Analysis of Excluded Question Data

During analysis of the collected participant data, we excluded questions collected from the first article because the first article was used as a practice round. Some of the reasons for excluding the first article were that participants would not know what to pay attention to during their first article and following questions. This could cause them to perform poorly during the first article questions compared to the second and third article. This difference in expectation could cause unreliable results within the questions asked after the first article. Below we provide equivalent analysis that was done within the paper, but this time including data from first article questions.

Table 1: General Question Model

Variables	*Estimate*	*Std. Error*	*t value*	*p*
(Intercept)	0.3238	0.1417	2.285	0.023
Question Time	0.0000	0.0004	0.158	0.874
Article Time	0.0006	0.0001	4.952	**<0.001**
Article Knowledge	-0.0083	0.0154	-0.536	0.591
Article Word Count	0.0000	0.0001	0.195	0.845
English Proficiency	0.0516	0.0162	3.173	**0.001**
Education Level	0.0290	0.0109	2.654	**0.008**
Age	0.0006	0.0004	1.349	0.177
Platform [Prolific]	-0.0408	0.0219	-1.864	0.062
Image Presence [True]	-0.0866	0.0193	-0.447	0.654
Reading Level	-0.0036	0.0094	-0.229	0.703

Table 2: Visual Question Model

Variables	*Estimate*	*Std. Error*	*t value*	*p*
(Intercept)	0.2076	0.1300	1.597	0.011
Question Time	-0.0015	0.0007	-2.109	**0.035**
Article Time	0.0005	0.0001	4.290	**<0.001**
Article Knowledge	0.0085	0.0155	0.547	0.584
Article Word Count	0.0000	0.0009	-0.577	0.565

English Proficiency	0.0500	0.0168	2.966	**0.003**
Education Level	0.0185	0.0112	1.645	0.100
Age	0.0002	0.0004	0.492	0.623
Platform [Prolific]	-0.0667	0.0227	-2.938	**0.003**
Image Presence [True]	-0.0035	0.0198	-1.181	0.070
Reading Level	0.0012	0.0079	1.530	0.129
Visual Question Consistency	0.0635	0.0237	2.677	**0.007**

Table 3: Image Question Model

Variables	*Estimate*	*Std. Error*	*t value*	*p*
(Intercept)	0.3816	0.1498	2.547	0.011
Question Time	-0.0035	0.0007	-4.741	**<0.001**
Article Time	0.0004	0.0001	3.422	**<0.001**
Article Knowledge	0.0155	0.0148	1.048	0.294
Article Word Count	-0.0001	0.0001	-0.801	0.425
English Proficiency	0.0310	0.0147	2.110	**0.035**
Education Level	0.0101	0.0100	1.015	0.310
Age	0.0001	0.0004	0.301	0.763
Platform [Prolific]	-0.0486	0.0202	-2.401	**0.016**
Image Presence [True]	0.1405	0.0187	7.494	**<0.001**
Reading Level	0.0063	0.0105	0.598	0.551

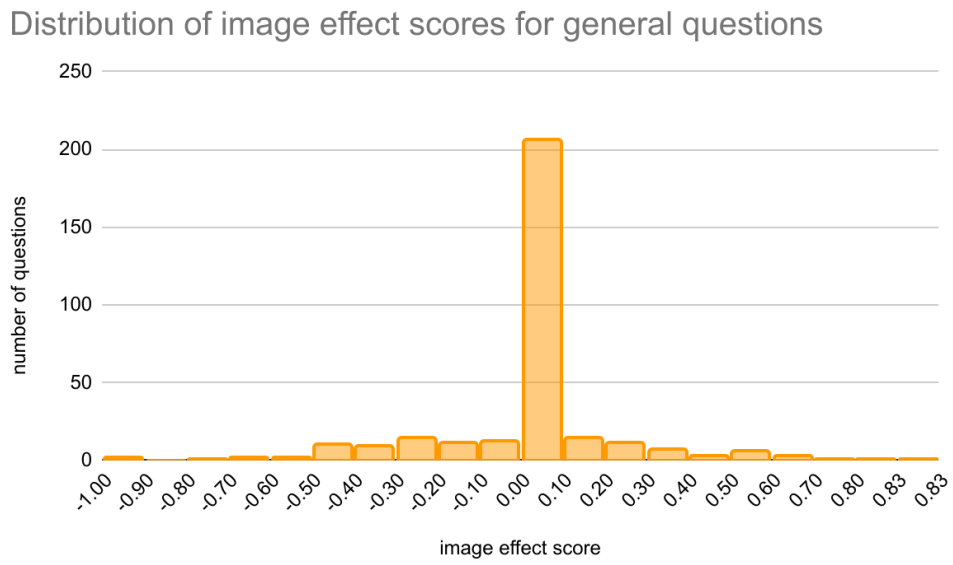
With the included data taken from first article questions all of the variables mentioned within the original paper come up significant here as well. Article time, education level, and english proficiency are significant in general questions as seen in **Table 1**. Question time, article time, english proficiency, and visual question consistency are significant in visual questions as seen in **Table 2**. Question time, article time, and image presence are similarly significant in image questions as seen in **Table 3**. All of these variables have the same affects as mentioned in the original paper.

Including first article questions does bring up the significance of platform in both visual **(Table 2)** and image questions **(Table3)**, and the significance of English proficiency for image questions

alone **(Table3)**. In this analysis the platform variable indicates that prolific participants had worse knowledge acquisition for both visual and image questions. Further, with higher english proficiency participants did better on image questions based on this larger dataset.

More on image qualitative analysis from section 5.3

One author performed the manual inspection for section 5.3. They inspected the top 15 image/article/question combinations for general questions, and the bottom 15 image/article/question combinations for image questions. A full qualitative analysis, with an explicit coding system, was out of the scope of this work. While the limited number of responses per question might affect the generalizability of these results, the range and distribution of image effect values reflect the main findings in our paper: for the large majority of questions, the image effect is 0, and only a small number of questions have a non-zero image effect, equally distributed across the positive and negative spectrum. Conversely, we see a negative image effect for a small proportion of image questions only, with the accuracy of the large majority of those questions being positively affected by the presence of an image (see below plots). This explanation was added to the results section.



For each general/visual/image question, we have one score that is either 0 if the question was answered correctly, or 1 otherwise. This makes scores quite comparable across questions of different types.

All image recognition questions are by definition image-consistent, as they are designed to test participants' ability to identify the image that best depicts an article topic. Therefore, our method would not apply to this case.

To evaluate image quality, we looked at whether the images fit the image quality criteria from the manual of style on Wikipedia. These include basic photographic quality rules, such as exposure and clutter, as well as more generic rules to optimally represent the article's subject. E.g., "An image of a white-tailed eagle is useless if the bird appears as a speck in the sky.". While manually labeling the quality of all images was out of the scope of this paper, we ranked all questions by "image effect" score, looked at the bottom 15, and evaluated their quality as well as other aspects (image type, question text, etc).

During which period was Newgrange built?

When was the Worldwide Plaza completed?

When was "La Scapigliata" painted?

(a) General knowledge questions where images help with content retention.

What is the material of Donatello's David?

How tall can totem poles be?

(b) Visual knowledge questions where images help content retention.

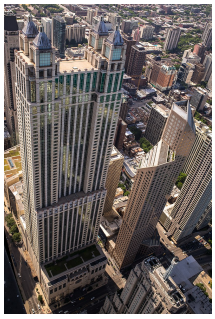
Figure 2: Qualitative analysis–examples of images that are helpful to answer generic questions. (a) Visual arts characteristics for general knowledge questions. (b) Visually consistent images for visual knowledge questions.

Low quality (*Spectral Bat*)

High quality (*Leafy Seadragon*)

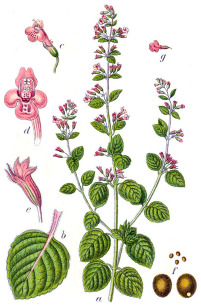
(a) Image quality

Part (*900 North Michigan*)



Whole (*900 North Michigan*)

(b) Part VS Whole



Illustrated version (*Nepeta*)

Photographed version (*Nepeta*)

(c) Photo VS Illustration

Figure 3: In-depth analysis - examples of images that are harmful to answer the image question. (a) Low Quality images. (b) Images that describe only a part of the object. (c) Photographed objects in the article vs illustrations in the question.