.3 3.8 4
0.99	67%		Kid Mum Bake Apron Daughter	9 9 7 4 3	4.7 4.4 4.3 3 4.7
0.99	79%		Boat Lake Fast Speedboat Life-vest	24 8 5 3 2	4.6 3.9 4.2 4.7 3
0.99	67%	 on. Not until he had run at ould he be able to sit quietl ess and the indifference of ity. He was alive and pu sluggish, sluttish spirits w he seemed infinitely better	Word Text Print Newspaper English	19 11 8 4 3	4.6 4.5 4.5 3.5 4.3
0.98	78%		Money Stack Dollar Cash Bill	24 11 11 8 6	4.6 4.6 3.9 5 4.7

b)

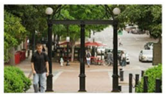
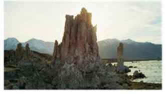
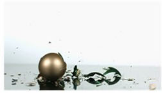
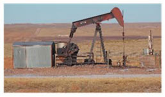
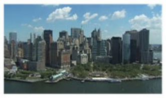
Image IA	Rarely reported	Image	5 words with highest IA		5 words with lowest IA	
			Word	IA (Fre/Con)	Word	IA (Fre/Con)
0.78	76%		Sidewalk Street Park Public Wait	1 (5/3.8) 0.99 (11/4.3) 0.99 (8/3.8) 0.99 (2/4) 0.98 (2/4)	Calm Na Fast Arbitrary Tree	0.24 (2/3.5) 0.25 (4/1) 0.45 (2/1) 0.49 (3/1) 0.61 (3/3.3)
0.79	79%		Mountain Rock Tall Peak Calm	1 (18/3.6) 1 (17/4.2) 1 (4/3.8) 1 (2/4.5) 0.99 (5/3.8)	Na Nature None Tree Water	0.24 (3/1) 0.34 (2/3.5) 0.46 (2/1) 0.57 (2/4) 0.64 (2/4)
0.79	78%		Ball Gold Graphic Ornament Reflection	1 (15/3.8) 1 (6/3.8) 1 (2/3.5) 0.99 (2/3.5) 0.99 (2/2.5)	Na Sea Paint People Not-guess	0.24 (3/1) 0.4 (2/4) 0.45 (2/2) 0.47 (2/3.5) 0.5 (4/1)
0.79	68%		Oil Drill Rig Crane Dig	1 (8/4.3) 1 (6/4.3) 1 (4/4.8) 1 (3/2.7) 1 (2/3)	Na Grass Field Man People	0.24 (2/1) 0.34 (2/4) 0.44 (2/4.5) 0.47 (2/4.5) 0.47 (2/3.5)
0.80	71%		Building Skyline City Skyscraper Natural	1 (13/4.3) 1 (4/4.8) 0.99 (12/3.8) 0.99 (2/5) 0.98 (3/2)	Na Man Flat Sky Beautiful	0.23 (2/1) 0.47 (2/2.5) 0.49 (2/3.5) 0.53 (3/2.7) 0.6 (2/2.5)

Figure 4. **Images with highest (a) and lowest (b) Image IA.** Five words with highest and lowest Word IA are also shown. The Word IA of all response words in (a) is 1. The figure shows that both images with highest and lowest Image IA have words whose Word IA is close to 1, accompanied with high confidence rating. The images in (b) have lower Image IA because they also have words with low Word IA. Fre. means the frequency of the word being reported (maximum = 30). Con. means the mean confidence ratings of the word across participants who reported it.

The rest of the top images with high Image IAs were all natural scenes. Among those words, although there are a few arguable ones, such as Text (11 out of 30 people reported) and Word (19), most words were local and detailed, including: Mum (9), Apron (4), Boat (24), Speedboat (3), and Life-vest (2), etc.

Figure 4(b) lists the five images that had the lowest Image IAs, ranging from 0.78 to 0.8. These scenes appear difficult to describe in words compared to the top images in Figure 4(a). Nonetheless, each image was reported with words whose Word IA was close to 1. This means that people agreed in describing these images with these highly specific words. Some of those words might be argued as global and coarse, such as Street (11), Tall (4), Natural (3) and City (12), whereas other words are more local and detailed, such as Ball (15), Crane (3), Dig (2), Drill (6), and Skyscraper (2). For these global and coarse words in Figure 4(b), we surmise that there are not many salient visual objects for detailed reports in the images. Thus, participants resorted to coarse descriptions, which are strongly shared by other people. As a result, the report frequency of these coarse descriptions are much higher than those for the rest of images, attaining high Word IA values.

Lower Image IA for Figure 4(b) is due to the result of many words that have lower Word IA, which were included to compute Image IA. There are some global and coarse words that can also be used to describe other images, such as Nature (2), Water (2), Sea (2), and Sky (3). There are also some meaningless words whose confidence rating is 1, such as “Na”, “Arbitrary”, and “None”, which are shared across images (i.e., low Word IA). The presence of these words indicates that some participants found it hard to describe the images using five distinct words. This is either due to the difficulty of the description of the image per se or due to the restriction of image viewing time, which we will investigate next.

Comparing the words whose Word IA is close to 1 in Figure 4(a) with those in Figure 4(b), we notice that the confidence ratings in Figure 4(a) tend to be higher than those in Figure 4(b). But the confidence ratings of the words with highest Word IA are much higher than those of the words with lowest Word IA within Figure 4(b), implying some correlation between confidence ratings and Word IA, which we will investigate next as well.

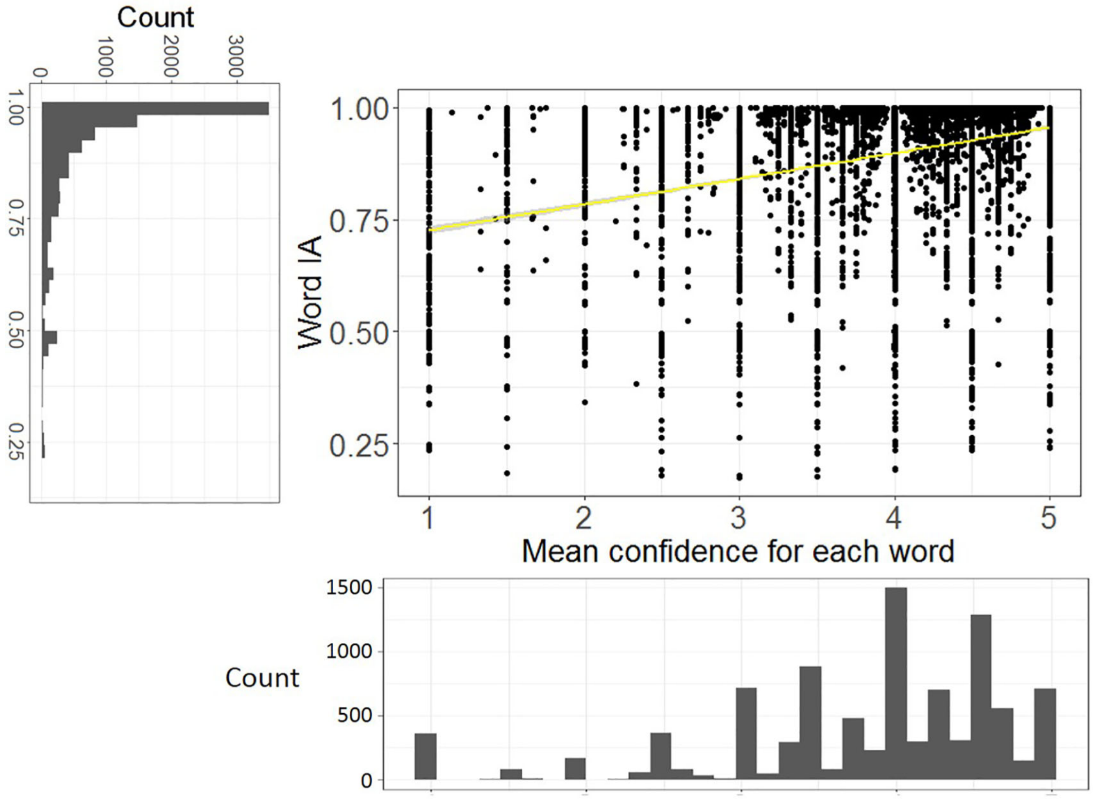


Figure 5. The higher Word IA (across SOA), the more confident people report with the words. X axis is the mean confidence (across participants who reported the word) for each word. Y axis is Word IA. 95% CI is also shown (gray shade).

We also analyzed the proportion of rarely reported words (the second column from the left for Figure 4(a) and (b). They were not included in the calculation of Image IA because their Word IA was not defined. If a word was reported by only one person, it could be truly a unique word to describe that particular image, but it could also be a random response that was not elicited by the stimulus images. A higher proportion of rarely reported words could indicate that it is harder to perceive describable information from the image. To test that, we computed the proportion of rarely reported words for each image, but we found no correlation between the proportion and Image IA, r = -.06, df = 418, p = .195, 95%CI = [-0.16, 0.03]. Whether rarely reported words are not random responses can be addressed in the future studies by recruiting a larger number of participants.

Word IA reflects mean confidence ratings for each word

As implied in Figure 4, high Word IA tends to be associated with high mean confidence ratings. If this correlation is reliable, Word IA can be taken as a proxy of conscious perception, reflecting metacognitive access to what participants see phenomenally (Matthews *et al.*, 2018; Seth *et al.*, 2008).

Figure 5 shows the scatter plot of Word IA (across SOA) and mean confidence rating for each word, with the distributions of those variables. There was a significant correlation between Word IA and mean confidence rating for each word, r = 0.33, df = 9461, p < .001, 95%CI = [0.32, 0.35]. Thus, our Word IA measure captures an aspect of conscious perception, that is, the degree of confidence they have in reporting what they see.

Even at the shortest SOA, Word IA correlates with confidence

It is plausible that the association between high Word IA and high confidence rating in Figures 4 and 5 are artefacts of our grouping of SOA. At longer SOAs (133 or 267 ms), it is likely that participants see details clearly and consciously with high confidence with high IA. At short SOA (67 ms), they might not see much and report random words with low

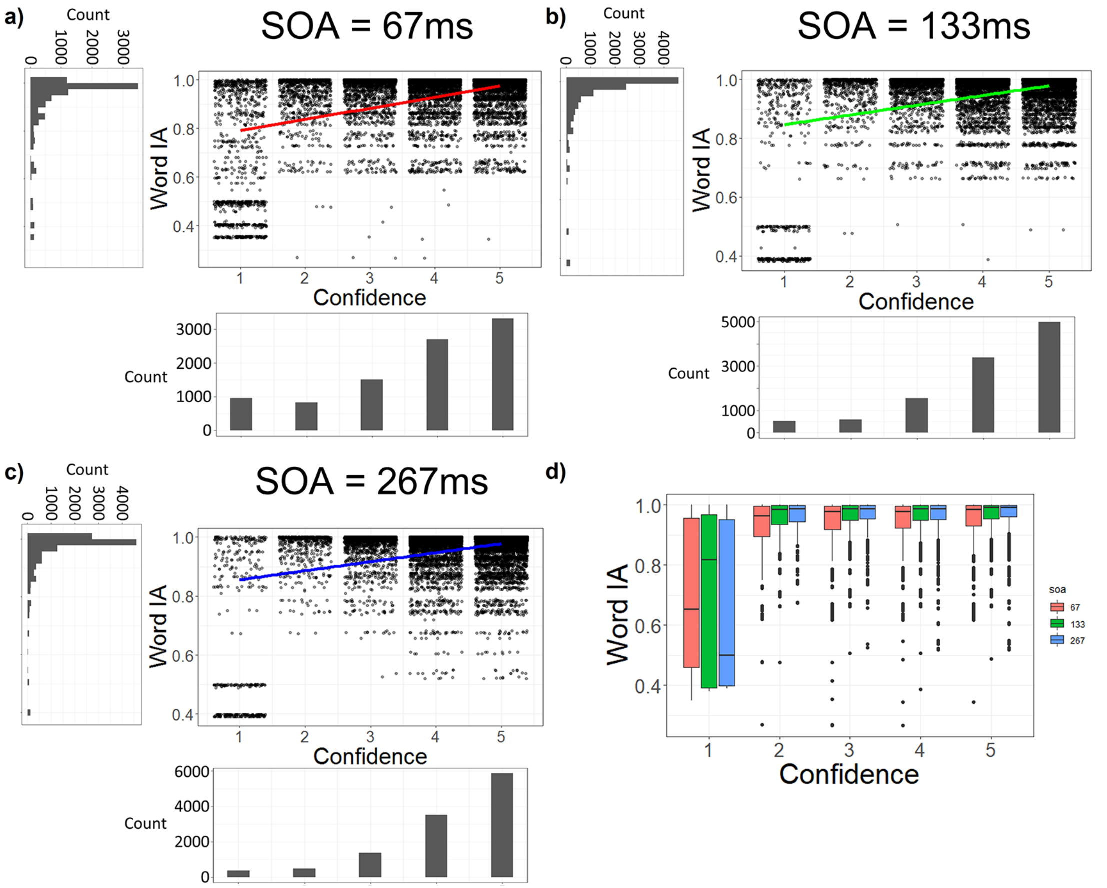


Figure 6. Confidence rating is correlated with Word IA at all three SOAs. a-c) The dot plots with jittering for confidence ratings. Note here, each dot represents a confidence rating from one participant for a given word. Word IA at SOA = 67 ms (a), 133 ms (b), and 267 ms (c). X axis is confidence rating. Y axis is Word IA. d) The boxplot showing Word IA for each SOA, grouped by confidence and SOA.

confidence with low IA. Such results would inflate the correlation between Word IA and confidence ratings. To rule out this possibility, we examined the correlation at each SOA.

However, even under the shortest SOA (= 67 ms), there was a significant correlation between Word IA and confidence rating: r = 0.43, df = 9330, p < .001, 95%CI = [0.41, 0.45] (Figure 6(a)). Figures 6(b) and (c) show the data for longer SOAs, which also showed significant correlations between confidence and Word IA: at SOA = 133 ms, r = 0.37, df = 11046, p < .001, 95%CI = [0.35, 0.38], and at SOA = 267 ms, r = 0.34, df = 11664, p < .001, 95%CI = [0.32, 0.35]. Therefore, confidence rating was correlated with Word IA at all three SOAs.

To control the difference among participants and images, we analysed the data with four multilevel regression models with the random effects being intercepts for participants and images. Our first model (Model 1) is the null model which only included the random effects of intercepts, adjusted Intraclass Correlation Coefficient (ICC) = 0.248, AIC = 57737, BIC = -57703. Our second model (Model 2) added the fixed effect of confidence rating on top of Model 1, adjusted ICC = 0.221, AIC = -61941, BIC = -61899. It showed that Word IA for each SOA can be explained by confidence rating, p < .001. The third model (Model 3) added the fixed effect of SOA on top of Model 2, adjusted ICC = 0.222, AIC = -62093, BIC = -62043. It showed that Word IA for each SOA can be explained by both confidence rating (p < .001) and SOA (p < .001). The fourth model (Model 4) is the full model which added the fixed effect of interaction between confidence rating and SOA on top of Model 3, adjusted ICC = 0.222, AIC = -62265, BIC = -62206. It showed that Word IA for each SOA can be explained by confidence rating (p < .001), SOA (p < .001), and the interaction between them (p < .001). According to the AIC and BIC of these models, Model 1 is the worst model, whereas Model 2, 3, 4 are similar, with Model 4 having the lowest AIC and BIC. However, according to the likelihood ratio test (Winter, 2013), there was a significant difference between Model 2 and 3, $\chi^2 = 153.81$, df = 1, p < .001. There was also a significant difference between Model 3 and 4, $\chi^2 = 173.75$, df = 1, p < .001. Therefore, Model 4 is the most preferred model, whose parameters are shown in Supplementary Table 2. The model shows that a higher SOA predicts higher Word IA.

The relationship between Word IA and SOA is visualised in Figure 6(d). Interpretation of results for confidence = 1 (about 5% of all responses) is somewhat complicated, because we explicitly asked participants to enter random words with confidence = 1 when they can't give five words. In fact, entries such as "none" and "NA" are often entered with confidence = 1. For SOA = 267 ms, we infer that the median Word IA is around 0.5 when confidence = 1, consistent with an idea that participants know that their report does not reflect the image. For SOA = 67 ms and 133 ms, however, the median Word IA are higher than 0.5 (0.65 and 0.82, respectively), possibly reflecting fleeting impressions that participants felt, which were actually shared with other participants in specific and selective ways as captured by Word IA. When confidence is higher than 1, Word IAs were nearly saturated (median > 0.95) for all SOAs. When we exclude responses with confidence of 1, there is still a significant positive correlation between SOAs and Word IA (p < 0.001, found by multilevel regression).

Beyond global and coarse reports revealed by Word IA – analysis of the highly similar image pairs and artificial images

Finally, we present somewhat unexpected, yet striking findings based on further analyses of two sets of images: highly similar image pairs and artificial image pairs.

As explained in Methods, we filtered the identical images over the available image sets (screenshots from videos) with both automatic and manual procedures. However, unintentionally, we retained 22 pairs and two trios of highly similar but slightly different natural images; see a) in Figure 7 and Supplementary Figure 2. Rather than removing the data from these images, we analyzed them with our analysis method to test its validity.

For simplicity, we removed one image from each trio and obtained 24 pairs of natural images in total. To quantify the degree of the similarity of free reports between two similar images, we computed "sum of square of error (SSE)". In short, SSE quantifies the similarity of the frequency of reports between two images presented to two separate populations of participants (see (b) in Figure 7). SSE is 0 when the reports are identical. It is expected to become large if it is computed between two different images.

The green line in (c) in Figure 7 demonstrates that two groups of participants gave a very similar set of words for these two similar images. This contrasts with the baseline (blue), where SSE is computed between randomly paired two images (for details, see Methods - Data analysis - shuffle the pairs of similar images). The fact that two distinct groups of N=30 participants responded highly similarly to the pairs of highly similar images lends support to our assumption for the IA measure: upon seeing similar images, a population of people give similar responses even when measured with an unconstrained free-report paradigm.

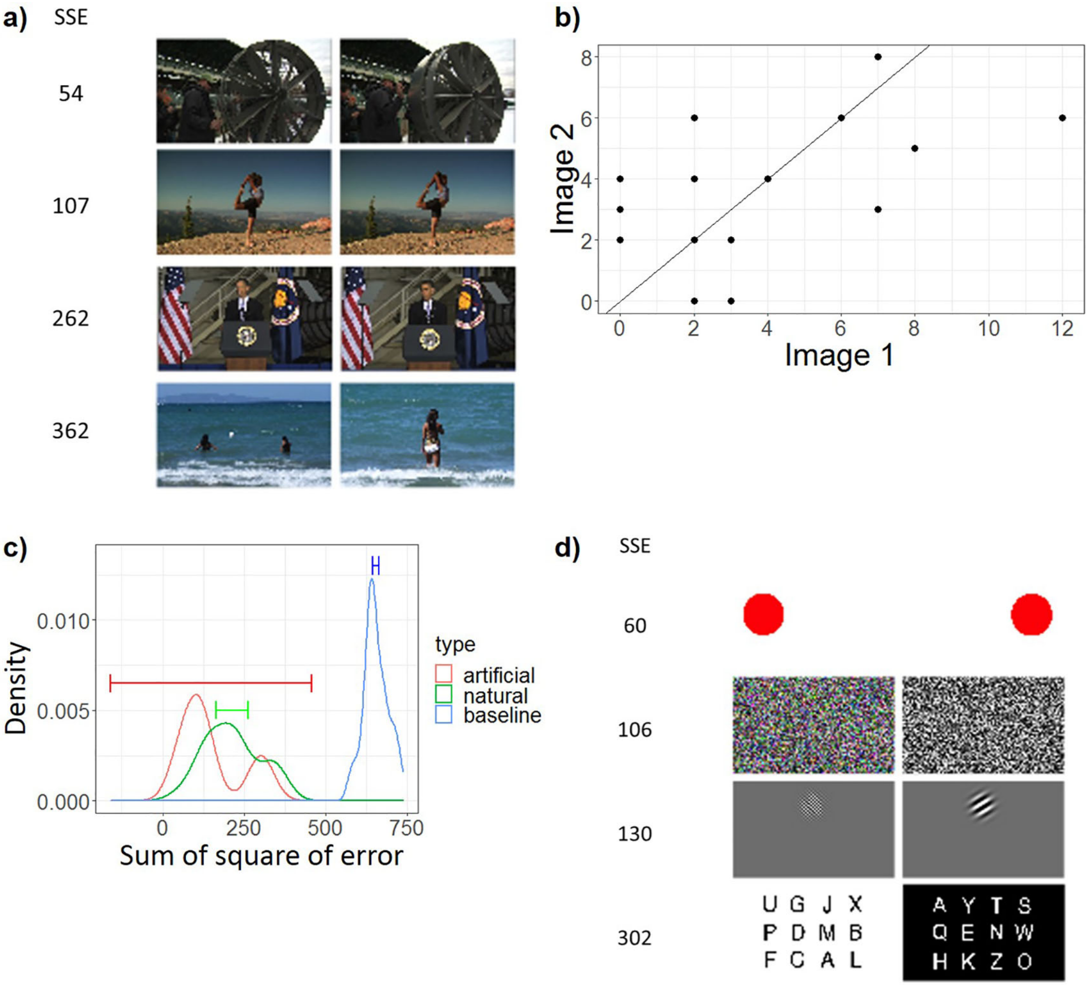


Figure 7. The Analysis of Highly Similar Images. a) Four example pairs of the highly similar natural images, including the pairs with lowest (54) and highest (362) Sum of Square of Error (SSE). For all highly similar natural images, see Supplementary Figure 2. b) An example plot of the frequency of all the unique response words from a pair of highly similar images. X axis is the frequency of response words from one image obtained from a set of N=30 participants. Y axis is the frequency of response words from the other similar image obtained from a separate set of N=30 participants. Each data point means a response word. The solid line is y = x, which means perfect consistency. c) The distributions of Sum of Square of Error (SSE) of pairs of two similar natural images (green) and artificial images (red). As baseline (blue), we computed SSE for randomly selected two images. X axis is SSE, which is the sum of squared distance between each data point and the y = x line in (b). Y axis is the probability density. 99%CI is shown by the line segment. d) SSE computed for four pairs of artificial images.

Secondly, we analysed the four pairs of artificial images (see (d) in Figure 7 and Supplementary Figure 3 for details), which have been used in previous psychophysical experiments.[3] Unlike natural images, we suspected that if these artificial images are presented to participants without any warning or expectation, they may not be able to see the contents very clearly (see Discussion). As we expected, participants did not really spontaneously report particular details of each simple image, which differentiates them (e.g., left vs. right, colored vs. black-and-white, high vs. low spatial frequency, left- vs. right-tilted). This is totally understandable given that these features only make sense when these pairs were presented in contrast and participants are invited to report their differences. Their spontaneous reports are more to do with the common attributes for these pairs, resulting in lower SSE.

One exception to this rule was the pair of the Sperling letter array, one of which was indeed featured in Figure 4(a), which achieved the highest Image IA among all the tested images. We expected that the dominant responses for these types of

[3] Because we had 412 natural scene images and only 8 artificial images (4 pairs), reported results about Word IA reflect mostly for natural images. We checked if any of the results changed significantly after removing artificial images, but they did not.

the images would be global and broad descriptions, such as "letters", "alphabets", but rarely the actual single letters, such as "U" or "A" (Haun *et al.*, 2017). Contrary to our expectation, each letter was reported rather frequently as shown in Figure 4(a) and Supplementary Figure 3. This is quite remarkable, given that in our experiment, participants were never told about this type of stimuli (see Discussion: Elimination of expectation). As a result, the SSE between the paired Sperling images (= 302) was much larger than the other three artificial image pairs (60-130).

Discussion

In the present study, we used a free-report paradigm (Figure 1) to examine a widely-supported view on rapid scene experience. In a traditional view, people are expected to process and report coarse, but not detailed, information about a briefly presented scene. Rather than presuming what words should qualify as "correct" answers, we used the words that other participants reported about the same image as the expected report. Based on this notion, we proposed a novel measure, termed "intersubjective agreement" (IA) (Figure 2). Word IA quantifies the specificity of a response word, by comparing the frequency of the response word under a given image with that under all other images. The frequency of the latter serves as the baseline report of the response word. With our IA measure, we demonstrated that participants report words that are maximally specific (IA = 1) across SOAs (as short as 67 ms) (Figures 3, 4). Further, we demonstrated Word IA computed across SOA and at a given SOA showed correlation with confidence ratings, thus reflecting the degree of conscious access (Figures 5, 6). The analysis of highly similar images (Figure 7) assured the validity of our paradigm. Our results challenge the notion that reports on a briefly viewed scene are coarse and not detailed.

Methodological novelties

Free-report paradigm with our novel IA measure. While the forced-choice tasks have revealed a great deal about various limits and scopes of the human visual system in rapid vision, they do not allow us to make inferences as to whether participants perceive more than what experimenters expected. Importantly, our paradigm is free from such a restriction imposed by us, the experimenters.

Our study is not the first application of the free-report paradigm to examine the nature of rapid visual perception. In a pioneering study, FeiFei *et al.* (2007) asked participants to write a short paragraph to describe the image of everyday scenes after briefly viewing it for 27 ms to 500 ms (masked). The responses were then analyzed by 5 scorers. Unfortunately, they introduced experimenters' bias at this evaluation stage. Specifically, the experimenters asked the scorers to evaluate if the response paragraph contained objects that they, the experimenters, defined for the scorers. As such, the objects that were not included in the experimenters' preset objects were not analysed.

Another issue of such assessor-based scoring is that the subjective experience of scorers is quite different from that of participants. In FeiFei *et al.* (2007), the raters saw the target image for a long duration moving their eyes to inspect various parts of the images. However, the participants in the main experiment saw the images briefly with minimal saccades. Thus, we cannot interpret what the participants actually experienced based on the scorers' evaluations.

We minimised these sources of experimenters' biases and confounds. We analysed all that participants reported about their experience based on what other participants reported in the same viewing condition. Importantly, our novel strategy can scale like forced-choice paradigms without manual inspection. With more participants per image set, our estimate becomes more robust.

Although not directly comparable, some of the previous forced-choice paradigms studied aspects of rapid scene experience. For example, Biederman *et al.* (2006) and Larson *et al.* (2014) showed that participants' responses were more accurate under a longer SOA. Similarly, in our study, Word IA for each SOA was positively associated with SOA (Supplementary Table 2). Importantly, this increase in IA was also accompanied with the increase in confidence and frequency of reports, consistent with previous studies as well (e.g., Fu *et al.*, 2016).

To enhance these strengths, we employed an online experimental platform to recruit a large number of participants and to use a large set of stimuli (Qianchen *et al.*, 2022). They are important prerequisites of our novel IA measure and our free-report paradigm. Note that in our preliminary experiment, we did not see any qualitative difference in the results between in-lab participants (N=10) and online participants (see Supplementary Figure 4).

Elimination of expectation. Another important extension is that our paradigm does not allow participants to expect what category of images will be shown. In previous studies, participants were told that the images would be one of a few categories (e.g., animal vs. non-animal; Fabre-Thorpe *et al.*, 2001). High accuracy in rapid scene processing (e.g., Evans & Treisman, 2005; Fabre-Thorpe *et al.*, 2001) may depend on expectation (McLean *et al.*, 2021; Sun *et al.*, 2017). In our paradigm participants could not expect the images, yet they still reported impressive details of the briefly presented images.

Confidence rating correlates with Word IA. Surprisingly, the detailed information accompanied with high confidence across images (Figures 4, 5 and 6).

The results with Sperling arrays (see Figure 4) were rather notable. Without any expectations, participants reported specific letters, which goes against an idea that conscious perception strongly depends on expectation (Mack *et al.*, 2017; Pinto *et al.*, 2015). Note these specific letters reported with high confidence were the ones that were included in the array.

Speaking of confidence rating, we are not aware of other studies that asked participants to provide confidence ratings together with words in a free-report paradigm. Our study is perhaps the first that combined a free-report and confidence rating. The words with high confidence were more specific and shared between participants (Figures 4, 5, and 6). If we consider what participants reported in this study as a "gist" of a scene, we can interpret it to mean that gists are something that we can metacognitively monitor and consciously access. In light of past studies that showed gists can be grasped without attentional amplification (Li *et al.*, 2002; Mack & Rock, 1998), this further supports the notion that conscious perception does not always require attentional amplification (Koch & Tsuchiya, 2007; Tsuchiya & Koch, 2016). Our study extends this notion with freely-reported image contents without expectation.

Importantly, we observed highly specific words (IA = 1) even in SOA = 67 ms condition (77% of images). Many specific words came with the highest confidence of 5. Non-specific words (IA < 0.6) were observed much less frequently across SOAs (30% for SOA = 67 ms, 19% for SOA = 133 ms, 17% for SOA = 267 ms). We predict this pattern of results would not be observed if the image was strongly masked to make the image invisible (Kaunitz *et al.*, 2011). Taken together, our findings require the revision of rapid scene experience from traditional "coarse and nonspecific" to "detailed and specific even without expectation".

Repetitions of images

What is the potential source of the discrepancy between ours and traditional findings? Given that we obtained similarly high IA for both artificial (letters) and natural stimuli, the source of discrepancy is unlikely to be solely due to the type of the stimuli.

Instead, we point out another important feature of our task. In our task, we never repeated the same image for a given participant. We suspect this is one source of the discordance as previous studies tended to use the same stimuli across trials (e.g., Kimchi, 1992; Navon, 1977). As Endress and Potter (2014) suggested, repeated presentation causes prospective interference, possibly biasing our attention and expectation to the features that distinguish trials (see also Kaunitz *et al.*, 2016; Qianchen *et al.*, 2022). It is worth revisiting previous studies but using stimuli that are never repeated across trials.

Presentation duration

What is the minimal duration that is required for such rapid recognition? When we designed this study, we were not confident if the online study can achieve reliable presentation of short duration stimuli. Thus, we opted for going rather conservative durations of 67 ms as minimal. Further, we did not expect that participants could make detailed word reports that correlate with confidence. Therefore, we did not push the limit. Future studies can test much shorter duration.

We performed some preliminary experiments with shorter SOAs online and confirmed its feasibility. To conduct such a study in the future, however, we recommend to confirm the results with in-lab participants to some extent. We also suggest that shorter SOA trials should be intermixed with longer SOA trials, so that participants can set an appropriate threshold for reports (Lin & Murray, 2014). Such experiments will address an important question on the temporal limit of intersubjectively shared experience upon viewing rapid visual scenes.

Future directions

Our novel paradigm with IA measure is highly adaptable for different research purposes. For example, any future researchers can use our current data as the normative data (our data is fully and freely available online for both the image sets [https://osf.io/q2cr8/] and the response words [https://osf.io/7spxd/]). Using them as the baseline, it would be easy to perform a study to test if some population of patients with known mental disorders, such as schizophrenia, would see and report the same words as the healthy population. Future studies could also utilize different stimulus sets to address different questions. For example, if we change the nature of the stimulus set into emotional movies (e.g., Cowen & Keltner, 2017), we can characterize commonality and differences of the nature of emotional experiences using free report paradigms without imposing any theoretical assumptions introduced by us, the experimenters. Also, the emotional intensity within each movie can be calculated using normative ratings of lexico-semantic factors (e.g., valence and arousal; Mohammad, 2018).

Finally, our paper has some implications to the theoretical debate on whether conscious experience is restricted by cognitive access or overflows such restriction (Block, 2007; Bronfman *et al.*, 2019). Previous debate on the overflow issue was almost always focused on the artificial stimulus situation introduced by Sperling (1960). Our paper contributes to this debate in two fresh ways. First, since the Sperling's initial introduction of the paradigm, it has been almost always "assumed" that what participants see in the "whole-report" condition is just a gist, such as "seeing everything", "alphabets", and "letters" (Haun *et al.*, 2017). Our empirical free-report challenges this notion. Even without any warning or expectation, participants report specific letters rather than generic gists (e.g., Figure 4a). Second, our paradigm opens up a possibility to examine the overflow issue with naturalistic stimuli. For that, we would need to develop equivalent experimental contrasts between the whole- vs. partial reports as in Sperling (1960), which is not impossible.

Conclusion

Contrary to a widely supported view that participants are more likely to perceive coarse information from a briefly presented natural scene image, we found that participants reported highly detailed descriptive words with high confidence. Our findings show that even without expectation, a brief glance of scene images is enough to provide us with a vivid conscious experience which can be metacognitively monitored. Our novel paradigm and measures can lead to an exciting avenue for future research that finds intersection on sensory psychology, big data, cognitive-linguistic and consciousness research.

Methods

Participants

We recruited 801 online participants from Amazon's Mechanical Turk (MTurk) who had a Human Intelligence Task (HIT) approval rating of more than 97% and more than 5,000 approved HITs. Among them, 131 participants did not complete the task and were excluded, leaving 670 valid participants. All the analysis presented in this paper is based on the first 30 participants for each of 20 groups (see below for details), that is 600 participants. MTurk has demonstrated its reliability and validity (Sheehan, 2017) by replicating lab effects in many research areas (Amir *et al.*, 2012; Crump *et al.*, 2013; Klein *et al.*, 2014; Paolacci *et al.*, 2010). Participants read the online explanatory statement and gave informed consent by pressing a key before they began the experiment. Each participant received US$3 as compensation for their time (30 minutes). Upon completion of the experiment, participants answered a few demographic questionnaires. Within the questionnaires, participants were allowed to answer 'NA' if they wished. We used TurkPrime (Litman *et al.*, 2017) to manage the recruitment and payment of participants from MTurk. Supplementary Figure 5 shows the demographics (sex, age, nationality, first and second language, number of years speaking English) of the participants. Ethics approval was obtained from the Monash University Human Research Ethics Committee (approved project ID 17674).

Apparatus

To create the online experiment, we used InquisitLab and InquisitWeb (version 5; Software, 2018). Inquisit's software provides highly accurate stimulus presentation time (within milliseconds; De Clercq *et al.*, 2003). To minimise the effect of a slow network, we asked participants to download the software package of InquisitPlayer, which then downloaded all experimental files to their computer before the experiment. During the experiment, InquisitWeb blocked the participant from interacting with other software programmes on the same computer, reducing the likelihood of poor data quality due to distractions. The functions of above software can be equivalently performed by PsychoPy and Pavlovia.

Stimuli

One of the authors (SN) captured 9120 still-frame naturalistic images from a series of online videos at Videoblocks. Each coloured image was in JPEG format with a resolution of 1920 × 1080 pixels. As a video contained multiple still-frame images captured within a one-second interval, some pictures were highly similar. After filtering the identical images, we obtained 570 images.

Among these images, we randomly selected 415 images for our experiment (shown in Supplementary Figures 6a - 6f). Among them, three images were used as practice images. In addition, we included eight images (four pairs) of artificially generated images to represent typical stimuli used in psychophysics experiments (i.e., random noise, Gabor patch, Sperling letter array (Sperling, 1960), a red circle; see (d) in Figure 7 and Supplementary Figure 3). As a result, we had 420 experimental images and three practice images. We randomly divided the experimental images into 20 blocks each with 21 images. Each participant saw the three practice images and a block of 21 images.

Procedure

Upon downloading the InquisitPlayer, the experiment began with a consent form. After giving the consent, participants received an instruction on how to perform the task. They were told that in each trial, they would briefly see an image. Their task was to enter five valid English words without any repeats to describe their impression of the image. They were allowed to use nouns, verbs, and adjectives. They were reminded of this instruction while they typed the words in each trial (see (b) in Figure 1). Upon these instructions, participants practiced the task for three trials (with the same three images across participants). The stimulus onset asynchronies (SOAs) for these trials were randomly assigned from 67, 133, or 267 ms, one trial each. These images were not analysed.

Each participant was tested in one block that contained 21 trials. Figure 1 shows the time course of a single trial, which started with 1 second of fixation, followed by a target image. The target image was shown for a variable duration until the five successive masking images, each presented for 60 ms. SOA between the target and the first mask was either 67 ms or 133 ms or 267 ms. Each mask was an image with 1920 × 1080 pixels, which we constructed by filling it with 16 × 16 pixel image patches randomly taken from the 420 experimental images.

After the mask, participants were presented with the response screen (see (b) in Figure 1). Participants gave response words as instructed and rated how confident they were in seeing each word in the image (1: Don't Know, 2: Guess, 3: Maybe, 4: Confident, 5: Very Confident). The response screen restricted participants to enter only alphabets and numerics. They were not allowed to use the empty space and asked to use a hyphen ('-') instead. If participants could not come up with 5 words, they were instructed to enter an arbitrary word with "Don't know" as the confidence rating.[4]

For each participant, we tested seven trials for each SOA (66 ms, 133 ms, and 267 ms), whose order was randomised across participants. Across 21 trials, participants never saw the same image more than once. For a given block of 21 images, we aimed to test a cohort of 30 participants so that each of the 21 images was presented to at least 10 participants at each SOA. Sometimes, MTurk recruited more than 30 participants. When this occurred, we took the first 30 participants' data for the subsequent analyses. For this reason, 70 participants were excluded so we had 600 participants in our data analysis.

Data cleaning

We performed minimal curation on the reported word as we did not want to inject our own bias into the data set. We provided all raw data so that researchers can examine the effects of any curation on our result. Our minimal curation procedure included the following.

If a participant entered the same word for a single image more than once, then we took only the first one with its confidence rating, and removed the second response (or later) from the analysis.

For each word, we applied three pre-processing steps. First, we converted the words to lowercase using the R package tm version 0.7-8). Second, we corrected misspelled words semi-automatically using the R library hunspell version 3.0) package (if the spell checker returned with one suggestion, we adopted it if relevant. Otherwise, we manually inspected the alternate suggestions and picked the most appropriate one). Finally, we performed word-lemmatisation using the R-package textstem version 0.1.4). Word-lemmatisation grouped similar words (e.g. plurals, verb tense), such as "child" and "children", "walked" and "walk", into a single base-form word so that the subsequent analysis would consider them as the same word.

After all these steps, we obtained 63,000 word responses (i.e., 420 images * 3 SOAs * 10 participants * 5 words) for the subsequent analysis. After combining the same words under the same images and the same SOA, we had 41,170 unique words. These word responses along with associated confidence ratings are available as a CSV file at https://osf.io/7spxd/.

Data analysis

Calculation of Word IA for each SOA. For each target image at a given SOA, there were 10 unique participants who saw it. For each of 50 reported words, we went through the following processes. First, we counted the number of participants out of the rest of nine participants who reported a target word. Second, for each of all other 419 images, at the same SOA condition, we counted the number of nine participants who reported the target word. We excluded one participant that has the same participant order for the other images to equate the number of the participants for this counting to be 9 per SOA. Third, based on the number of these participants, we computed % of report and % of cumulative report for both the target image and all other images (Supplementary Table 1). From these, we applied the logic of Signal Detection Theory (Green & Swets, 1966; Macmillan & Creelman, 2004). We used the cumulative frequency of reports to construct the receiver operating characteristic (ROC; see (c) and (b) in Figure 2. Fourth, we calculated the area under the

[4] In a pilot task, we asked participants to report up to 10 words per image. This task turned out to be difficult or tedious for many participants. Most participants seemed to be comfortable in reporting roughly up to 5 words per image. As we wanted to obtain as many words per image as possible, we settled on the 5 words.

ROC curve (AUC). If there were more than one participant who reported the same target word, we went through the same procedure and computed the mean AUC across them as Word IA at that SOA.

As we explained in the Results section, if the target word was reported only by one participant among 10 participants, it is called “rarely reported word” and its Word IA is not defined. Among the 41,170 unique (word, image, SOA) groups, 30,954 of them were only reported by one person and therefore didn’t have a defined Word IA, which left us 10,216 response words with a valid Word IA.

Calculation of Word IA across SOAs. We also computed Word IA by pooling the responses across SOAs. We did this to reduce the possibility of rarely reported words and to estimate the baseline rate of the reported words more robustly.

We performed the same analysis as Word IA for each SOA with the following exceptions. For each target image across SOAs, we had 30 unique participants who saw it, resulting in 150 reported words. For Word IA across SOAs, we counted the number of participants out of the rest of 27 participants who reported a target word. We removed one participant from each SOA group so that across SOAs analysis is not biased. Similarly, we did so for 27 participants to estimate the baseline with the other images. Other aspects of the calculation were the same as Word IA for each SOA.

Among the 30,911 unique (word, image) groups, 21,448 of them did not have a defined Word IA, which left us 9,463 response words that had a defined Word IA.

Exclusion of similar images in Word IA calculation. Although we tried to filter highly similar images, we noticed there were still some highly similar (but slightly different) image pairs in our image set (Supplementary Figure 2). While none of these pairs was presented to the same participant, word responses generated from these images could affect Word IA. For a given reported target word, the Word IA for one of the similar paired images will be generally lower because the target word is likely to be reported under the other paired image. To reduce the effect of this problem, when calculating the Word IA of responses to one of these images, we excluded the paired image from the baseline.

Multilevel regression models. The models were performed using the “lme4” package in R. The parameters were estimated using the maximum likelihood (ML) method.

Shuffle the pairs of similar images. To compare the consistency between pairs of highly similar images with random image pairs (in Results section: Beyond global and coarse reports revealed by Word IA), we constructed a “null distribution” as the baseline using a bootstrap method. We randomly selected one image from each of the 24 pairs and paired it with another image that was randomly selected from the 48 natural images (i.e., the 24 pairs of highly similar images) to obtain 24 random pairs of images. We then calculated the SSE (see Results section) of each pair and the mean of the 24 pairs. We repeated this process 100 times to obtain the “null distribution” of SSE.

Data availability

Underlying data

OSF: Human Annotations with Confidence Ratings on Natural Images. https://doi.org/10.17605/OSF.IO/7SPXD (Chuyin, *et al.*, 2021)

This project contains the following underlying data

gist_batch_v1_all_mt_raw.csv (the raw data)

gist_batch_v1_all_mt_summary.csv (all the participants who signed up the study on MTurk)

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Extended data

OSF: Natural Scene Images for the IA paper. https://doi.org/10.17605/OSF.IO/Q2CR8 (Nishimoto *et al.*, 2021).

This project contains the following extended data

587 natural scene images in JPEG.

8 artificial stimulus images in JPEG.

OSF: Supplementary materials for the IA paper. https://doi.org/10.17605/OSF.IO/U6QPM (Chuyin, *et al*., 2021).

This project contains the following extended data

Supplementary Tables and Figures.pdf (Supplementary Tables 1-2 and Supplementary Figures 1-7)

Supplementary tables.xlsx (Supplementary Tables in Excel)

Motivation of IA.pdf

Experiment code available from: https://doi.org/10.5281/zenodo.5794712 (Koh & mluu0010, 2021)

Processed data and analysis code available from: https://doi.org/10.5281/zenodo.5796303 (Chuyin & Koh, 2021)

Acknowledgement

A previous version of this article is published on PsyArXiv: https://psyarxiv.com/d2s38/ (https://doi.org/10.31234/osf.io/d2s38).

References

Amir O, Rand DG, Gal YK: **Economic games on the internet: the effect of $1 stakes.** *PLOS ONE.* 2012; **7**(2).
PubMed Abstract | Publisher Full Text

Bayne T, McClelland T: **Ensemble representation and the contents of visual experience.** *Philosophical Studies.* 2018; **176**(3): 733–753.
Publisher Full Text

Biederman I: **On the semantics of a glance at a scene.** *Perceptual Organization.* 1981; **213**: 253.

Biederman I, Rabinowitz JC, Glass AL, *et al.*: **ON the information extracted from a glance at a scene.** 2006. 1–4.

Block N: **Overflow, access, and attention.** *Behavioral and Brain Sciences.* 2007; **30**(5-6): 530–548.
Publisher Full Text

Bronfman ZZ, Jacobson H, Usher M: **Impoverished or rich consciousness outside attentional focus: Recent data tip the balance for Overflow.** *Mind & Language.* 2019; **34**(4): 423–444.
Publisher Full Text

Campana F, Tallon-Baudry C: **Anchoring visual subjective experience in a neural model: The coarse vividness hypothesis.** *Neuropsychologia.* 2013; **51**(6): 1050–1060.
PubMed Abstract | Publisher Full Text

Chuyin Z, Koh ZH, Gallagher R, *et al.*: **Human Annotations with Confidence Ratings on Natural Images.** 2021, December 20.
Publisher Full Text

Chuyin Z, Koh ZH, Gallagher R, *et al.*: **Supplementary materials for the IA paper.** 2021, December 22.
Publisher Full Text

Chuyin Z, Koh ZH: **PantheraTigrisAltaica/IntersubjectiveAgreement: The analysis code (v1.0).** *Zenodo.* 2021.
Publisher Full Text

Cowen AS, Keltner D: **Self-report captures 27 distinct categories of emotion bridged by continuous gradients.** *Proceedings of the National Academy of Sciences.* 2017; **114**(38): E7900–E7909.
PubMed Abstract | Publisher Full Text

Crump MJC, McDonnell JV, Gureckis TM: **Evaluating amazon's mechanical turk as a tool for experimental behavioral research.** *PLOS ONE.* 2013; **8**(3): e57410.
PubMed Abstract | Publisher Full Text

De Clercq A, Crombez G, Buysse A, *et al.*: **A simple and sensitive method to measure timing accuracy.** *Behavior Research Methods, Instruments, & Computers.* 2003; **35**(1): 109–115.
PubMed Abstract | Publisher Full Text

Dennett DC: **Heterophenomenology reconsidered.** *Phenomenology and the Cognitive Sciences.* 2007; **6**(1): 247–270.
Publisher Full Text

Endress AD, Potter MC: **Large capacity temporary visual memory.** *Journal of Experimental Psychology: General.* 2014; **143**(2): 548–565.
PubMed Abstract | Publisher Full Text

Evans KK, Treisman A: **Perception of objects in natural scenes: Is it really attention free?.** *Journal of Experimental Psychology: Human Perception and Performance.* 2005; **31**(6): 1476–1492.
PubMed Abstract | Publisher Full Text

Fabre-Thorpe M, Delorme A, Marlot C, *et al.*: **A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes.** *J Cogn Neurosci.* 2001; **13**(2): 171–180.
PubMed Abstract | Publisher Full Text

Fei-Fei L, Iyer A, Koch C, *et al.*: **What do we perceive in a glance of a real-world scene?.** *Journal of Vision.* 2007; **7**(1): 10–29.
PubMed Abstract | Publisher Full Text

Fu Q, Liu YJ, Dienes Z, *et al.*: **The role of edge-based and surface-based information in natural scene categorization: evidence from behavior and event-related potentials.** *Consciousness and cognition.* 2016; **43**: 152–166.
PubMed Abstract | Publisher Full Text

Green DM, Swets JA: *Signal detection theory and psychophysics.* New York: Wiley; 1966; vol. **1**: 1969–1912.

Greene MR, Botros AP, Beck DM, *et al.*: **What you see is what you expect: Rapid scene understanding benefits from prior experience.** 2015; **77** (4): 1239–1251.
Publisher Full Text

Greene MR, Fei-Fei L: **Visual categorization is automatic and obligatory: Evidence from stroop-like paradigm.** *Journal of Vision.* 2014; **14**(1): 14–14.
PubMed Abstract | Publisher Full Text

Haun AM, Tononi G, Koch C, *et al.*: **Are we underestimating the richness of visual experience?.** *Neuroscience of Consciousness.* 2017; **2017**(1): 817–814.
PubMed Abstract | Publisher Full Text

Hollingworth A: **Does consistent scene context facilitate object perception?.** *Journal of Experimental Psychology: General.* 1998; **127**(4): 398–415.
PubMed Abstract | Publisher Full Text

Kaunitz LN, Kamienkowski JE, Olivetti E, *et al.*: **Intercepting the first pass: rapid categorization is suppressed for unseen stimuli.** *Frontiers in psychology.* 2011; **2**: 198.
PubMed Abstract | Publisher Full Text

Kaunitz LN, Rowe EG, Tsuchiya N: **Large capacity of conscious access for incidental memories in natural scenes.** *Psychological Science.* 2016; **27**(9): 1266–1277.
PubMed Abstract | Publisher Full Text

Kimchi R: **Primacy of wholistic processing and global/local paradigm: A critical review.** *Psychological bulletin.* 1992; **112**(1): 24–38.
PubMed Abstract | Publisher Full Text

Klein RA, Ratliff KA, Vianello M, *et al.*: **Investigating variation in replicability.** *Social psychology.* 2014; **45**(3): 142–152.
Publisher Full Text

Koch C, Tsuchiya N: **Attention and consciousness: two distinct brain processes.** *Trends in cognitive sciences.* 2007; **11**(1): 16–22.
PubMed Abstract | Publisher Full Text

Koffka K: **Perception: An introduction to the gestalt-theorie.** *Psychological Bulletin.* 1922; **19**(10): 531–585.
Publisher Full Text

Koh ZHmluu0010: **zhaokoh/gist-generation: First release (final version) with license (v1.0.1).** *Zenodo.* 2021.
Publisher Full Text

Larson AM, Freeman TE, Ringer RV, *et al.*: **The spatiotemporal dynamics of scene gist recognition.** *Journal of Experimental Psychology: Human Perception and Performance.* 2014; **40**(2): 471–487.
PubMed Abstract | Publisher Full Text

Li FF, VanRullen R, Koch C, *et al.*: **Rapid natural scene categorization in the near absence of attention.** *Proceedings of the National Academy of Sciences.* 2002; **99**(14): 9596–9601.
PubMed Abstract | Publisher Full Text

Lin Z, Murray SO: **Priming of awareness or how not to measure visual awareness.** *Journal of Vision.* 2014; **14**(1): 27–27.
Publisher Full Text

Litman L, Robinson J, Abberbock T: **TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences.** *Behavior research methods.* 2017; **49**(2): 433–442.
PubMed Abstract | Publisher Full Text

Loschky LC, Larson AM: **The natural/man-made distinction is made before basic-level distinctions in scene gist processing.** *Visual Cognition.* 2010; **18**(4): 513–536.
Publisher Full Text

Mack A, Clarke J, Erol M, *et al.*: **Scene incongruity and attention.** *Consciousness and cognition.* 2017; **48**: 87–103.
PubMed Abstract | Publisher Full Text

Mack A, Rock I: **Inattentional blindness: Perception without attention.** *Visual attention.* 1998; **8**: 55–76.

Macmillan NA, Creelman CD: *Detection theory: A user's guide.* Psychology Press; 2004.
Publisher Full Text

Malcolm G, Nuthmann A, Schyns P: **Beyond gist: Diagnostic information changes with level of scene categorization.** *Journal of Vision.* 2012; **12**.
Publisher Full Text

Matthews J, Schröder P, Kaunitz L, *et al.*: **Conscious access in the near absence of attention: critical extensions on the dual-task paradigm.** *Philosophical Transactions of the Royal Society B: Biological Sciences.* 2018; **373**(1755): 20170352.
PubMed Abstract | Publisher Full Text

McClelland T, Bayne T: **Concepts, contents, and consciousness.** *Neuroscience of Consciousness.* 2016; **2016**(1): niv012–niv019.
PubMed Abstract | Publisher Full Text

McLean D, Renoult L, Malcolm GL: **Expectation-Based Gist Facilitation: Rapid Scene Understanding and the Role of Top-Down Information.** *bioRxiv.* 2021.
Publisher Full Text

Mohammad S: **Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words.** *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 2018. (pp. 174–184).

Navon D: **Forest before trees: The precedence of global features in visual perception.** *Cognitive Psychology.* 1977; **9**(3): 353–383.
Publisher Full Text

Nishimoto S, Gallagher R, Koh ZH, Chuyin Z, Tsuchiya N: **Natural scene images for the IA paper.** 2021, December 21.
Publisher Full Text

Oliva A, Torralba A: **Building the gist of a scene: The role of global image features in recognition time series forecasting: The easy methods.** *Progress in Brain Research.* 2006; **155**: 23–36.
Publisher Full Text

Paolacci G, Chandler J, Ipeirotis PG: **Running experiments on amazon mechanical turk.** *Judgement and Decision making.* 2010; **5**: 411–419.

Pinto Y, van Gaal S, de Lange FP, *et al.*: **Expectations accelerate entry of visual stimuli into awareness.** *Journal of Vision.* 2015; **15**(8): 13–13.
PubMed Abstract | Publisher Full Text

Qianchen L, Gallagher RM, Tsuchiya N: **How much can we differentiate at a brief glance: Revealing the truer limit in conscious contents through the Massive Report Paradigm (MRP).** *Royal Society Open Science.* 2022; **9**(5): 210394.
Publisher Full Text

Rosch E: *Principles of categorization.* The MIT Press; 1998; 1–22.

Seth AK, Dienes Z, Cleeremans A, *et al.*: **Measuring consciousness: relating behavioural and neurophysiological approaches.** *Trends in Cognitive Sciences.* 2008; **12**(8): 314–321.
PubMed Abstract | Publisher Full Text

Sheehan KB: **Crowdsourcing research: Data collection with amazonMechanical turk.** *Communication Monographs.* 2017; **85**(1): 140–156.
Publisher Full Text

Software, M: **Inquisit 5.** 2018.

Sperling G: **The information available in brief visual presentations.** *Neuropsychology.* 1960; **74**(11): 1.

Sun Q, Ren Y, Zheng Y, *et al.*: **Superordinate level processing has priority over basic-level processing in scene gist recognition.** *i-Perception.* 2016; **7**(6): 204166951668130.
PubMed Abstract | Publisher Full Text

Sun Y, Zhang Z, Wu B: **The impact of contextual expectation on rapid natural scene recognition.** *Acta Psychologica Sinica.* 2017; **49**(5): 577–589.
Publisher Full Text

Thorpe S, Fize D, Marlot C: **Speed of processing in the human visual system.** *Nature.* 1996; **381**(6582): 520–522.
Publisher Full Text

Thunell E, Thorpe SJ: **Memory for repeated images in rapid-serial-visual-presentation streams of thousands of images.** *Psychological Science.* 2019; **30**(7): 989–1000.
PubMed Abstract | Publisher Full Text

Tsuchiya N, Koch C: **The relationship between consciousness and top-down attention**.*The neurology of consciousness.* Academic Press; 2016; (pp. 71–91).
Publisher Full Text

Walther DB, Caddigan E, Fei-Fei L, *et al.*: **Natural scene categories revealed in distributed patterns of activity in the human brain.** *Journal of Neuroscience.* 2009; **29**(34): 10573–10581.
PubMed Abstract | Publisher Full Text

Winter B: **A very basic tutorial for performing linear mixed effects analyses.** *arXiv preprint arXiv:1308.5499.* 2013.
Reference Source

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 21 September 2022

https://doi.org/10.5256/f1000research.137770.r148424

✓

Qiufang Fu
State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

The authors have successfully addressed all my concerns. I would like to thank the authors for their thorough responses.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Implicit learning, subliminal perception,

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 08 August 2022

https://doi.org/10.5256/f1000research.79226.r144465

?

Qiufang Fu
State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

This study combined a free-report and confidence rating to investigate how detailed information people can report after briefly seeing an image. Each image was presentation for 67, 133, or 267 ms, and followed by a dynamic masks for 300 ms. Six hundred and seventy participants attended this experiment online. A novel IA (Intersubjective Agreement) was developed to quantify how specifically the response words were used to describe the target image. The results showed that participants reported highly specific and detailed aspects of the briefly shown image, and IA is positively correlated with confidence.

I found that the novel paradigm is very interesting and the new findings are intriguing. However, as the method is novel, the authors might need to elaborate it more on the novel index and how the procedure can measure what they aimed to measure.

My concerns are mainly as follows:

1. For each image, participants were asked to enter five words and reported their confidence for each word. Why were participants asked to report five words rather than one or two or three or four words? Is it possible that this manipulation led participants to give more specific information about each image as the scene gist was unique while there were various kinds of specific information? I am wondering whether there were a sequence effect of IA on the five words. For example, is it possible that the IA for the first word were less specific than that for the other words?

2. The authors used a wide array of natural and artificial images but did not report IAs separately for natural images and artificial images. However, the authors mentioned “the current study aims to introduce a novel methodology to investigate how detailed information participants can report after briefly seeing a natural scene image” in the abstract. Thus, it would be better to analyze IAs for natural and artificial images separately.

3. Three SOAs were adopted in the current study. Previous studies indicated the accuracy increased with longer SOA in the forced-choice paradigms. I am wondering how the IA changed with SOA in the current study?

4. Given that when MTurk recruited more than 30 participants for a given block of 21 trials, the authors took the first 30 participants’ data for the subsequent analyses. There were 20 blocks, and thus the data from 600 rather than 670 participants were included. Is it right? Would you please make this more clear.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Implicit learning, subliminal perception,

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Aug 2022

Zhang Chuyin, Monash University, Melbourne, Australia

Thank you for reviewing our paper and providing valuable comments! We will respond to your concerns one by one.

1, *For each image, participants were asked to enter five words and reported their confidence for each word. Why were participants asked to report five words rather than one or two or three or four words?*

Reply: We agree that the number of response words is a key manipulation in our paradigm. To be honest, it is not easy to answer this "why" question. We imagine that the results might have been different if we asked participants to report 1, 2, 3, or 4 words. As a novel task, we had to start from somewhere.

In the pilot version of the experiment, we asked participants to report up to 10 words per image. This task turned out to be difficult or tedious for many participants. Meanwhile, most participants seemed comfortable in reporting roughly up to 5 words per image (including an option of reporting random words with confidence rating of "Don't know"). As we wanted to obtain as many words per image as possible, we settled on the five words. This information has been added as a footnote as follows:

- Footnote: "In a pilot task, we asked participants to report up to 10 words per image. This task turned out to be difficult or tedious for many participants. Most participants seemed to be comfortable in reporting roughly up to 5 words per image. As we wanted to obtain as many words per image as possible, we settled on the 5 words."

We believe that most aspects of the results would be replicated even if we reduce the number of words into 4, 3, 2, or 1. But, which aspects of the results stay the same or differ will be an empirical question. Having said that, repeating all experiments in this paper with the reduced report number of words will consume a lot of time and resources. Thus unless we have a specific question to address, we hope not to include this in this paper. We hope the reviewer would agree with this.

2, *Is it possible that this manipulation led participants to give more specific information about each image as the scene gist was unique while there were various kinds of specific information? I am wondering whether there were a sequence effect of IA on the five words. For example, is it possible that the IA for the first word were less specific than that for the other words?*

Reply: Based on what we expect from the literature, we agree with the reviewer that participants may first report the scene gist, then more specific information. To test this, we directly analysed the sequence effect on Word IA as you suggested (see Figure R1a). Although Word IA are almost saturated to the maximum value of 1, we can still see Word IA tends to be slightly higher for earlier-reported words at all three SOAs with SOA = 67ms having strongest effect of the sequence (This is confirmed statistically: a negative correlation between report order and Word IA, r = -0.056, p < .001. The interaction between SOAs and report sequence is significant, p < .001). The report sequence didn't influence confidence much even when we looked at separately for each SOA (Figure R1b).

Thus, if we assume that Word IA captures something about "specific information" (we believe it is, as it is only high when the word is reported to the target image, but not to the other images), the report sequence does not matter much (even at the 5th location with SOA = 67ms, median Word IA is > 0.96). Perhaps, to see the report sequence effects clearly, we may need to use threshold level natural scenes (e.g., with SOA = 17 or 33 ms) in the future study.

See Figure R1 linked here

Figure R1. Report sequence effect analysis. a) The boxplot showing Word IA, grouped by report sequence and SOA. b) Confidence, grouped by report sequence and SOA.

Because this finding is counter-intuitive and possibly going opposite of what has been established in the field using artificial stimuli, we checked if this result is true only for natural stimuli. Please be mindful that given we included only 8 artificial stimuli, the following analysis is limited.

In particular, we checked the two Sperling letter arrays (1960). With this letter array stimuli, many researchers would agree that individual letters, such as {U, P, L, G, J} in Figure 4a, are specific responses, whereas descriptions about more generic aspects, such as alphabet, letter, are more like scene gist. Thus, focusing on these images makes it easier to check the report-sequence vs specificity hypothesis. For both types of reports, we calculated the percentage of responses at each report sequence (see Figure R2).

See Figure R2 linked here

Figure R2. The percentage of responses given at each report sequence for generic and specific responses.

Under our experimental condition, it appears that participants tend to report specific letters first, then generic impressions later. It is interesting to note that this result was obtained despite the fact that we included the Sperling array in the midst of natural scenes, which would not let participants to expect specific letters. Having said that, due to the limited amount of the data, a chi-square test found no significant association between report sequence and the type of response, $X^2(4) = 6.03$, p = .1966 (N = 30 participants for each of two images). Thus, this observation is quite preliminary.

In summary, we infer that there is a very weak sequence effect that participants tend to report specific words first. This won't influence our main finding that participants are able to report very specific information at a brief glance of the images.

3, *The authors used a wide array of natural and artificial images but did not report IAs separately for natural images and artificial images. However, the authors mentioned "the current study aims to introduce a novel methodology to investigate how detailed information participants can report after briefly seeing a natural scene image" in the abstract. Thus, it would be better to analyze IAs for natural and artificial images separately.*

Reply: We agree that natural images and artificial images might show different results. However, because we had 412 natural scene images and only 8 artificial images (4 pairs), reported results reflect mostly for natural images. However, the overall trends of the results are quite similar as we summarised in Figure 7. In Figure 4a, we also analysed the Sperling's letter arrays in particular, because they have been widely studied by previous studies. Our Figure 4a shows that the letter array has the highest Image IA (mean Word IA under an image), and almost all the responses were very specific. Having said that, we checked if the results changed after removing artificial images. We confirmed that they did not change. We added this in results as follows:

- Footnote: "Because we had 412 natural scene images and only 8 artificial images (4 pairs), reported results about Word IA reflect mostly for natural images. We checked if any of the results changed significantly after removing artificial images, but they did not."

4, *Three SOAs were adopted in the current study. Previous studies indicated the accuracy increased with longer SOA in the forced-choice paradigms. I am wondering how the IA changed with SOA in the current study?*

Reply: We note that in Fu et al.'s (2016) categorization task, participants' accuracy increased from ~70% (SOA = 40ms) to ~90% (SOA = 213ms). Thus we checked to see if there is a significant relationship between SOA and Word IA, as a proxy of "accuracy" in our study (Note that Word IA should not be taken as "accuracy" in the usual sense, where experimenters define what the answer should be. Rather IA measures the agreement or similarity of experience among participants, which may or may not match with what experimenters' expected answers).

We added a Figure 6d into the paper (see Figure R3) to show this relationship more clearly. Interpretation of results for confidence = 1 (~5% of all responses) is somewhat complicated, because we explicitly asked participants to enter random words with confidence = 1 when they can't give five words. In fact, entries such as "none" and "NA'' are often entered with confidence = 1. For SOA = 267ms, the median Word IA is around 0.5 when confidence = 1. This is consistent with an idea that participants know that their report does not reflect the image. For SOA = 67ms and 133ms, however, the median Word IA are higher than 0.5 (0.65 and 0.82, respectively), possibly reflecting fleeting impressions that participants had, which were actually shared with other participants in specific and selective ways as captured by Word IA. When confidence is higher than 1, Word IAs were nearly saturated (median > 0.95) for all SOAs. (There is a significant positive correlation between SOAs and Word IA ($p < 0.001$, multilevel regression), when we exclude confidence = 1).

Perhaps, to see the SOA effect clearly, we may need to use threshold level natural scenes (e.g., with SOA=17 or 33 ms) in the future study.

See Figure R3 linked here

Figure R3 (Figure 6d). The boxplot showing Word IA, grouped by confidence and SOA.

5, *Given that when MTurk recruited more than 30 participants for a given block of 21 trials, the authors took the first 30 participants' data for the subsequent analyses. There were 20 blocks, and thus the data from 600 rather than 670 participants were included. Is it right? Would you please make this more clear.*

Reply: Thank you for catching this mistake. It's correct that we analysed only 600 participants' responses. We've clarified this in the revision.

Competing Interests: NA.

Reviewer Report 07 February 2022

https://doi.org/10.5256/f1000research.79226.r120530

✓ **Marius Usher**
School of Psychology, Tel Aviv University, Tel Aviv, Israel

The manuscript examines the amount of information that observers can extract from visual images at very fast rates (67/133/267 ms masked). The researchers present a large set of complex scenes, which observers can not generate expectations about, and rely on a verbal free response method (rather than the traditional forced response) to obtain information about what the observers could extract from the brief and masked display. In order to quantify the accuracy of the responses the authors developed a novel index called Intersubjective Agreement (IA), which reflects the specificity of the responses to a target image. The results indicate that most of the responses show a high IA and correspond to visual details (small object, or letter in a scene), a finding which goes against the traditional wisdom that people can only extract broad (gist or category type) information from such brief displays. Moreover, the IA is correlated with confidence, indicating metacognitive conscious access.

I find the question that this study addresses an important one and the method innovative and solid. The writing is also quite clear, so I believe this paper could be of interest to the research community. Below I will enter a number of comments, which the authors could address to make their paper more clear and easier to follow, and to explore its implications.

Methods

- p. 4. I wonder if the term Word IA, should be Word-Image IA (as it seems to link to a Image-word pair).
- p. 5. "Word-IA are high". Is the point that the values are closed to 1? Is the possible range (0, 1)? It may help to specify this.
- Fig 2c. There should be a legend for the colors here.
- Also it would help to clarify the motivation why one has to rely on such a complex method (ROC curves) for the index, instead of just computing a ration of responses given to the target image over All responses (to all images). Usually ROC curves are used when there is a subjective-criterion difference. It was not very clear why this complex method was needed here. Finally, the construction of Fig 2c may be explained better. (Those cumulative probabilities appears to only have binary 0/1 values...)
- It would help if one could also provide some index of the information that one can extract from the images. (Perhaps this could integrate the IA with the number of words reported...)

Results

There are very interesting results reported here. But to me the most important would be to see how the information extracted from the image (see above) or at least the IA, varies with the presentation rate. While the paper reports that IA are high even at the fastest rate (67 ms), I did not see a Figure showing the dependency on the rate. Finally, it would be quite interesting to know what the temporal limit at which the capacity to extract information from rapid displays breaks down? (I suppose this will need to take place at fast enough rates, of < 20 ms or so). Showing such data will also help to better appreciate the results at 67 ms, as the more rapid presentation would serve as a control, relative to which the present results can be compared.

Implications

Finally, it may be interesting to discuss implications to theoretical debates such as the Overflow in Consciousness research.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cognitive psychology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 21 Aug 2022

Zhang Chuyin, Monash University, Melbourne, Australia

Thank you for reviewing our paper and providing valuable comments! We will respond your concerns one by one.

1, *p.v4. I wonder if the term Word IA, should be Word-Image IA (as it seems to link to a Image-word pair).*

Reply: We agree with you that Word-Image IA is more precise and consistent with what we meant by Word IA. However, Word-Image IA is a bit long and the abbreviation of WIIA is not intuitive to grasp. On the other hand, Word IA is concise with a minimal use of abbreviations, IA. Thus, we opted for using Word IA throughout. We've added the following to make it clear:

- Footnote: "To be precise, we should perhaps use "Word-Image IA" instead of "Word IA" as it links to an image-word pair. Just for the simplicity and readability, we use "Word IA"."

2, *p. 5. "Word-IA are high". Is the point that the values are closed to 1? Is the possible range (0, 1)? It may help to specify this.*

Reply: Yes, the possible range of Word IA is [0, 1]. By "high" we mean the values are close to 1. We've clarified this as follows:

- Page 5: "as we mentioned above, Word IA ranges from 0 to 1. Most words that were reported had very high Word IA (close to 1) across SOAs."

3, *Fig 2c. There should be a legend for the colors here.*

Reply: Thank you. We've added in a figure legend for Figure 2c (and newly added e). We also paste the modified explanation of Figure 2 as follows:

- ***See Figure R1 linked here***
- **Figure R1 (Figure 2).** e) Another example for the target word "grass" after viewing (a), whose Word IA = 0.84. They are constructed in the same way as (c) and (d). This example shows values on the red line are not binary.

4, *Also it would help to clarify the motivation why one has to rely on such a complex method (ROC curves) for the index, instead of just computing a ration of responses given to the target image over All responses (to all images). Usually ROC curves are used when there is a subjective-criterion difference. It was not very clear why this complex method was needed here.*

Reply: We had empirical and theoretical motivations to introduce the ROC curves instead of a simpler method, like the one you suggested.

Empirically, a simpler ratio scale is not easy to interpret. If we define it as: Ratio1 = # of responses given to the target / # of responses to all images (including the target), then, it ranges from 0 to 1. But the distribution is extremely concentrated around 0 and 1 (Figure R2).

Theoretically, this method is problematic in that any word that was reported in many images, such as "grass", can never obtain high value, even if one particular image was reported with "grass" for all of N = 30 people. If "grass" was reported 300 times in total, then this ratio1 index will be 0.1 (= 30/300). Thus, we need some way to normalise the baseline report rate of the word.

See Figure R2 linked here
Figure R2. Histogram for the distribution of ratio1 (X axis). Ratio1 = [# of responses given to the target / # of responses to all images (including the target)]. We calculated ratio1 for all unique (word, image) pairs.

This leads to the 2nd type of ratio scheme, Ratio2 = % of responses given to the target / % of responses to all images (including the target).

This comes closer to what our ROC is measuring. However, Ratio2 = $[X_{target}/30]/[X_{all}/(420 \times 30)] = (X_{target}/X_{all}) \times 420$ = Ratio1 × 420. Thus, it can go from 0 to 420 and interpretation of ratio 2 is even harder than ratio 1. We need some alternative methods.

At first, we wanted to invent an index that is free of subjective criteria, or confidence rating, as you suggested, using ROC. Let us consider "eiffel-tower" in Fig 2a as an example. 13 out of 30 participants reported it under the target image, and 1 participant reported it under other images. Then, our implementation of the ROC goes as follows:

1. Order the responses according to the confidence ratings to the target image, then to compute a cumulative probability. This serves like a cumulative hit in a standard ROC. (Figure R3a - green line)
2. Repeat the same procedure as 1) but for all-non target images, which serves like a

cumulative false alarm. (Figure R3a - red line)

3. Construct the ROC. (Figure R3b)

See Figure F3 linked here

Figure R3. a) Cumulative proportion (y-axis) for the target (green) and other images (red). We first order the responses according to the confidence, and count the number of responses for each confidence rating. X axis is confidence rating. If participants did not report “eiffel-tower”, we counted them as 0 confidence rating response, so that cumulative proportion always reaches 1 for both hit (green) and false alarm (red). b) A ROC curve constructed from Figure R3a. X axis is the false alarm rate, which comes from the red line in Figure R3a. Y axis is the hit rate, which comes from the green line in Figure R3a. AUC here is 0.72.

From the ROC, we computed the area under the ROC, which we call “AUC by confidence”. For the example of “eiffel-tower”, AUC by confidence = 0.72.

We calculated AUC by confidence for every response word, compared against Word IA in Figure R4a.

See Figure R4 linked here

Figure R4. a) The scatterplot showing the relationship between Word IA (x axis) and AUC by confidence (y axis). Red line is the identity. b) The strong correlation between AUC by confidence and report frequency under the target image. c) Histogram of AUC by confidence (x-axis is the same as panel b).

In Figure R3, we showed the cumulative proportions and their corresponding ROC. As we count the proportion of people who did NOT report “eiffel-tower” as confidence = 0, the maximum area under the ROC (regardless of the false alarm) is set by the % of people who reported “eiffel-tower” in the target image, which is 13/30 = 43%. This corresponds to the highest black point in Figure R3b on the y-axis. Although “eiffel-tower” was reported only 1 out of 419 images × 30 responses (nearly 0 false alarm), this means that “eiffel-tower” can attain only up to AUC of 0.72. This means, generally speaking, AUC by confidence can reach 1 when and only when all 30 participants reported the word under the target image.

Figure R4a shows a moderate positive correlation between AUC by confidence and Word IA. However, Figure R4b shows an extremely tight correlation between AUC by confidence and the report frequency of the word for the target image. In the case of the “eiffel-tower”, the report frequency of 13 (out of 30) indeed predicts AUC by confidence at 0.72. As a result, their distributions are quite different. AUC by confidence is highly concentrated around 0.55 (Figure R4c).

Another feature we see in Figure R4a is that AUC by confidence hardly goes below 0.45. This is because our participants rarely reported the same word across many images. In other words, the baseline report rate of any given word is very low. Among all reported words, “man” was reported most frequently, but it was only around 12% of possible report rate (around 1600 times, out of the possible 420 × 30=12600 times in total under all images).

This leads to our proposal of IA. Figure 2c in the main text shows that we use % of participants for cumulative probability to construct ROC. This allows us to characterise a word such as “eiffel-tower” with IA~1.0, which signals that this word is effectively used for this image, but not for the other images. This also explains why in Figure R4a, we see many words with AUC by confidence of ~0.6-0.7 having Word IA of 0.75-1.0.

Given the importance of explaining the motivations behind our new IA measure, we decided to include this as a supplementary material. We explicitly mentioned this in the Results section as follows:

- ○ Footnote: "We described our motivations and rationale behind our definition of IA over other simpler methods, such as ratio of reported words for the target vs all images in Supplementary Material (see https://osf.io/nvfhs)."

5, *Finally, the construction of Fig 2c may be explained better. (Those cumulative probabilities appears to only have binary 0/1 values…)*

Reply: You are correct in pointing out that both green and red lines in Figure 2c go from 0 to 100%. As long as we do not group multiple images as a target image, the green line always goes from 0 to 100% as it shows the cumulative probability of hit. The red line (other-image) might be different for other examples. We’ve added Figure 2e and 2f (example of “grass”) to show that cumulative possibilities on the red line take values between 0 and 1. This is because the word “grass” was reported in many images with different frequencies (50%, 30%, …). We’ve also added a sentence into the legend of Figure 2c as follows:

- ○ (Figure 2c) “The value on this curve doesn’t have to be 0 or 1, it jumps from 0 to 1 because no one reported “eiffel-tower” under the other images. ”

6, *It would help if one could also provide some index of the information that one can extract from the images. (Perhaps this could integrate the IA with the number of words reported…)*

Reply: We agree with you that it would be great to come up with some index of information that one can extract from a given image. In our results section, we proposed “Image IA” as such an estimate. Image IA is the mean of Word IA under a specific image.

Thanks for your suggestion to integrate IA and the number of words reported, which we did not try before. While there can be many ways to integrate them, as an initial attempt, we tested if the sum of the Word IA (called “Sum IA”) per a particular image can work as the desired index. Below, we compare our original Image IA and Sum IA.

Ideally, the index should be lower for a simple and abstract image, which would be hard to extract specific information and to report many words that are specific for that image. Also, the ideal index should be high if an image contains many distinct, clear and nameable objects.

Somewhat surprisingly, we found that Image IA and Sum IA are uncorrelated (see Figure R5). But the relationship between them highly depends on the number of unique words reported under an image. In Figure R5, we display 4 examples at the extreme corners (highest/lowest Image IA/Sum IA).

By looking at the top left image, we have an impression that high Sum IA (= 26.82) may be related to many words that can describe the image (e.g., rice, wheat, field, farm, yellow, brown, machine etc). However, lower Image IA (= 0.84) may be related to the fact that each word is not very specific to this image and is often reported in other images.

Now, looking at the bottom right image, it may have a high Image IA (= 0.99) as it is accompanied with very specific words that are not reported in other images (e.g., boat, lake, fast, speedboat, life-vest. See Figure 4a). But as this image does not contain many objects, it may have led to smaller Sum IA (=12.82).

But it is hard to say why the top right dog image was described with many (Sum IA = 28.6) specific words (Image IA = 0.95) while the wheel image at the bottom left was described with few (Sum IA = 11.7) and less specific words (Image IA = 0.84).

Of course, we need to be careful in our inference because our impression here is based on the prolonged and unlimited duration of image viewing. Our IAs are based on the reports by participants who viewed them only briefly (67ms, 133ms or 266ms, masked).

See Figure R5 linked here
Figure R5. The relationship between Image IA and Sum IA. There is no correlation between Image IA and Sum IA, but the relationship highly depends on the number of unique words reported under an image. Each colour (dots and lines) means a number of unique words reported under an image. Four example images with their Sum IA (below the image) and Image IA (in the bracket) are also shown.

We also examined other images. Sup Fig 6 and 7 ordered 420 images by their Image IA and Sum IA, respectively (https://osf.io/m7az6). For the convenience of the reviewer, in Figure R6, we show the 42 images with the highest Sum IA, the 42 images with the lowest Sum IA, and their Image IA. While our impressions above may be largely true, there are also notable exceptions. At this stage, we cannot say which of Image IA and Sum IA captures "the information one can extract" better. Each index is likely to be capturing different aspects of how people experience the scenes at a brief moment and report them in an intersubjectively consistent manner. To say more, we would need more research in the future, which we plan to do.

See Figure R6 linked here
Figure R6. a) 42 images with the highest Sum IA, ranked by Sum IA. The Sum IA and Image IA (in the bracket) of each image is shown under the image. b) 42 images with the lowest Sum IA, ranked by Sum IA. The Sum IA and Image IA (in the bracket) of each image is shown under the image.

7, *There are very interesting results reported here. But to me the most important would be to see how the information extracted from the image (see above) or at least the IA, varies with the presentation rate. While the paper reports that IA are high even at the fastest rate (67 ms), I did not see a Figure showing the dependency on the rate.*

Reply: In order to increase the number of observations for each image, we calculated our Image IA (and Sum IA above) across SOA. It means we used N = 30 participants across three SOAs under each image. So the Image IA and Sum IA values above are the same for different SOAs.

To address how these values depend on SOAs, we sacrificed the precision and calculated them for each SOA (where we had only N = 10 participants, Figure R7). From the figures we can see that both Sum IA and Image IA at SOA = 67ms are slightly lower than them at SOA = 133 and 267ms (ANOVA: for Image IA, F(2, 809.03) = 71.27, p < .001; for Sum IA, F(2, 836.19) = 41.67, p < .001).

See Figure R7 linked here
Figure R7. a) The boxplot showing Sum IA per SOA at each SOA. b) The boxplot showing Image IA per SOA at each SOA.

8, *Finally, it would be quite interesting to know what the temporal limit at which the capacity to extract information from rapid displays breaks down?*

Reply: We agree with you that going shorter SOAs will be very interesting.

We performed some preliminary experiments with shorter frame rates online. We confirmed its feasibility, but to be confident about the results, we believe that we need to intermix shorter SOAs with longer SOAs and test many participants with many images. Showing images with only shorter SOAs will be feasible to add to our data, but we expect that many participants will start setting a high threshold for reporting when there are no easy trials included (e.g., Lin & Murray 2014 Journal of Vision). Thus, to avoid such an effect, we probably need to replace our 67 ms condition with a shorter condition, while keeping 133 and 266ms conditions. To test the entire image set, this will mean that we need to run another 670 subjects to add the 33 ms (or 17ms) condition. Perhaps, we need to start its feasibility from the in-lab face to face experiment. Overall, we agree that this will be an important direction of research we plan to pursue, but given the time it would take to complete, the length and the density of the current paper, we prefer to do so in the future.

Reflecting on this, we decided to add the following under “Presentation duration” in Discussion:

- “We performed some preliminary experiments with shorter SOAs online and confirmed its feasibility. To conduct such a study in the future, however, we recommend to confirm the results with in-lab participants to some extent. We also suggest that shorter SOA trials should be intermixed with longer SOA trials, so that participants can set an appropriate threshold for reports (Lin & Murray 2014 Journal of Vision). Such experiments will address an important question on the temporal limit of intersubjectively shared experience upon viewing rapid visual scenes.”

9, *Finally, it may be interesting to discuss implications to theoretical debates such as the Overflow in Consciousness research.*

Reply: We agree that our paper has some implications to the theoretical debate on the overflow in consciousness research. Our paper contributes to this debate in two ways. First, is about the Sperling paradigm, which has been the central to the overflow debate.

In Sperling's paradigm (1960), participants are briefly shown an array of 12 letters in 3 rows and 4 columns. Generally, they are believed to report having an impression of seeing all letters. However, under the whole-report condition, where they are asked to report as many letters as they can, they cannot report no more than 4 individual letters accurately. However, when cued on a particular row, after the array disappears, they can report all letters accurately.

Our results in Figure 4a cast doubt on this uncontested description of the results in the whole-report condition (See also Figure R6b, Supplementary Material 6 and 7 for two Sperling arrays, with both very high Image IA and Sum IA). Our participants in fact reported highly specific letters ({U, P, L, G, J} in Figure 4a) more frequently than the general impression of "letter array", "alphabets".

This is consistent with some reviews (e.g., Haun et al., 2017; Bronfman et al. 2019), which propose that participants are actually perceiving more than previously thought.

Second, our experiment opens up a possibility to extend the discussion of the overflow debate in more naturalistic situations, while most discussions so far centred around the results obtained with highly artificial stimuli. Nonetheless, our current study has implemented more like the whole-report condition of the Sperling paradigm. By extending it with a pre/post cued partial-report paradigm, we can expect that our paradigm provides a novel methodology to study the overflow debate with a large set of natural scene images. We included these points under "Future directions" in Discussion as follows:

Page 14: "Finally, our paper has some implications to the theoretical debate on whether conscious experience is restricted by cognitive access or overflows such restriction (Block, 2007 BBS; Bronfman et al., 2019 Mind & Language). Previous debate on the overflow issue was almost always focused on the artificial stimulus situation introduced by Sperling (1960). Our paper contributes to this debate in two fresh ways. First, since the Sperling's initial introduction of the paradigm, it has been almost always "assumed" that what participants see in the "whole-report" condition is just a gist, such as "seeing everything", "alphabets", and "letters" (Haun et al., 2017). Our empirical free-report challenges this notion. Even without any warning or expectation, participants report specific letters rather than generic gists (e.g., Figure 4a). Second, our paradigm opens up a possibility to examine the overflow issue with naturalistic stimuli. For that, we would need to develop equivalent experimental contrasts between the whole- vs. partial reports as in Sperling 1960, which is not impossible."

Competing Interests: NA.