Eye movements predict test scores in online video education

Jens Madsen[1], Sara U. Julio[1], Pawel J. Gucik[1], Richard Steinberg[2], and Lucas C. Parra[1]

[1] *Department of Biomedical Engineering, City University, New York, NY*
[2] *School of Education and Department of Physics, City University, New York, NY*

Abstract

Experienced teachers pay close attention to their students, adjusting their teaching when students seem lost. This dynamic interaction is missing in online education. We propose to measure attention to online videos remotely by tracking eye movements, as we hypothesize that attentive students follow videos similarly with their eyes. Here we show that inter-subject correlation of eye movements during instructional video presentation is substantially higher for attentive students, and that eye movements are predictive of individual test scores on the material presented in the video. These findings replicate for videos in a variety of production styles, for intentional and incidental learning and for recall and comprehension questions alike. We reproduce the result using standard web cameras in a classroom setting, and with over 1,000 participants at-home without the need to transmit user data. Our results suggest that online education could be made adaptive to a student's level of attention in real-time.

Introduction

We have known for a long time that attended stimuli are easier to remember [1]. The point of gaze is an overt indicator of where we focus our attention [2, 3]. Therefore, the point of gaze is indicative of what we might recall in the future [4]. When students are not following the relevant teaching material, then there is a good chance that they are not paying attention and that they will perform poorly in subsequent exams. Experienced teachers know this and adjust the interaction with students accordingly [5]. During online education this immediate feedback is lost. Here we suggest that standard web cameras could be used to monitor attention based on eye movements. When they lose focus they can then choose to redo portions of a course, take a break, or start over. Eye tracking has been extensively used to evaluate online media, including user interfaces, advertising or educational material [6, 7]. These studies often focus on the content of eye fixations in static media, to determine, for example, whether users look at a specific graphic or whether they read a relevant text [8, 9]. This approach requires detailed analysis and interpretation of the specific online content, and cannot be used routinely to evaluate individual students. Evaluating the content of eye fixations is particularly complicated for dynamic stimuli such as instructional video, which is increasingly abundant online. Here we focus on dynamic video and whether students "follow" that dynamic content, in the literal sense of following with their eyes.

Previous studies have shown that during video presentation, brain signals of viewers respond similarly [10, 11, 12], in particular for dynamic videos that tell a story [13]. This similarity of responses, measured as intersubject correlation of the time courses of brain activity, is predictive of whether subjects subsequently remember the content of a story [14, 15], or how they perform in subsequent test [16]. In particular, it measures whether the video engages a viewer's attention [17]. We also know that eye movements are correlated across subjects during video presentation [18, 19], and that dynamic, well-produced movies and video advertising elicits higher intersubject correlation of eye-movements [20, 21, 22]. What has not been established yet is whether this eye-movement correlation similarly depends on attention, or whether it is predictive of learning. Much of our eye movements during video seem to be explained by simple low-level rules [20] and do not differ much even when movies are presented backwards in time [18]. We hypothesize, however, that online instructional videos guide eye movements in a similar way across students, but only if students are paying attention. We predict that eye-movement correlation between subjects is predictive of retention of the material presented in the video. The alternative hypothesis is that the stimuli drive eye movements without engaging a student's mind meaningfully in the material. One may also argue that static stimuli, while not reliably guiding eye movements, may nonetheless engage students minds [23, 24].

Here we test our hypothesis in different learning scenarios with instructional videos produced in different styles. Specifically, we measure intersubject correlation (ISC) of eye movements combined with pupil size, recorded while students watch short instructional videos. We test whether ISC is modulated by attention by measuring ISC during normal vs. distracted viewing. We investigate whether ISC of individual students is predictive of their performance in subsequent tests in a laboratory setting. We demonstrate how this approach can be used to measure attention remotely in online education, using subjects own computers, without the need to transfer data from the user, thus preserving online privacy. We test this in a classroom and on a large cohort of subjects at home.

Results

Effects of attention on eye movements during video presentation

We recruited 60 subjects to participate in a series of experiments where they were asked to watch 5 or 6 short instruc-

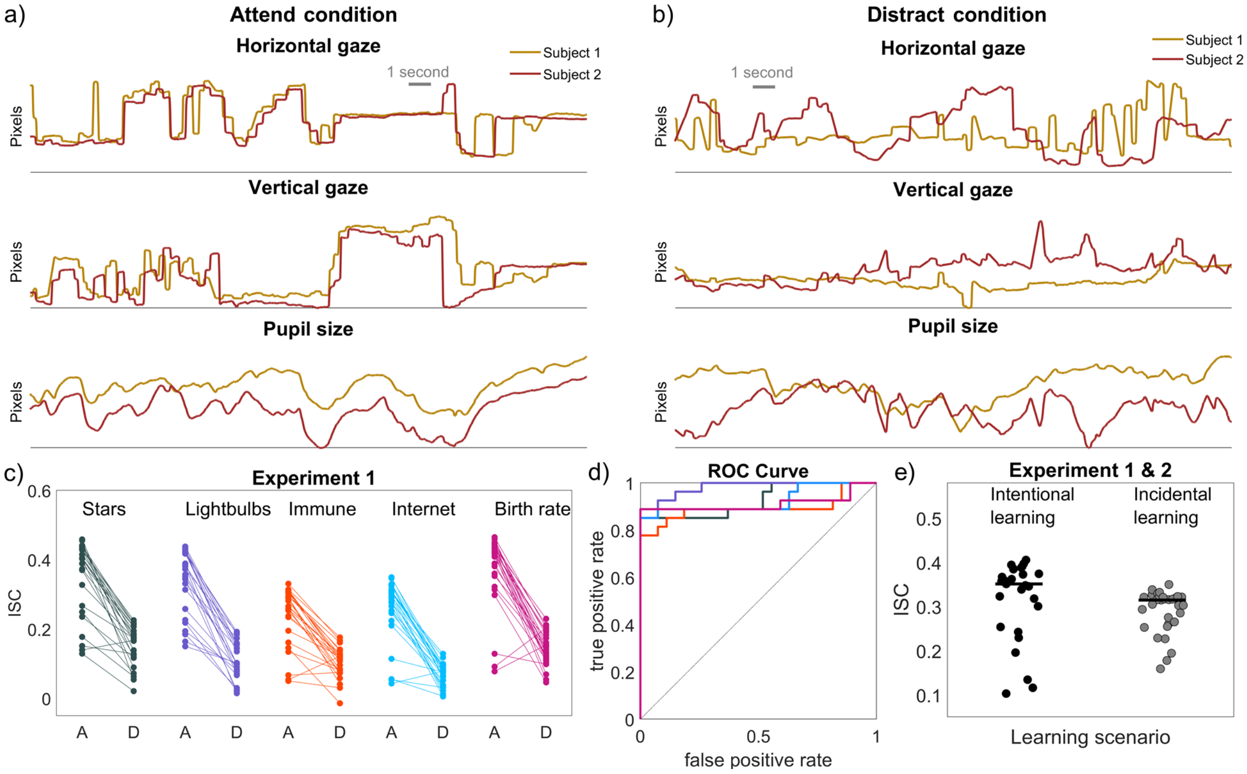

Figure 1: **Intersubject correlation of eye movements modulated by attention when watching instructional videos. a)** Two subjects' gaze position and pupil size follow each other during attentive viewing. **b)** The same two subjects viewing the same segment of video while distracted by a counting task. **c)** The intersubject correlation (ISC) of eye movement is measured as the mean of ISC of vertical and horizontal gaze position and pupil size. Values for each subject are shown as dots for all videos in Experiment 1. Each dot is connected with a line between two different conditions namely when subjects were either attending (A) or were distracted (D) while watching the video. **d)** the receiver operator curve for deciding whether a subject is attending or distracted based on their ISC. **e)** Intentional learning shows a higher ISC. Each dot is the average ISC for each subject when they watched all instructional videos in the attend condition using either the intentional or incidental learning style.

tional videos in the laboratory while we monitored their eye movements. The videos covered a variety of topics related to physics, biology and computer science (Tab. 1). Some feature a teacher writing on a board, while others use more modern storytelling using animations or the popular writing-hand style. A first cohort of subjects (N=27, 17 females, age 18-53 mean=26.74, standard deviation SD=8.98) watched 5 short instructional videos, after each video they took a test with questions related to the material presented in the videos, which they were informed were going to come. After watching the videos and answering questions they watched the videos again. To test for attentional modulation of ISC, in the second viewing subjects performed a serial subtraction task (count in their mind backwards in steps of seven starting from a random prime number between 800 and 1000). This is a common distraction task in visual attention experiments [4]. During the first attentive viewing eye movement of most subjects are well correlated (Fig. 1a), during the second, distracted viewing they often diverge (Fig. 1b). The same appears to be true for the fluctuations of pupil size. To quantify this, we measure the Pearson's correlation of these time courses between subjects. For each student we obtain an intersubject correlation (ISC) value as the average correlation of that subject with all other subjects in the group. We further average over the three measures taken, namely, vertical and horizontal gaze position as well as pupil size. This ISC is substantial during the normal viewing condition (Fig. 1c; ISC median=0.32, interquartile range IQR=0.12, across videos) and decreases in the second distracted viewing (ISC median=0.11, IQR=0.07). Specifically, a three-way repeated measures ANOVA shows a very strong fixed effect of the attention condition ($F(1,231)$=749.06, p=1.93$\cdot10^{-74}$) a fixed effect of video ($F(4,231)$=32.29, p=2.23$\cdot10^{-21}$) and a random effect of subject ($F(26,231)$=9.21, p=1.62$\cdot10^{-23}$). This confirms the evident variability across films and subjects. The effect of attention, however, is so strong that despite the variability between subjects one can still determine the attention condition near perfectly from the ISC of individual subjects (Fig. 1b). Specifically, a receiver operator characteristic curve for determining attentional state has an area under the curve of $Az = 0.944 \pm 0.033$ (mean ± SD over videos).

Motivation modulates intersubject correlation of eye movements

To test the effect of motivation we repeated the experiment, but this time subjects did not know that they would be quizzed on the content of the videos. The two conditions thus constitute intentional and incidental learning which are known to elicit different levels of motivation [25]. As expected, we find a higher ISC in the intentional learning condition (ISC median=0.325, IQR=0.12, N=27) as compared to the incidental learning condition (ISC median=0.317, IQR=0.06, N=30) (Fig. 1e; two-tailed Wilcoxon rank sum test: z=2.67, p=7.68$\cdot10^{-3}$). This suggests that lower motivation in the incidental learning condition resulted in lower attentional levels and thus somewhat less correlated eye movements and pupil size. The increased motivation in the intentional learning condition is also reflected in the increased test scores as compared to the incidental learning condition (Fig. 2c; intentional learning score=65.22 ± 18.75 points, N=27, incidental learning score = 54.53 ± 15.31 points, N=31; two-sample t-test: t(56)=2.39, p=0.02, d=0.63).

Correlated eye movements as predictors of test scores

In the previous experiments we confirmed the hypothesis that if subjects are distracted the ISC of eye movements and pupil size is reduced. Given the well-established link between attention and memory we therefore expect that ISC will be predictive of how much each subject retained from the instructional video. We tested this hypothesis by quizzing subjects after they had watched the video using a short four alternative forced-choice questionnaire (11–12 questions). Students that watched the video performed significantly better than naïve students (65.2% ± 18.8% versus naïve: 45%±8.8%; t(56)=-5.37 p=1.58$\cdot10^{-6}$; see Methods section for details). Importantly we find a strong correlation between ISC and test scores across subjects for all videos we tested (Fig. 1b; r=0.61 ± 0.06, SD across 5 videos, p<3.60$\cdot10^{-3}$). This is the case regardless of whether students were aware they would be tested or not (Intentional: r(25)=0.61, p=7.51$\cdot10^{-4}$, Incidental: r(29)=0.58, p=5.87$\cdot10^{-4}$). Evidently subjects with lower ISC performed poorer on the tests (e.g. subject 3 in Fig. 2a). Inversely, subjects with correlated eye movements obtain higher test scores (e.g. subject 1 & 2 in Fig. 2a). Basically, if subjects do not follow the dynamics of the video with their eyes, they have not paid attention and as a result their test scores are lower. Alternatively, subjects with prior knowledge on the material were more interested, and thus paid more attention.

Different styles of instructional videos

The effect we observed was true for all 5 videos tested (in Experiment 1: Intentional and in Experiment 2: Incidental). The style of these five videos consisted of either animation (lightbulbs, immune, internet) or showed a hand drawing figures (stars, birth). To test whether this effect is robust across different types of video styles, we performed an additional experiment on a new cohort of 30 subjects (Experiment 3; 22 females, 8 males, age 18-50, mean=25.73, SD=8.85 years). All subjects watched 6 videos on different topics produced in three different styles (two videos each): a real-live presenter along with animation, a presenter writing on a glass board, and writing-hand with animation. Despite the different visual appearance and dynamic, we still find a strong correlation between ISC and test scores for all three styles (Fig. 2d, Animation & Presenter: r(27)=0.53, p=3.1$\cdot10^{-3}$), Animation & Writing hand: r(28)=0.51, p=3.7$\cdot10^{-3}$), Glassboard & Presenter: r(27)=0.46, p=0.01).

Recognition and comprehension questions

It is possible that attention favors recognition of factual information, but that questions probing for comprehension of the material require the student to disengage from the video to process the content "offline". We therefore included in Experiment 3 comprehension questions (41 out of a total

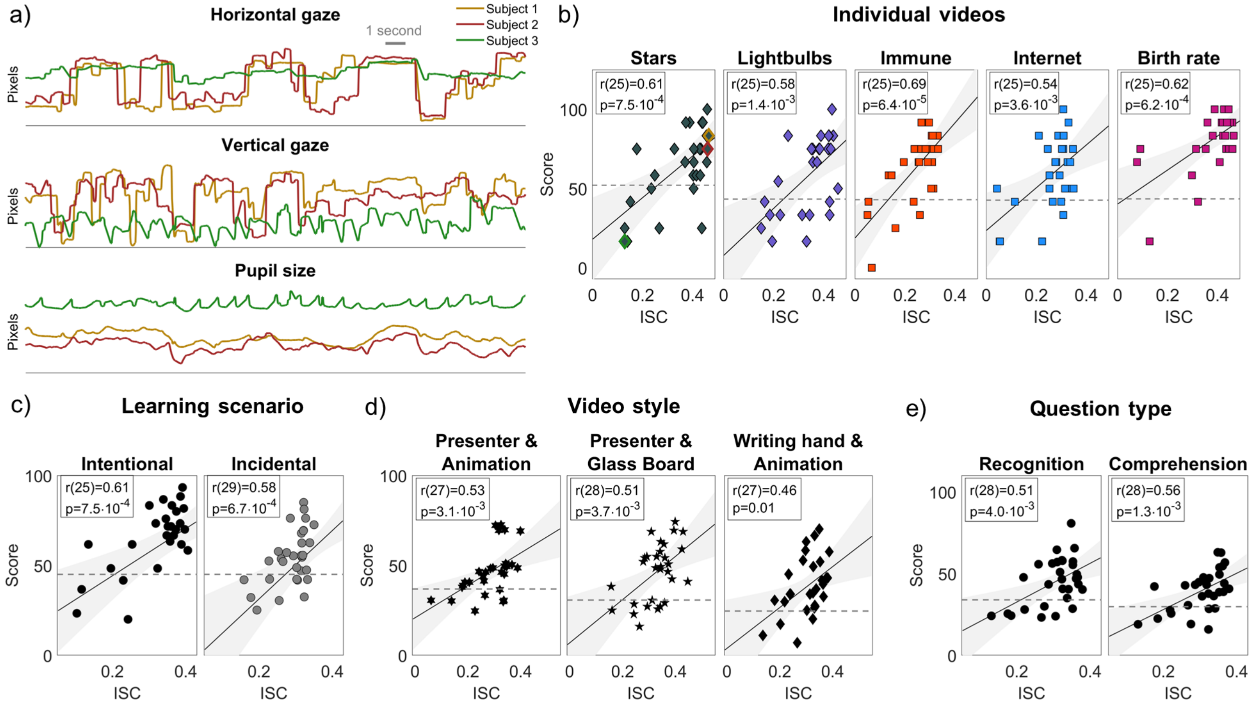

Figure 2: **Intersubject correlation of eye movements during instructional videos predicts learning performance.** **a)** Eye movements of three representative subjects as they watch Why are Stars Star-Shaped?. Two high performing subjects have similar eye movements and pupil size. A third, low performing student does not match their gaze position or pupil size. **b)** Intersubject correlation of eye movements (ISC) and performance on test taking (Score) for each of five videos in Experiment 1. Each dot is a subject. The high and low performing subjects (subjects 1-3) from panel (a) are highlighted for the Stars video. Dotted lines represent performance of subjects naïve to the video. **c)** Same as panel (b) but averaging over the 5 videos. The data was collected in two different conditions: During intentional learning (Experiment 1) where subjects knew they would be quizzed on the material. During incidental learning (Experiment 2) where subjects did not know that quizzes would follow the viewing. **d)** Videos in three different production styles (Experiment 3) show similar correlation values between test scores and ISC. Each point is a subject where values are averaged over two videos presented in each of the three styles. (See Fig. S2 for results on all 6 videos.) **e)** A similar effect is observed for different question types. Here each point is a subject with test scores averaged over all questions about factual information (recognition) versus questions requiring comprehension. ISC were averaged over all 6 videos in Experiment 3.

of 72 questions across the 6 videos). Overall subjects did similarly on the comprehension questions as compared to the recognition questions (Fig. 2e) and we find a significant correlation with ISC for these comprehension questions (r(28)=0.56,p=1.3$\cdot 10^{-3}$), and we again find a correlation with recognition performance (r(28)=0.51, p=4.0$\cdot 10^{-3}$). These correlation values do not differ significantly (asymptotic z-test after Fisher r-to-z conversion, p=0.52) suggesting that comprehension and recognition are both affected by attention. Indeed, test scores for comprehension and recognition questions are significantly correlated across subjects (r(28)=0.62 (p=2.34$\cdot 10^{-4}$)). Predicting comprehension performance may have important implications for educational practice.

Capturing eye movements online at scale using standard web cameras

Thus far all experiments were performed in a laboratory setting with a research grade eye-tracker. To test the approach in a realistic setting we developed an online platform that can operate on a large scale of users. The platform relies on standard web cameras and existing eye tracking software that can run on any web browser [26]. The software operates on the remote computer of the users and captures gaze position. In one experiment we recruited 82 students (female=21, age 18-40, mean=19.6, SD=2.7 years) from a college physics class to participate after their lab sessions using the desktop computers available in the classroom (Experiment 4: Classroom). In another experiment we recruited 1012 participants (female=443, age 18-64, mean=28.1, SD=8.4 years) on MTurk and Prolific. These are online platforms that assign tasks to anonymous subjects and compensate them for their work (Experiment 5: At-home). The subjects used the webcam on their own computers emulating the at-home setting typical for online learning. The gaze position data collected with the web camera is significantly noisier than using the professional eye tracker in the lab (Fig. 3a). To quantify this, we compute the accuracy of gaze position when subjects are asked to look at a dot on the screen (Fig. 3b). As expected, we find a significant difference in gaze position accuracy between the laboratory and the classroom (two-sample t-test t(69)=-7.73, p=6.3$\cdot 10^{-11}$) and a significant difference between the classroom and the at-home setting (t(242)=-2.46, p=0.01). Despite this signal degradation we find a high correlation between the median gaze position data for laboratory and classroom data (Horizontal gaze: r=0.87 ± 0.04; Vertical gaze: r=0.75 ± 0.04) and laboratory and at-home (Horizontal gaze: r=0.91 ± 0.04; Vertical gaze: r=0.83 ± 0.04).

Predicting test scores in a classroom and at home using web cameras

To preserve online privacy of the users we propose to evaluate eye movements remotely by correlating each subject's eye movements with the median gaze positions (Fig. 3a). Instead of ISC with all members of the group, we thus compute the correlation with the median position locally without the need to transmit individual eye position data (see Methods). To compensate for the loss of the pupil signal we now also measure the correlation of eye movement velocity, which is high when subjects move their gaze in the same direction, regardless of absolute gaze position (see Methods). We combine these eye movement metrics by taking a weighted average of the vertical, horizontal and velocity ISC (wISC; see Methods). We find that this wISC of eye-movement robustly correlates with subsequent test scores (Fig. 3; Tab. S1) despite the lower quality of the gaze position data. In fact, the correlation of wISC with test scores for the classroom (Fig. 3c; r=0.46 ± 0.16, p<0.01) are comparable to the values in the laboratory experiments (r = 0.59 ± 0.08, all p<0.01; compare to Fig. 2b). The at-home experiment had also highly significant correlation between wISC and subsequent test scores (Fig. 3d; r=0.47 ± 0.08, p<3.9$\cdot 10^{-8}$). The prediction accuracy of the test score is 14.59% ± 16.86% (median across videos, IQR across all videos and subjects), which is equivalent to 1.75 out of 12 questions. We can essentially predict how well a student is going to perform on a test by comparing their eye movements to the median eye movements.

Inherent ability versus attentional state

So far we have argued that variable attention modulates performance across students, and that attention is also reflected in eye movements. But it is also possible that an inherent ability, or trait of the students causes them to perform differently and this affects their eye movements as well. Thus, correlation between ISC and test scores could be induced by a variable trait, and not a variable state of the students. To assess these options, we asked students for their grade point averages in their college courses (GPA is available for experiments 2, 3 and 4), and asked subjects to perform a digit span test (in Experiment 2 and 3). The GPA is well-known to predict individual test scores [27], and the digit span test is a simple test for working memory capacity [28]. We found a significant correlation across subjects between the digit span test and the test scores for the videos (Experiment 2: r(24)=0.41, p=0.04; Experiment 3; r(28)=0.39 , p=0.03). We also found a significant correlation between the students self-reported GPA and the scores they obtained (Experiment 2: r(27)=0.38, p=0.04, Experiment 3: r(22)=0.37, p=0.06; Experiment 4: r=0.38 +- 0.10, p<0.04). However, we found no correlation between the eye movement measures and either the digit span test (Experiment 2: r(24)=-0.12, p=0.56; Experiment 3: r(28)=0.27, p=0.16) nor the GPA (Experiment 2: r(24)=0.03, p=0.86; Experiment 3: r(28)=0.24, p=0.25). Given that ISC correlates with performance this suggests that in addition to attentional state, test-taking performance is affected by an inherent trait or ability. However, this trait may have no effect on attentional state, or at least it is not reflected in the ISC of eye movements.

Subject or stimuli effect

We have so far shown that the level of attention as measured by ISC can predict how well students will perform on a test quiz. However, students could already be familiar with a topic covered in one of the instructional videos, and hence already know the answers to the quiz questions. This could lead the student to pay more attention to the videos, due to their interest in the topic (stimuli effect). On the other hand it could

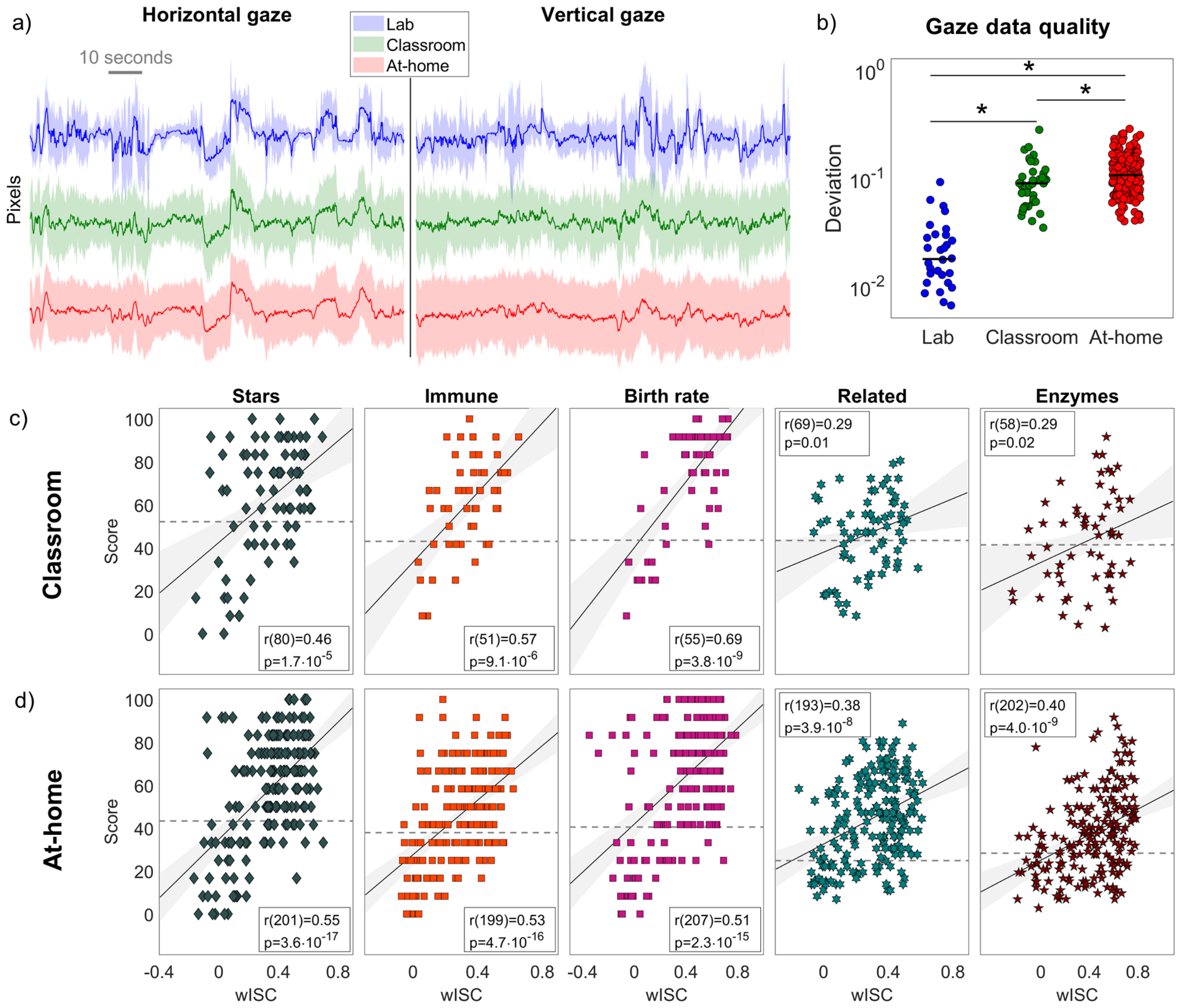

Figure 3: **Weighted inter-subject correlation of eye movement measured using low-cost web camera predicts test scores. a)** Gaze position for 'Immune' video in Laboratory, Classroom and At-home settings. Median and interquartile range are taken across subjects (solid line and grayed area respectively). **b)** Deviation of gaze position when subjects looked at 4 "validation" dots presented in sequence on the corners of the screen, collected in the Laboratory, Classroom and At-home settings for the first video shown to subjects (see Methods). *indicates a significant difference in means. **c)** Eye-movement wISC is predictive of performance in the Classroom. **d)** Eye-movement wISC is predictive of performance in the At-home setting.

be that some students just pay more attention throughout the experiment and as a consequence they will both have higher ISC and test scores (subject effect). We test the *subject effect* by subtracting the subject average of score and ISC from the ISC and Scores of the test quizzes. If the correlation between ISC and test scores is driven by the subject, by subtracting the subject average, the effect would disappear. This is what we find, there is no significant correlation between ISC and Score in any of the three experiments (Fig. S6a-c; Experiment 1: r(25)=-0.25, p=0.21, Experiment 2: r(29)=-0.16, p=0.40, Experiment 3: r(28)=-0.05, p=0.78), suggesting there is a strong subject effect. If the correlation was due to familiarity with a topic, we can test the *stimuli effect* by subtracting the stimuli average of score and ISC from the ISC and test scores. We find significant correlation between ISC and Score in all three experiment (Fig. S6d-f; Experiment 1: r(25)=0.71, p=3.8$\cdot 10^{-5}$, Experiment 2: r(29)=0.54, p=1.9$\cdot 10^{-3}$, Experiment 3: r(28)=0.59, p=5.5$\cdot 10^{-4}$), suggesting there is no stimuli effect.

Discussion

We found that eye movements during viewing of instructional videos are similar between students, but only if they are paying attention. The effect is strong, allowing one to detect with a few minutes of gaze-position data if a student is distracted. Consequently, and as predicted, we find that students performed well in subsequent quizzes if their eyes followed the material presented during the video in a stereotypical pattern. We replicated this finding in two subsequent laboratory experiments, where we confirmed that the effect persists when students do not expect to be quizzed, and that the effect does not depend on the specific type of video or the type of questions asked. The results also replicate in a classroom setting and in a large scale online experiment with users at home using standard web cameras. By correlating to the median gaze-positions one can avoid transmitting personal data over the internet. Thus we conclude that one can detect students' attentional engagement during online education with readily available technology. In fact, we can predict how well a student will perform on a test related to an instructional video, by looking at their eyes while maintaining online privacy.

Note that our study was purely correlational. It is possible that stronger students can both follow the video better and also perform better in the test, without the need for a direct link between the two. We build an analytic model assuming a common cause for inter-subject correlation and test performance. While we refer to this common cause as "attention", it really can refer to any internal state of a subject that may have a causal effect on test scores and eye movements such as alertness, interest, motivation, engagement, fatigue, etc. This causal model explained the data more accurately than a simple correlation. But ultimately, our study did not control attention prospectively and thus cannot conclusively answer the direction of this relationship.

Even if our causal model is correct, it is not clear that this common cause is a state of the student or a generic ability. In fact, performance correlated with GPA and working memory capacity. In contrast, these traits did not correlate with ISC of their eye movements. This suggests that a portion of the test scores are affected by a generic ability of the student, in addition to the attentional state that can be gleaned from their eye movements.

We tested for recall of factual information presented in the videos. Performance on these questions naturally depend on attention to the presentation of this factual information. For questions requiring comprehension, instead, it may be that students need time to think about material quietly without being absorbed by the video. Yet, we did not find a degradation in the ability to predict test scores from eye movement for the comprehension questions. However, a more nuanced analysis and larger sample size may be needed to establish a difference in our ability to predict comprehension and recall performance.

ISC of eye movements varied significantly between subjects and videos. The variability between subjects is to a certain degree predictive of different test scores and thus we can ascribe it to genuine differences in attention. However, there is a significant variability in ISC across subjects even in the distracted condition, suggesting that baseline levels of ISC do vary between subjects, irrespective of attention. ISC also differed significantly among videos. This again could be due to different levels of attention that the videos elicit, but it could also be due to differences in visual dynamic (slower videos may drive eye movements less vigorously). Therefore, one caveat of using ISC is that its values should always be compared against a baseline that calibrates for these differences among videos [29].

A number of previous studies show the merits of using eye tracking to evaluate online education. For instance, when learning from pictures and written text, fixation times and re-reading predict learning performance [30]. Showing the instructor's face while talking seems to help students' attend to the material [31, 32], but there are mixed results on whether this is actually beneficial for learning [33, 34]. These types of results required careful analysis of the exact content that is fixated upon. The method presented here assesses whether students are paying attention without the need for specific information about the contents of the video.

Our analysis also included pupillary responses. That this should correlate between subjects is not surprising as it is strongly driven by luminance changes in the visual stimulus. The novel observation is that this correlation is modulated by attention. This may be a consequence of the similar eye movements, as this will lead to similar luminance fluctuations in foveal vision, or may result from the effects of attention [31] or arousal [32] on pupillary response. The present finding differs from the extensive literature on pupil size which attempt to link pupillary response to specific events. For example, pupil size predicted reliably which stimuli were recalled, in particular for emotionally arousing stimuli [35]. Pupil size has also been linked with cognitive effort, for instance, the effort associated with holding multiple items in working memory [36]. In contrast to this traditional work on event-related pupil dilation we did not have to analyze the specific content of the stimulus. As with the eye movements, we can simply use other viewers as a reference to determine if the pupil size is correlated, and if it is, anticipate high test scores.

Online education often struggles to persistently engage students' attention, which may be one of the causes for low retention [37]. Student online engagement is often measured in terms of the time spent watching videos [38], mouse clicks [39] or questionnaires [40], and some important lessons have been gained from these outcome measures. For instance, videos should be short, dynamic, and show the face of the instructor talking with enthusiasm [38]. Our recent work has focused on measuring attentional engagement of the students by measuring their actual brain activity [16]. Attempts to record brain signals in a classroom have been made [41, 42], but typically require help from research personnel and may thus not be practical, particularly at home. The method we have presented here opens up the possibility to measure not just time spent with the material, but the actual engagement of the student's mind with the material, regardless of where they are. With adequate data quality one may be able to even adapt the content in real time to the current attentional state of the student. In particular, for synchronous online course, where students participate at the same time, real-time feedback to the teachers may allow them to adapt to students' level of attention in real-time, much like real teachers in real classrooms. The internet has turned attention into a commodity. With video content increasing online, remote sensing of attention to video at scale may have applications beyond education, including entertainment, advertising, or politics. The applications are limitless.

Methods

Participants

1182 subjects participated in one of five different experimental conditions. The first two experiments tested the learning scenario of online education, namely intentional learning (Experiment 1, N=27, 17 females, age 18-53 M=26.74, SD=8.98, 1 subject was removed due to bad data quality) and incidental learning (Experiment 2, N=31, 20 females, age range 18-50, mean 26.20, SD 8.30 years; 3 subjects were removed due to bad signal quality). Experiment 3, was designed to investigate the effect of different video styles and assessment types (N=31, 22 females, age 18-50, M=25.73, SD=8.85 years; 2 subjects were removed due to bad signal quality). Participants for the laboratory Experiments 1-3 were recruited from mailing lists of students at the City College of New York and local newspapers ads (to ensure a diverse subject sample). Experiment 4 was designed to replicate the findings from the laboratory in a classroom setting. Participants were all enrolled in the same physics class at the City College of New York (N=82, female=21, age 18-40, M=19.6, SD=2.7 years). Experiment 5 replicated the finding from the laboratory in a home setting. Amazon Mechanical Turk and Prolific was used to recruit subjects (N=1012, 473 female, age range 18-64, M=28.1, SD=8.4 years). Subjects of Experiments 1-4 only participated in a single experiment, i.e. they were excluded from subsequent Experiments. In Experiment 5 subjects were allowed to participate in more than one assignment so the total count are not unique subjects. The experimental protocol was approved by the Institutional Review Boards of the City University of New York. Documented informed consent was obtained from all subjects for laboratory experiments. Internet-based informed consent was given by subjects that were recruited for the online experiments.

Stimuli

The five video stimuli used in Experiments 1, 2, 4 and 5 were selected from the 'Kurzgesagt – In a Nutshell' and 'minute physics' YouTube channels. They cover topics relating to physics, biology, and computer science (Table 1, Range: 2.4 – 6.5 minutes, Average: 4.1 ± 2.0 minutes). Two of the videos ('Immune' and 'Internet') used purely animations, where 'Boys' used paper cutouts and handwriting. 'Bulbs' and 'Stars' showed a hand drawing illustrations aiding the narrative. The six video stimuli used in Experiments 3-5 were selected from 'Khan Academy', 'eHow', 'Its ok to be smart' and 'SciShow'. The videos cover topics related to biology, astronomy and physics (Table 1, Duration: 4.2 – 6 minutes long, Average: 5.15 ± 57 seconds). They were specifically chosen to follow recommendations from a large scale MOOC analysis [38]. The three styles chosen were based on popular styles from YouTube. 'Mosquitoes' and 'Related' produced in the 'Presenter & Animation' style shows a presenter talking as pictures and animations are shown. 'Planets' and 'Enzymes' were produced in the 'Presenter & Glass Board' style and shows a presenter drawing illustrations and equations on a glass board facing the viewer. 'Capacitors' and 'Work energy' used the 'Animation & Writing hand' style.

Procedure

Laboratory experiments

In Experiment 1 (intentional learning), subjects watched a video and answered afterwards a short four-alternative forced-choice questionnaire. The subjects were aware that they would be tested on the material. The test covered factual information imparted during the video (11 – 12 recall questions). Examples of questions and answer options can be found in Tab. 1. In Experiment 2 (incidental learning) subjects were not aware that they would be tested or asked questions regarding the material. They first watched all 5 videos, and subsequently answered all the questions. In Experiment 3, subjects were informed that questions regarding the material would be presented after each video and followed the procedure of Experiment 1, using a different set of stimuli. The order of video presentation, questions and answer options were randomized for all three experiments. Common for Experiments 1-3, after subjects had watched all video stimuli and answered questions, they watched all the videos again in a distracted condition using the same order as the attend condition. In this condition participants counted backwards, from a randomly chosen prime number between 800 and 1000, in decrements of 7. This task aimed to distract the subjects from the stimulus without requiring overt responses and is based on the serial subtraction task used to assess mental capacity and has previously been used to assess attention [7].

Online experiments

The web camera experiments (Experiments 4 and 5) were carried out using Elicit (???), a framework developed for online

Experiment #	Title	Abbreviation	Video Style	Duration (min:sec)	URL-ending*	Topic area	Example question	Example answer choices
1,2,4,5	Why are Stars Star-Shaped?	Stars	Animation & Writing hand	3:28	VVAKFJ8VVp4	Physics	What causes "suture lines"?	1. Where the fibers that make up the eye's lens meet 2. Health problems 3. Short-sightedness 4. All options are correct
1,2	Why Do We Have More Boys Than Girls?	Birth rate	Animated	2:48	3IaYhG11ckA	Biology	What is the ratio of boys to girls born worldwide?	1. 106:100 2. 100:100 3. 96:100 4. None of the options are correct
1,2,4,5	The Immune System explained	Immune	Animated	6:48	zQGOcOUBi6s	Biology	What is the main job of the macrophage cell?	1. To kill enemies 2. To cause inflammation 3. To activate cells 4. All options are correct
1,2	How modern Light Bulbs work	Bulbs	Animation & Writing hand	2:57	oCEKMEeZXug	Physics	Which gas exists in halogen light bulbs?	1. Hydrogen bromide 2. Mercury chloride 3. Nitrogen bromide 4. Nitrogen fluoride
1,2	Who invented the Internet? and why?	Internet	Animated	6:32	21eFwbb48sE	Computer Science	What was the goal of the first network?	1. Optimizing processor usage 2. Facilitating communication 3. Sharing research materials 4. Espionage
3	What if we killed all the Mosquitoes?	Mosquitoes	Presenter & Animation	4.21	e0NT9i4Qnak	Biology	Anopheles is the primary vector for:	1. Malaria 2. Dengue 3. Yellow fever 4. Zika
3,4,5	Are we all related?	Related	Presenter & Animation	6.03	mnYSMhR3jCI	Biology	How much of the human DNA is coded into proteins?	1. 2% 2. 98% 3. 80% 4. 30%
3	Dielectrics in Capacitors	Capacitors	Writing hand & Animation	5.46	rkntp_3cZl4	Physics	What happens when a dielectric is inserted in the capacitor which is in a circuit with a battery?	1. Charge Q increases 2. Charge Q decreases 3. Voltage V increases 4. Voltage V decreases
3	Work and the work-energy principle	Work energy	Writing hand & Animation	6.26	30o4omX5qfo	Physics	A person pushes a box along a horizontal floor at a constant speed. The net work done on the box is:	1. Zero 2. Positive 3. Negative 4. It depends
3,4,5	How do people measure Planets and Suns?	Planets	Presenter & Glass Board	4.23	bYgV9nvgJ3E	Astronomy	As the size of a star increases, the angular measurement needed in the stellar parallax technique...	1. Does not matter 2. Increases 3. Decreases 4. Stays the same
3,4,5	What function does an Enzyme have?	Enzymes	Presenter & Glass Board	4.29	lkRZKqDdwzU	Biology	What is the value of the activation energy in the example shown in the graph?	1. 3 2. 7 3. 4 4. 5

Table 1: Experiment, title, abbreviation, style, duration, web address, and example questions and answer choices. URL beginning with `https://www.youtube.com/watch?v=`

experiments. In Experiment 4 (classroom) students used the same computers they use for their class exercises. From the Elicit webpage subjects could select which video they wanted to watch from a list of 5 videos. Subjects were given a short verbal instruction besides the written instructions that were provided through the website. In Experiment 5 (at-home) subjects could select HITs (Amazon Mechanical Turk assignments) or assignments (Prolific) that contained a single video with questions and otherwise followed the same procedure as Experiment 4. For both Experiment 4 and 5, subjects were informed that there would be questions regarding the material after the video. They first received instructions regarding the procedure, performed the webcam calibration to enable tracking of their eye movements, watched a single video and answered a four-alternative choice questionnaire for that video. Subjects were allowed to perform more than one assignment, i.e. view more than one video and answer questions. In Experiment 5 subjects were additionally shown a short instruction video on how to calibrate the webcam to track eye movements.

Online eye tracking using web cameras

The webcam-based gaze position data was recorded using WebGazer [26]. WebGazer runs locally on the subject's computer and uses their webcam to compute their gaze position. The script fits a wireframe to the subject's face and captures images of their eyes to compute where on the screen they are looking. Only the gaze position and the coordinates of the eye images used for the eye position computation were transmitted from the subject's computer to our web server. In order for the model to compute where on the screen the participant is looking, a standard 9-point calibration scheme was used. Subject had to achieve a 70% accuracy to proceed in the experiment. Note that here we did transfer user data to the server for analysis. However, in a fully local implementation of the approach no user data would be transmitted. Instead, median eye positions of a previously recorded group would be transmitted to the remote location and median-to-subject correlation could be computed entirely locally.

Preprocessing of webcam-based gaze position data

WebGazer estimates point of gaze on the screen as well as the position and size of the eyes on the webcam image. Eye position and size allowed us to estimate the movement of the subject in horizontal and vertical directions. The point of gaze and eye image position & size were upsampled to a uniform 1000Hz, from the variable sampling rate of each remote webcam (typically in the range of 15-100Hz). An inclusion criteria for the study was that the received gaze position data should be sampled at at least 15Hz in average. Missing data were linearly interpolated and the gaze positions were denoised using a 200ms and 300ms long median filter. Movements of the participant were linearly regressed out of the gaze position data using the estimated position of the participant from the image patch coordinates. This was done since the estimated gaze position is sensitive to movements of the subject (we found this increased the overall ISC). Subjects that had excessive movements were removed from the study (16 out of 1159 subjects; excessive movement is defined as 1000 times the standard deviation of the recorded image patch coordinates in the horizontal, vertical and depth directions). Blinks were detected as peaks in the vertical gaze position data. The onset and offset of each blink were identified as a minimum point in the first order temporal derivative of the gaze position. Blinks were filled using linear interpolation in both the horizontal and vertical directions. Subjects that had more than 20% of data interpolated using this method was removed from the cohort (14 out of 1159 subjects). We could not compute the visual angle of gaze since no accurate estimate was available for the distance of the subject to the screen. Instead, gaze position is measured in units of pixels, i.e. where on the screen the subject is looking. Since the resolutions of computer screens varies across subjects, the recorded gaze position data in pixels were normalized to the width and height of the window the video was played in (between 0 and 1 indicating the edges of the video player). Events indicating end of the video stimuli ("stop event") were used to segment the gaze position data. The start time for each subject was estimated as the difference between the stop event and the actual duration of the video. This was done, since the time to load the YouTube player was variable across user platforms.

Estimate of the quality of gaze position

To compute the quality of the gaze position data, subjects were instructed to look at a sequence of 4 dots in each corner of the screen, embedded in the video stimuli before and after the video. The actual dot position on the subjects screen was computed and compared to the captured eye gaze position of the WebGazer. The deviation was computed as the pooled deviation of the recorded gaze position from the position of the dot, while the subject looked at each dot. Poor data quality is indicated by higher deviation. Furthermore, subjects with low quality calibration were identified by computing the spatial difference of recorded gaze position data of opposing dots in the horizontal and vertical direction when they were looking at the 4 dots. If the difference in recorded gaze position between dot pairs were in average negative the subject was excluded (135 of 1159).

Preprocessing of laboratory gaze position data

In the laboratory (Experiments 1-3) gaze position data was recorded using an Eyelink 1000 eye tracker (SR Research Ltd. Ottawa, Canada) at a sampling frequency of 500 Hz using a 35mm lense. The subjects were free to move their heads, to ensure comfort (no chin rest). A standard 9-point calibration scheme was used utilizing manual verification. To ensure stable pupil size recordings, the background color of the calibration screen and all instructions presented to the subjects were set to be the average luminance of all the videos presented during the experiment. In between each stimulus presentation a drift-check was performed and tracking was recalibrated if the visual angular error was greater than 2 degrees. Blinks were detected using the SR research blink detection algorithm and remaining peaks were found using a peak picking algorithm. The blink and 100ms before and after were filled with linearly interpolated values.

Intersubject correlation and attention analysis of gaze position data

Intersubject correlation of eye movements is calculated by (1) computing the Pearson's correlation coefficient between a single subject's gaze position in the vertical direction with that of all other subjects while they watched a video. (2) obtaining a single ISC value for a subject by averaging the correlation values between that subject and all other subjects (ISC) (3) and then repeating steps 1 and 2 for all subjects, resulting in a single ISC value for each subject. We repeat step 3 for the horizontal eye movements $\text{ISC}_{horizontal}$ and the pupil size ISC_{pupil}. To obtain the measure used for laboratory experiment we averaged the three ISC values which we call $\text{ISC}=(\text{ISC}_{vertical}+\text{ISC}_{horizontal}+\text{ISC}_{pupil})/3$. The ISC values for the attend and distract conditions, were computed on the data for the two conditions separately. To test whether ISC varies between the attend and distract conditions, a three-way repeated measures ANOVA was used with fixed effect of video and attentional state (attend vs. distract) and random effect of subject. As an additional measure the receiver operating characteristic curve (ROC) was used. Each point on the curve is a single subject. To quantify the overall ability of ISC to discriminate between attend and distract conditions the area under the ROC curve is used (AUC). To test for the effect motivation has, ISC was computed for each video in the attend condition and averaged across all videos. Since the distribution was not Gaussian, we tested for a difference in median ISC values with a Wilcoxon rank sum test. To test for the effect of video style on the attentional modulation of ISC we performed a three-way repeated measures ANOVA. The random effect was subject and fixed effects were stimuli, attentional condition and video style.

Weighted intersubject correlation of eye movements

For the experiments with the web camera in the classroom and at-home we compute for each time point in the video the median gaze position across all subjects (Fig. 3a). We then compute the Pearson's correlation coefficient of that median time course with the gaze position of each subject. We refer to this as median-to-subject correlation, $\text{MSC}_{vertical}$ and $\text{MSC}_{horizontal}$. Note that in principle this can be computed with the median gaze positions previously collected on a sample group for each video. To compute this remotely without transmitting the gaze data of individual users, one would transmit this median gaze positions to the remote user of the online platform (two values for each time point in the video). MSC can then be computed locally by the remote user. We additionally compute MSC for the velocity of eye movements as follows. First we compute movement velocity by taking the temporal derivative of horizontal and vertical gaze positions using the Hilbert transform. We form two-dimensional spatial vectors of these velocity estimates (combining Hilbert transforms of horizontal and vertical directions). These vectors are normalized to unit length. The median gaze velocity vectors is obtained as the median of the two coordinates across all subjects. The median-to-subject correlation of velocity, $\text{MSC}_{velocity}$, is then computed as the cosine distance between the velocity vectors of each subject and the median velocity vector, averaged over time. Finally, we combine the three MSC measures to obtain a single weighted intersubject correlation value for each subject: $\text{wISC} = w_1\text{MSC}_{vertical} + w_2\text{MSC}_{horizontal} + w_3\text{MSC}_{velocity}$. The weights w_i are chosen to best predict test scores with the constraint that they must sum up to 1 and that they are all positive. This is done with conventional constrained optimization. The constraints insure that the wISC values are bounded between -1 and 1. To avoid a biased estimate of predictability we optimize these weights for each subject on the gaze/score data leaving out that subject from the optimization, i.e. we use leave-one out cross-validation.

Student learning assessment

Four-choice, multiple-choice questions were used to assess the performance of students (Score). Test performance was calculated as the percentage correct responses each student gave for each video. For questions that had multiple correct options, points were given per correct selected options and subtracted per incorrect selected option. The questionnaires were designed in pilot experiments to yield an even distribution of answer options from subjects that had not seen the videos. All questions and answer options can be found here. To estimate the baseline difficulty of the questions, separate naïve cohorts of subjects were given the same questions without seeing the videos. Two different cohorts were recruited from the City College of New York to compare against the cohorts recruited for Experiments 1-4 (Experiment 1,2 and 4, N=26; Experiment 3, N=15) and a third from Prolific to compare against the at-home experiment cohort (Experiment 5, N=25). When evaluating the different learning styles (incidental and intentional learning) in Experiments 1 and 2, students' scores and ISC values were averaged across all videos. ISC was compared to student test performance by computing the Pearson's correlation coefficient between ISC and test performance. Similarly, to test the effect of video style, the ISC and scores for each subject were averages for the videos produced in different styles and correlated using Pearson's correlation. Testing the connection between ISC and test scores on each individual video, subjects' scores were compared with the ISC using Pearson's correlation. To test whether there is a significant difference in correlation between comprehension or recall questions and ISC we used the same ISC values and performed a test between correlation values with a shared dependent variable [43]. Testing how well eye-movement ISC can predict the performance of students on tests regarding the material in the online setting, we use leave-one-out cross validation. We estimate the attention model (see Supplement for description)on all subjects leaving but one subject's ISC values and their corresponding test scores. We then estimate how well ISC predicts the test score on the left out subject. We do this for all subjects and compute the median absolute deviation between the prediction and the actual score. To test if our eye-movement ISC model is statistically better than a naïve model (only predicting the average score), we subtract the prediction errors of the two models and perform a two-sided sign test.

References

[1] D. J. Murray and Helen E. Ross. Vives (1538) on memory and recall. *Canadian Psychology/Psychologie canadienne*, 23(1):22–31, 1982.

[2] Heiner Deubel and Werner X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36(12):1827–1837, June 1996.

[3] James E. Hoffman and Baskaran Subramaniam. The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6):787–795, January 1995.

[4] Geoffrey R. Loftus. Eye fixations and recognition memory for pictures. *Cognitive Psychology*, 3(4):525–551, October 1972.

[5] Charlotte E. Wolff, Niek van den Bogert, Halszka Jarodzka, and Henny P. A. Boshuizen. Keeping an Eye on Learning: Differences Between Expert and Novice Teachers' Representations of Classroom Management Events. *Journal of Teacher Education*, 66(1):68–85, January 2015.

[6] Hans-Jürgen Bucher and Peter Schumacher. The relevance of attention for selecting news content. An eye-tracking study on attention patterns in the reception of print and online media. *Communications*, 31(3):347–368, 2006.

[7] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052, 2008.

[8] David A. Slykhuis, Eric N. Wiebe, and Len A. Annetta. Eye-Tracking Students' Attention to PowerPoint Photographs in a Science Education Setting. *Journal of Science Education and Technology*, 14(5):509–520, December 2005.

[9] Fang-Ying Yang, Chun-Yen Chang, Wan-Ru Chien, Yu-Ta Chien, and Yuen-Hsien Tseng. Tracking learners' visual attention during a multimedia presentation in a real classroom. *Computers & Education*, 62:208–220, March 2013.

[10] Jacek P. Dmochowski, Paul Sajda, Joao Dias, and Lucas C. Parra. Correlated Components of Ongoing EEG Point to Emotionally Laden Attention – A Possible Marker of Engagement? *Frontiers in Human Neuroscience*, 6, 2012.

[11] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, 303(5664):1634–1640, March 2004.

[12] K. Lankinen, J. Saari, R. Hari, and M. Koskinen. Intersubject consistency of cortical MEG signals during movie viewing. *NeuroImage*, 92:217–224, May 2014.

[13] Uri Hasson, Ohad Landesman, Barbara Knappmeyer, Ignacio Vallines, Nava Rubin, and David J. Heeger. Neurocinematics: The Neuroscience of Film. *Projections*, 2(1):1–26, June 2008.

[14] Samantha S. Cohen and Lucas C. Parra. Memorable Audiovisual Narratives Synchronize Sensory and Supramodal Neural Responses. *eNeuro*, 3(6):ENEURO.0203–16.2016, November 2016.

[15] Uri Hasson, Orit Furman, Dav Clark, Yadin Dudai, and Lila Davachi. Enhanced Intersubject Correlations during Movie Viewing Correlate with Successful Episodic Encoding. *Neuron*, 57(3):452–462, February 2008.

[16] Samantha S. Cohen and Jens Madsen, Gad Touchan, Denise Robles, Stella F. A. Lima, Simon Henin, and Lucas C. Parra. Neural engagement with online educational videos predicts learning performance for individual students. *Neurobiology of Learning and Memory*, 155:60–64, November 2018.

[17] Samantha S. Cohen, Simon Henin, and Lucas C. Parra. Engaging narratives evoke similar neural activity and lead to similar time perception. *Scientific Reports*, 7(1):4578, July 2017.

[18] U. Hasson, E. Yang, I. Vallines, D. J. Heeger, and N. Rubin. A Hierarchy of Temporal Receptive Windows in Human Cortex. *Journal of Neuroscience*, 28(10):2539–2550, March 2008.

[19] John M. Franchak, David J. Heeger, Uri Hasson, and Karen E. Adolph. Free Viewing Gaze Behavior in Infants and Adults. *Infancy*, 21(3):262–287, 2016.

[20] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, and Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10):28–28, August 2010.

[21] Christoforos Christoforou, Spyros Christou-Champi, Fofi Constantinidou, and Maria Theodorou. From the eyes and the heart: a novel eye-gaze metric that predicts video preferences of a large audience. *Frontiers in Psychology*, 6, 2015.

[22] Kate Burleson-Lesser, Flaviano Morone, Paul DeGuzman, Lucas C. Parra, and Hernán A. Makse. Collective Behaviour in Video Viewing: A Thermodynamic Analysis of Gaze Position. *PLOS ONE*, 12(1):e0168995, January 2017.

[23] Mary Hegarty, Sarah Kriz, and Christina Cate. The Roles of Mental Animations and External Animations in Understanding Mechanical Systems. *Cognition and Instruction*, 21(4):209–249, December 2003.

[24] Richard E. Mayer, Mary Hegarty, Sarah Mayer, and Julie Campbell. When Static Media Promote Active Learning:

Annotated Illustrations Versus Narrated Animations in Multimedia Instruction. *Journal of Experimental Psychology: Applied*, 11(4):256–265, 2005.

[25] Frank W. Schneider and B. L. Kintz. An analysis of the incidental-intentional learning dichotomy. *Journal of Experimental Psychology*, 73(1):85–90, 1967.

[26] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence - IJCAI 2016*, January 2016.

[27] Benjamin S. Bloom. *Human characteristics and school learning.* Human characteristics and school learning. McGraw-Hill, New York, NY, US, 1976.

[28] Alan S. Kaufman and Elizabeth O. Lichtenberger. *Assessing Adolescent and Adult Intelligence.* John Wiley & Sons, August 2005.

[29] Jason J. Ki, Simon P. Kelly, and Lucas C. Parra. Attention Strongly Modulates Reliability of Neural Responses to Naturalistic Narrative Stimuli. *Journal of Neuroscience*, 36(10):3092–3101, March 2016.

[30] Sheng-Chang Chen, Hsiao-Ching She, Ming-Hua Chuang, Jiun-Yu Wu, Jie-Li Tsai, and Tzyy-Ping Jung. Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities. *Computers & Education*, 74:61–72, May 2014.

[31] Sebastiaan Mathôt and Stefan Van der Stigchel. New Light on the Mind's Eye: The Pupillary Light Response as Active Vision. *Current Directions in Psychological Science*, 24(5):374–378, October 2015.

[32] Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008.

[33] Logan Fiorella, Andrew T. Stull, Shelbi Kuhlmann, and Richard E. Mayer. Instructor presence in video lectures: The role of dynamic drawings, eye contact, and instructor visibility. *Journal of Educational Psychology*, pages No Pagination Specified–No Pagination Specified, 2018.

[34] Margot van Wermeskerken, Susanna Ravensbergen, and Tamara van Gog. Effects of instructor presence in video modeling examples on attention and learning. *Computers in Human Behavior*, 89:430–438, December 2018.

[35] Anne Bergt, Anne E. Urai, Tobias H. Donner, and Lars Schwabe. Reading memory formation from the eyes. *European Journal of Neuroscience*, 47(12):1525–1533, 2018.

[36] Tepring Piquado, Derek Isaacowitz, and Arthur Wingfield. Pupillometry as a Measure of Cognitive Effort in Younger and Older Adults. *Psychophysiology*, 47(3):560–569, May 2010.

[37] Sara Isabella de Freitas, John Morgan, and David Gibson. Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. *British Journal of Educational Technology*, 46(3):455–471, 2015.

[38] Philip J. Guo, Juho Kim, and Rob Rubin. How Video Production Affects Student Engagement: An Empirical Study of MOOC Videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 41–50, New York, NY, USA, 2014. ACM. event-place: Atlanta, Georgia, USA.

[39] Marian Petre and Mary Shaw. What's the Value Proposition of Distance Education? *ACM Inroads*, 3(3):26–28, September 2012.

[40] Chin Choo Robinson and Hallett Hullinger. New Benchmarks in Higher Education: Student Engagement in Online Learning. *Journal of Education for Business*, 84(2):101–109, November 2008.

[41] Suzanne Dikker, Lu Wan, Ido Davidesco, Lisa Kaggen, Matthias Oostrik, James McClintock, Jess Rowland, Georgios Michalareas, Jay J. Van Bavel, Mingzhou Ding, and David Poeppel. Brain-to-Brain Synchrony Tracks Real-World Dynamic Group Interactions in the Classroom. *Current Biology*, 27(9):1375–1380, May 2017.

[42] Andreas Trier Poulsen, Simon Kamronn, Jacek Dmochowski, Lucas C. Parra, and Lars Kai Hansen. EEG in the classroom: Synchronised neural recordings during video presentation. *Scientific Reports*, 7:43916, March 2017.

[43] James H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251, 1980.

Data Availability

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The author(s) declare no competing interests.

Author contributions

J.M. and L.C.P. designed the study, J.M. collected the data with help from S.U.J. and P.J.G, J.M. analyzed the data, L.C.P. build and implemented the analytic model, J.M. and L.C.P. wrote the manuscript, R.S. and S.U.J. designed the questions for the student assessment.

Acknowledgements

We would like to thank Iain Bryson from *Freeroaming Solutions* for his invaluable help in developing Elicit, which allowed us to carry out the online experiments. Without him this would not have been possible. We would like to acknowledge the help of James Hedberg for supporting collection of data from the physics classroom cohort. We thank Uri Hasson, Ofer Tchernichowski and Hernan Makse for helpful comments on an early version of this manuscript. We would like to acknowledge National Science Foundation Grant DRL-1660548 for supporting this project.

Supplement

Details of figure 3

Intersubject correlation of eye movements: a measure of attentional engagement

Intersubject correlation of eye movements is modulated by attention

In Figure 1c of the main manuscript we showed the result of the attention manipulation task on the ISC of eye movements and pupil size when subjects watched video stimuli. Here we extend this analysis to Experiments 2 and 3.

For Experiment 2 (Incidental learning, Figure S1a), we perform the same three-way repeated measures ANOVA and find a significant main effect of subject (F(29,257)=4.80, p=1.65e$\cdot10^{-12}$)), a significant main effect of stimuli ($F(4,257)$=42.08, p=4.00e$\cdot10^{-27}$)) and importantly, a significant main effect of attentional state ($F(1,257)$=1213.27, p=2.53$\cdot10^{-99}$)). This suggests that regardless of the different instructions given to the subjects in the two learning scenarios, the measure of ISC is still able to discern the two attentional conditions.

In Experiment 3 six new videos were selected and the attention manipulation task was again used to test the ability of ISC to discern attentional states (Figure S1b). We perform a three-way repeated measures ANOVA and find a significant main effect of subject ($F(28,287)$=5.41, p=1.49$\cdot10^{-14}$), a significant main effect of stimuli ($F(5,287)$=19.88, p=5.15e$\cdot10^{-17}$)) and importantly, a significant main effect of attentional state ($F(1,287)$=1357.63, p=8.21e$\cdot10^{-111}$)). This indicates the robustness of ISC of eye movements working for a total of eleven different videos.

Intersubject correlation of eye movements modulated by attention for multiple video styles

As education is moving to the online domain, teaching material is growing in abundance and so are the different video styles. We wanted to test if ISC of eye movements as a measure of attentional engagement generalizes to some of the most popular video styles, which are found on the major educational channels of YouTube. In Experiments 1 and 2 subjects watched educational videos produced using the 'Animations' (N=3) and 'Writing hand & Animation' styles (N=2). In both experiments subjects (N=27,30) watched the videos both in an attentive and distracted condition. The resulting ISC scores for the two conditions can be seen on Figure S1c).

Importantly, in Experiment 1 we find that ISC is modulated by attention for both video styles (Figure S1c, left), performing a three-way repeated measures ANOVA with subjects as a random effect and attentional state and video style as fixed effects. We find a significant main effect of attentional state (F(1,260)=638.60, p=5.63$\cdot10^{-72}$)) but no significant main effect of video style (F(1,260)=0.43, p=0.51).

These finding are replicated in Experiment 2 using a different learning style on a new cohort (Figure S1c, middle). Here we perform the same two-way repeated measures ANOVA test and find a main effect of attentional state (F(1,289)=875.74, p=1.83$\cdot10^{-89}$)) and no significant effect of video style (F(1,289)=0.38, p=0.54). This indicates that with this small sample we do not see any effect of video style on the ability of ISC to discriminate between attentional states.

In Experiment 3 we extended our analysis by including 2 additional video styles, namely 'Presenter & Animation' and 'Presenter & Glass Board' (Figure S1c, right). Despite these very different video styles we again find a robust discrimination of attentional state with the eye movement ISC (F(1,287)=1365.79, p=4.03$\cdot10^{-111}$)). However, in this case we do find a main effect of video style (F(2,287)=20.57, p=4.47$\cdot10^{-9}$)). We attribute this to the general dynamics of the video, where some have high spatial dynamics whereas others are more static eliciting less eye movements. Despite these differences, regardless of video style or learning scenario we find a robust discrimination between attention conditions replicated on three different cohorts.

Intersubject correlation of eye movements predicts test scores

In the main text we report a significant correlation between test score and ISC and show the results for Experiment 1 (in Figure 2b). Here we show the same results for Experiments 2 and 3 (Figure S2a and S2b respectively).

In the main text we analyzed video style in Experiment 3. Here we report similar results for Experiment 1 (Figure S2c, left) for video styles 'Animation' (r=0.67, p=1.1$\cdot10^{-4}$), N=27) and 'Writing hand & Animation' (r=0.62 ,p=5.4$\cdot10^{-4}$), N=27). We reproduce these findings in Experiment 2 using a new cohort and learning scenario (Figure S2c, right). We find a significant correlation for 'Animation' (r=0.40, p=0.03, N=30) and 'Writing hand & Animation' (r=0.50, p=5.0$\cdot10^{-3}$), N=30).

Modeling attention as common cause and measurement noise

The test scores were determined with a short quiz of only 11-12 questions. This makes the scores inherently noisy. Noise in the eye tracking data also seems to have affected the accuracy of the ISC metric. To determine how these noise sources, limit our ability to predict test scores we formulated a probabilistic model (Supplement). This model aims to match observed distribution of test scores and ISC (scatter plots in Fig. 2) and 3). The simplest model captures the correlation between scores and ISC assuming they are normally distributed (Fig. S4a, Gaussian model). We also build a model based on our hypothesis that attention causally affects eye movements as well as test scores (Fig. S4b, Attention model). We fit both models using maximum likelihood optimization (see Supplement for details) and find that the causal attention model is significantly more likely than the simple correlation model describing the data (Fig. S4c). The parameters of the attention model are consistent with independent empirical observations, such as the baseline performance of naïve subjects (Fig. S4d) or the noise estimates of the eye movements (Fig. S4e). According to the model the differing performance in predicting test scores is well explained by a change in signal-to-noise ratio (SNR, Fig. S4f). SNR is an estimate of the variance in

Experiment	Stimuli	Subjects	wISC	Score (%)	wISC vs. Score	Prediction error (% score*)
4: Classroom	Stars	N=82, Female 18, Age 19.7 ± 2.7	0.33 ± 0.20	62.14 ± 26.68	r=0.46 (p=1.8·10^{-5})	19.6 ± 19.4 (p=0.18)
4: Classroom	Immune	N=53, Female 13, Age 19.2 ± 1.2	0.33 ± 0.15	59.62 ± 22.35	r=0.55 (p=2.7·10^{-5})	12.5 ± 16.6 (p=0.33)
4: Classroom	Birth rate	N=57, Female 16, Age 19.3 ± 1.2	0.42 ± 0.22	72.17 ± 24.99	r=0.69 (p=5.3·10^{-9})	13.2 ± 15.0 (p=0.02)
4: Classroom	Related	N=71, Female 16, Age 19.5 ± 2.8	0.27 ± 0.17	46.35 ± 18.06	r=0.28 p=0.02	15.0 ± 16.7 (p=4.8·10^{-4})
4: Classroom	Enzyme	N=60, Female 17, Age 19.2 ± 1.1	0.38 ± 0.25	43.24 ± 21.92	r=0.33 p=0.01	15.4 ± 18.2 (p=0.43)
5: At-home	Stars	N=203, Female 86, Age 29.8 ± 8.8	0.35 ± 0.20	58.83 ± 24.73	r=0.55 (p=3.6·10^{-17})	14.8 ± 18.0 (p=2.6·10^{-4})
5: At-home	Immune	N=201, Female 87, Age 27.4 ± 8.3	0.28 ± 0.18	45.61 ± 21.89	r=0.53 (p=4.7·10^{-16})	12.8 ± 15.6 (p=7.4·10^{-3})
5: At-home	Birth rate	N=209, Female 88, Age 27.2 ± 7.8	0.41 ± 0.24	65.19 ± 26.71	r=0.51 (p=2.3·10^{-15})	17.5 ± 18.0 (p=3.7·10^{-3})
5: At-home	Related	N=195, Female 85, Age 27.9 ± 8.2	0.30 ± 0.19	44.57 ± 19.29	r=0.38 (p=3.9·10^{-8})	14.6 ± 14.7 (p=0.15)
5: At-home	Enzyme	N=204, Female 97, Age 28.1 ± 8.7	0.42 ± 0.25	37.72 ± 18.86	r=0.40 (p=4.0·10^{-9})	10.8 ± 15.8 (p=0.08)

Table S1: **Details of Figure 3.** Experiment, stimuli, subjects, weighted Intersubject Correlation (wISC), score, correlation between wISC and score, leave-one-out cross validated Median Absolute Deviation (MAD). Mean ± SD are taken across subjects. *p-values indicate test for whether predicted score is significantly different than naive median score prediction using sign test. z-score is given where this test is approximate.

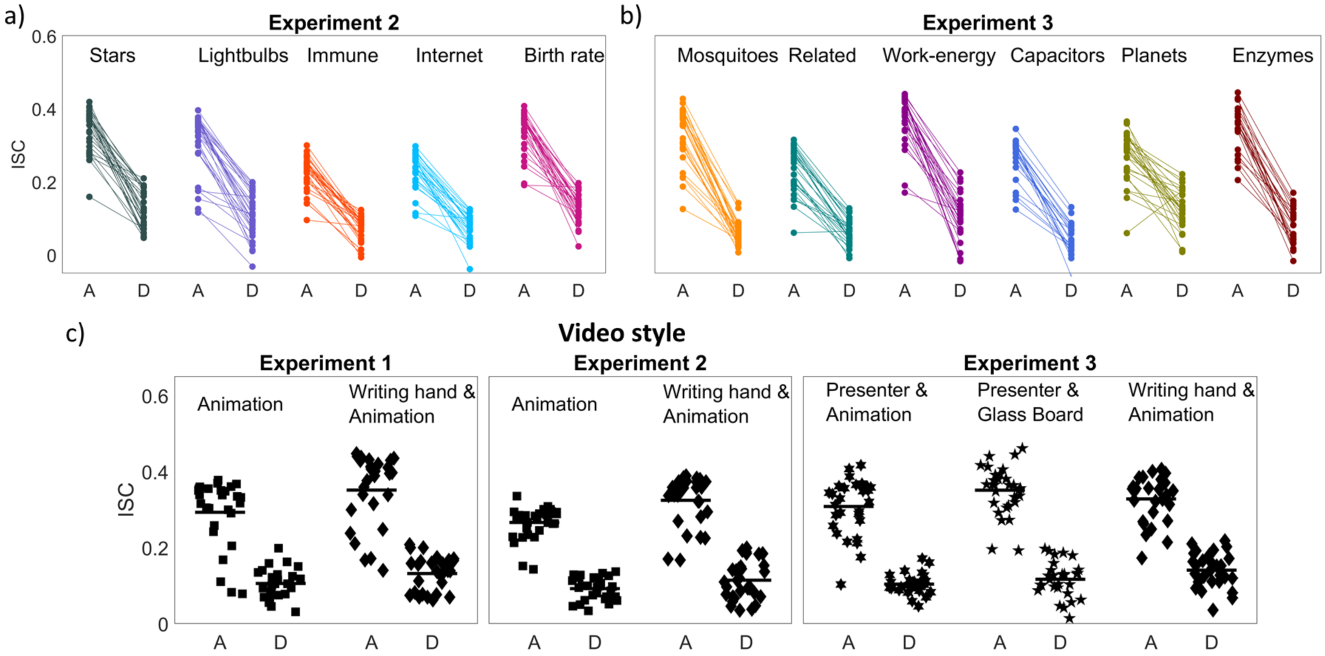

Figure S1: **Intersubject correlation of eye movements and pupil size is modulated by attention when watching educational videos. a-b)** The intersubject correlation (ISC) values for each subject are shown as dots for all video tests in Experiment 2 (panel a and 3 (panel b). Each dot is connected with a line between two different conditions, namely, when subjects were either attending (A) or were distracted (D) while watching the video. **c)** The intersubject correlation (ISC) values for each subject averaged across videos of different styles for Experiments 1-3.

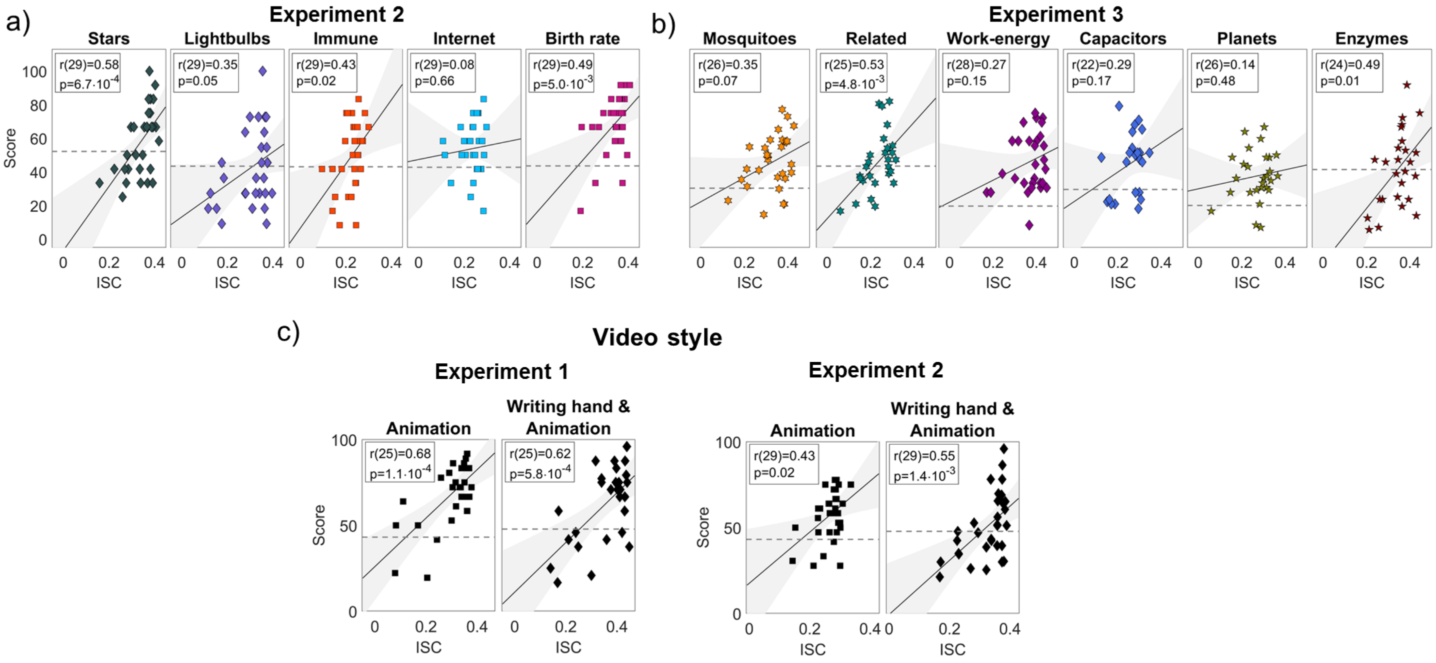

Figure S2: **Intersubject correlation of eye movements predict test scores. a)** The relation between intersubject correlation of eye movements and pupil size and performance on test taking (Score) for each of five videos of Experiment 2. Each dot is a subject. **b)** same as panel (a) but for the six videos in Experiment 3. **c)** same as panel (a) but averaging over the video produced in the 'Animation' and 'Writing hand & Animation' styles for Experiments 1 and 2.

attention over the variance in the noise of the attention measure (eye movement ISC). If SNR could be improved, then prediction performance could be substantially improved (Fig. S4f). However, with only 12 test questions there is a limit to this prediction performance as the test scores are inherently noisy. For instance, despite the relatively low eye tracking noise in the laboratory experiment (both measured and estimated; Fig. S4f) the model suggests that predictability is limited to r<0.7 (Fig. S4g), which is consistent with the empirical data (Fig 2b). Thus, in order to achieve better prediction of test scores one would need to increase the number of questions to obtain a more reliable assessment of student performance (Fig. S4g). In summary, a larger number of test questions, lower noise in our estimate of the attentional state of the subjects, and larger variance in attention across subjects are all expected to contribute to better prediction of test scores.

Probabilistic model of relationship between ISC and test scores

The data likelihood

The goal of this Supplement is to formulate a probability model for the observed data, namely, the test scores k_i and the correlation values c_i of eye movement. These are measured for subjects $i = 1 \dots N$. First we assume that these measures are independent across subjects, so that the data likelihood factorizes:

$$L(\theta) = p(k_1, \dots k_N, c_1 \dots c_N|\theta) = \prod_{k=1}^{N} p(k_i, c_i|\theta) \qquad (1)$$

Here $p(k, c|\theta)$ is the joint probability density of the data given parameters θ. Next we propose two generative models for this joint density. One is a straightforward bivariate Gaussian density that captures the correlation between the two variables. We refer to this as the "Gaussian model" (Figure S3a). The other will incorporate a common cause that leads to observations k and c through specific processes. We refer to that as the "Attention model" (Figure S3b).

Gaussian model

The canonical approach to describing the correlation between two variables as a probability density is the bivariate Gaussian density

$$p(k, c|\theta) = \frac{1}{2\pi|\Sigma|} \exp\left(-\frac{1}{2}([k, c] - \mu)\Sigma^{-1}([k, c] - \mu)^T\right) \qquad (2)$$

The maximum likelihood solution for the parameters $\theta = \{\mu, \Sigma\}$ can be found in any statistics text book. They are the sample mean and sample covariance:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}[k_i, c_i] \qquad (3)$$

$$\Sigma = \frac{1}{N-1}\sum_{i=1}^{N}([k_i, c_i] - \mu)([k_i, c_i] - \mu)^T \qquad (4)$$

Note that we have here 5 parameter that have been fit to the data: $\mu = [\mu_k, \mu_c]$, and $\Sigma = [\sigma_k^2, \sigma_{kc}; \sigma_{kc}, \sigma_c^2]$.

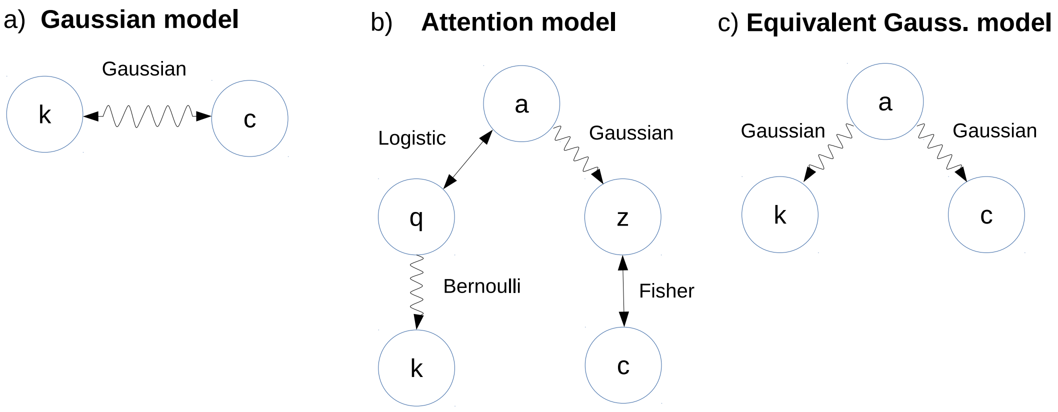

Figure S3: **Alternative models to explain test scores k and ISC values c.** Wavy lines indicates probabilistic relationship between variables. Straight line indicates deterministic relationship. Two-sided arrows indicate that the relationship is invertible. **a)** The two variables are related by a bivariate Gaussian distribution. **b)** Variable a indicates "attention", or more generally, an unobserved internal state of the subject that affects test scores and eye movements. Variable q captures the odds of answering a question correctly, i.e. questions to the subjects are Bernoulli trials with odds q, and the sum of correct answers is test scores k. Variable z is the Fisher transformed version of correlation value c. It is assumed that this variable captures attention, except additive Gaussian noise. **c)** In the Gassian model the correlation of k and c can be equivalently formulated as the result of a common driving factor a that is also Gaussian distributed.

Attention model

The second approach is an explicit formulation of our hypothesis that attention affects both test-taking performance and how similar eye movements are to the group of subjects. We assume that variables do not otherwise affect each other, i.e. eye movement don't directly affect test scores, and evidently test scores can not affect eye-movements that happened in the past. In this view, the joint density can be written as

$$p(k,c) = \int p(k,c|a)p(a)\,da = \int p(k|a)p(c|a)p(a)\,da. \quad (5)$$

Here variable a quantifies the level of attention, and it is distributed across subjects according to $p(a)$. We do not take the word "attention" here too literal. From a modeling point of view, this variable captures any internal state of the subject that affect performance as well as eye movements. This could include alertness, engagement, interest or fatigue. For instance, a low value for a could be due to fatigue or lack of interest in the material. As an internal state of the subject, a is an unobserved variable, so we integrate over it. We assume that once conditioned on attention a, the score k and eye correlation c become independent with $p(k|a)$ representing the probability that at a given attention level a students obtains a score of k in the subsequent exam, and $p(c|a)$ is the probability that at a given attention level the eye movements correlates with other subjects at a value c. We will now specify an analytic model for each of these terms.

We assume that attention a is normally distributed in our cohort with standard deviation σ_a and zero mean, where zero represents an average level of attention, while positive and negative values represent more or less than average attention.

$$p(a) = \frac{1}{\sqrt{2\pi}\sigma_a} \exp\left(-\frac{a^2}{2\sigma_a^2}\right) \quad (6)$$

The exam consists of a series of yes/no questions, lets say, n questions. Assume that a student has a chance q of getting each question right. The number of correct answers k is then Bernoulli distributed:

$$p(k|n,q) = \begin{pmatrix} n \\ k \end{pmatrix} q^k (1-q)^{n-k} \quad (7)$$

To link attention to performance we assume that high attention levels will give students a high chance of answering questions correctly (close to $q = 1$) and low attention will cause poor odds (close to $q = 0$). Denote with θ_a the level of attention at witch a subject reaches even odds of answering correctly ($q = 0.5$), and let β_a be the sensitivity of performance on changing levels of attention. Then we can represent odds of correctly answering a question as a function of attention as follows:

$$q(a) = \text{logistic}(\beta_a(a - \theta_a)) \quad (8)$$

where we used the logistic function to capture the transition from probability 0 to probability 1:

$$\text{logistic}(x) = \frac{\exp(x)}{1 + \exp(x)} \quad (9)$$

Now we turn our attention to the distribution of correlation values c. The density of correlation values is well characterized by a normal distribution after the Fisher z-transformation:

$$z = \text{atanh}(c) \quad (10)$$

We will therefore work with the density $p(z|a)$ instead of $p(c|a)$. The two can be related after appropriate scaling:

$$p(c|a) = \frac{1}{1 - c^2} p(z(c)|a) \quad (11)$$

Given our finding that attention strongly modulates the correlation values we assume that z is directly determined by attention. However, we have seen that there are different noise levels in measuring eye movements, and so we will assume that z carries an independent additive noise

$$z = a + n \tag{12}$$

with n normally distributed with mean μ_n and standard deviation σ_n. Given this normal noise, z given a is distributed with the following density

$$p(z|a) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(z-\mu_n-a)^2}{2\sigma_n^2}\right) \tag{13}$$

In total, the joint distribution of the model can be written as

$$p(k,c|\theta) = \frac{1}{1-c^2}\begin{pmatrix} n \\ k \end{pmatrix}\frac{1}{2\pi\sigma_n\sigma_a} \int da q(a)^k(1-q(a))^{n-k}\exp\left(-\frac{(z(c)-\mu_n-a)^2}{2\sigma_n^2} - \frac{a^2}{2\sigma_a^2}\right) \tag{14}$$

The parameters of this joint density are now $\theta = \{\sigma_a, \beta_a, \theta_a, \sigma_n, \mu_n, \}$. To estimate these 5 parameters we will again use maximum likelihood optimization.

For the purpose of finding the optimal parameters first recall that n and a are independent and that a is zero mean. Therefore the following constraints apply to the parameters:

$$\mu_z = \mu_n \tag{15}$$

$$\sigma_z^2 = \sigma_a^2 + \sigma_n^2 \tag{16}$$

Now, μ_z and σ_z^2 can be estimated directly from the sample data:

$$\mu_z = \frac{1}{N}\sum_{i=1}^{N} z_i \tag{17}$$

$$\sigma_z^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_i-\mu_z)^2 \tag{18}$$

With these two constraints determined by the data, we really only have 3 degrees of freedom remaining for optimization. Parameter μ_n is directly specified by μ_z (15). Parameters σ_a and σ_n^2 will be reduced to a single parameter, namely, the logarithm of the signal to noise ratio:

$$\lambda = \log\left(\frac{\sigma_a^2}{\sigma_n^2}\right) \tag{19}$$

by leveraging the constraint (16) the two variances parameterized with a single parameter λ:

$$\sigma_a^2 = \frac{\sigma_z^2}{1+e^{\lambda}} \tag{20}$$

$$\sigma_n^2 = \frac{\sigma_z^2}{1+e^{-\lambda}} \tag{21}$$

So in total we directly measure μ_n and optimize for the log-likelihood with respect to parameters $\{\lambda_a, \beta_a, \theta_a\}$. Optimization of three unconstrained parameters can be done fairly efficiently numerically.

Unfortunately the integral over a in (14) has no obvious solution and we thus evaluated it numerically during optimization. To do this with equal accuracy for arbitrary parameter values we perform the following variable substitution: $a' = a - \mu_n$ and write the integral as

$$\int da'\, q(a'-\mu_n)^k(1-q(a'-\mu_n))^{n-k}\exp\left(-\frac{(a'-\mu_{a'})^2}{2\sigma_{a'}^2}\right) \tag{22}$$

with

$$\sigma_{a'}^2 = \frac{\sigma_n^2\sigma_a^2}{\sigma_z^2}, \tag{23}$$

$$\mu_{a'} = \frac{z\sigma_a^2+\mu_n\sigma_n^2}{\sigma_z^2} \tag{24}$$

In practice the integral is executed as a sum by evenly dividing the range $a' \in [\mu_{a'} - 3\sigma_{a'}, \mu_{a'} + 3\sigma_{a'}]$ into discrete steps of $\Delta a' = 0.1\sigma_{a'}$.

Comparison between models

Here we want to make a few concluding remarks in comparing these two models.

Both the Gaussian model and the Attention model have 5 free parameters that are fit to the data (Figure S4a and S4b respectively). According to the Akaike Information Criterion one can choose the preferred model by comparing the maximum likelihood value obtained on the data. In fact, since both models have the same number of parameters no correction is needed for the degrees of freedom and one can directly evaluate the likelihood ratios. These ratios are reported in Figure S4c of the main paper for all experimental data. We find that the Attention model is significantly more likely in all cases (likelihood of Attention model over Gaussian model is larger than 1).

The parameters of the two models quantify similar properties of the distribution: μ_n and μ_c take on identical roles, and σ_c is captured by σ_a and σ_n; θ_a affects the mean score μ_k and β_a and σ_a affect σ_k; finally the larger σ_n (more noise), the smaller will be the correlation between k and c, as captured by σ_{kc}.

The variables of the Attention model, once fit to the empirically observed data (performance scores and eye movement ISC) seem to correctly capture empirical observations that were assessed independently of this data. For instance, the performance of naive subjects decreases with the estimated attention threshold θ_a (Figure S4d). This is expected as θ_a intends to capture the difficulty of questions (how much attention is required to answer half the questions right in average). Thus, videos with easy tests should have a low estimate for θ_a and videos with harder tests should have a high estimate. Additionally, the measured deviation from fixation dots increases with the estimated noise σ_n (Figure S4e). This is expected as σ_n capture the noise in the eye-movement ISC. Thus, inaccurate eye tracking data should lead to noise in the measured eye-movement ISC. Finally, the estimated signal-to-noise ratio σ_a/σ_n across videos follows the expected relationship with correlation (of score vs ISC) as predicted by the model (Figure S4f). While these estimates are not independent from the data used for the parameter fit, it does show internal consistency of the modeling approach across videos and supports

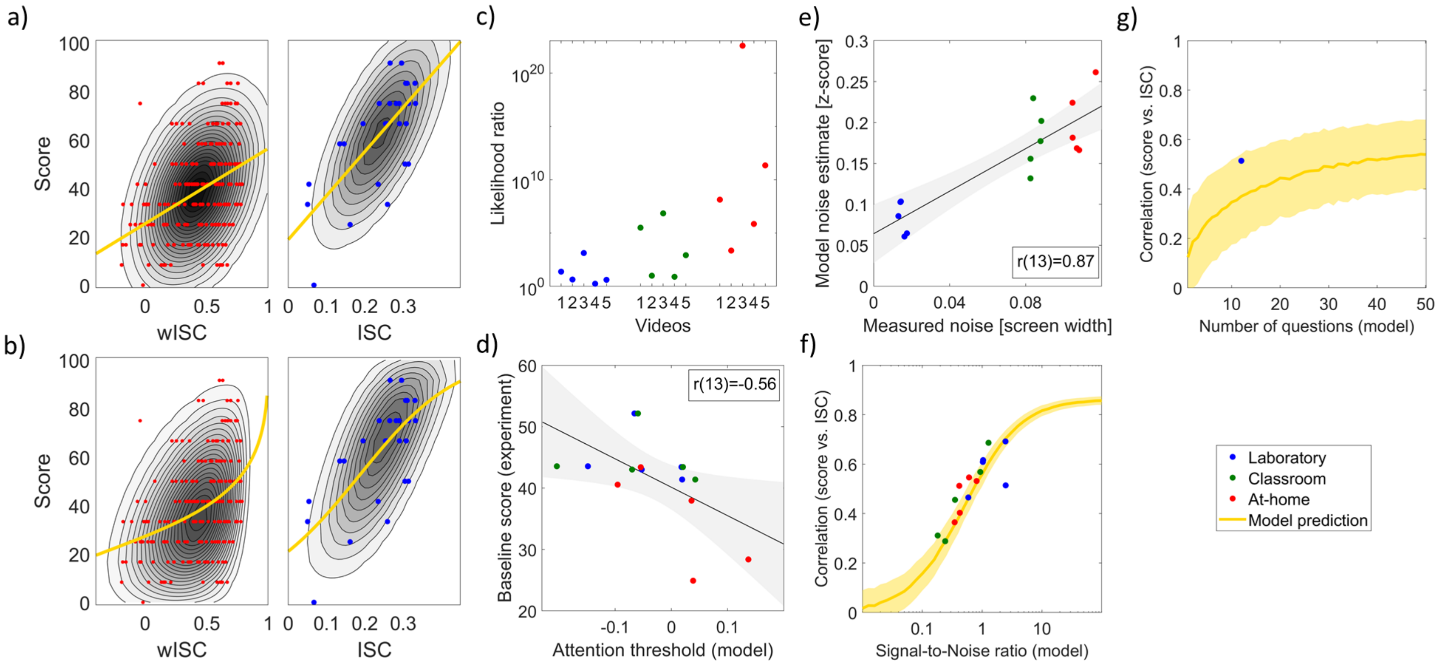

Figure S4: **Analytic model of test scores and ISC of eye movements: a)** Gaussian model fit to laboratory and at-home experiments for videos ‘Immune’ (blue dots) an ‘Birth rate’ (red dots) respectively . This model assumes a bivariate Gaussian distribution. Contour lines indicate the likelihood according to the model. **b)** Fit of the Attention model for the same data. This model assumes that score is Bernoulli distributed and ISC follows a Fisher z-score (see Supplementary Fig. S3). **c)** likelihood ratio between the Attention model and the Gaussian mode for all data from experiments in laboratory (Figure 2b), classroom (Figure 3c), and at-home (Figure 3d). Values larger than 1 indicate that attention model is a better fit. **d)** Performance of naive subjects, i.e. baseline test scores, compared to the estimated attention threshold in the model θ_a. This threshold parameter is the level of attention required to achieve 50% correct answers. **e)** comparison of estimated noise σ_n and measured noise (Deviation as in Figure 3b). **f)** correlation between eye-ISC and test scores as a function of estimated signal-to-noise ratio, σ_a^2/σ_n^2. The variability observed in the empirical data is consistent with changing SNR. **g)** correlation between ISC and test scores as a function of the number of questions for the laboratory experiment. Predictability of test scores increases with number of questions.

the argument that score/ISC correlations differ due to differing SNR levels.

The Gaussian model can be equivalently described in terms of a common cause (Figure S3c), i.e. the correlation between variables k and c is introduced by an unobserved common cause a. In the case of normally distributed variables the Gaussian model (Figure S3a) and the Equivalent model (Figure S3c) are mathematically indistinguishable. The Gaussian model (Figure S3a) can also be described as normal distributed variable k affecting c with some Gaussian noise, or vice versa. So one can not discern the direction of causality or a common cause for Gaussian data. For the Attention model the direction of influence can not be readily reversed. Thus, a better fit of the Attention model may be suggestive of a causal relationship via an unobserved common cause.

The Attention model takes the bounded nature of the variables k and c explicitly into account, whereas the Gaussian model does not. Thus, it is no surprise that the attention model has a higher likelihood on the empirical data as it does not allocate any probability mass outside the valid data range. In contrast, the Gaussian distribution has non-zero probability outside the valid range of the data. So a better fit may simply reflect that the data is more carefully modeled. Indeed, if the Bernoulli distribution is replaced by a Gaussian and the logistic and Fisher transformations by a linear transformation, then the attention model of Figure S3b reduced to the equivalent Gaussian model of Figure S3c. Ultimately the two models are not very different except for this features of a more careful model of the observed variables.

Gaze position data collected in at-home experiment

To provide a sense of the raw data here we display gaze position collected using web cameras in the at-home for one of the videos ('Birth rate') for all subjects (Figure S5). Each row is a subject and the intensity is the eye gaze position. Subjects are sorted by how well they did on the test that followed the video. It seems clear high-performing subjects have a stereotypical pattern of eye movements and this pattern disappears as performance drops.

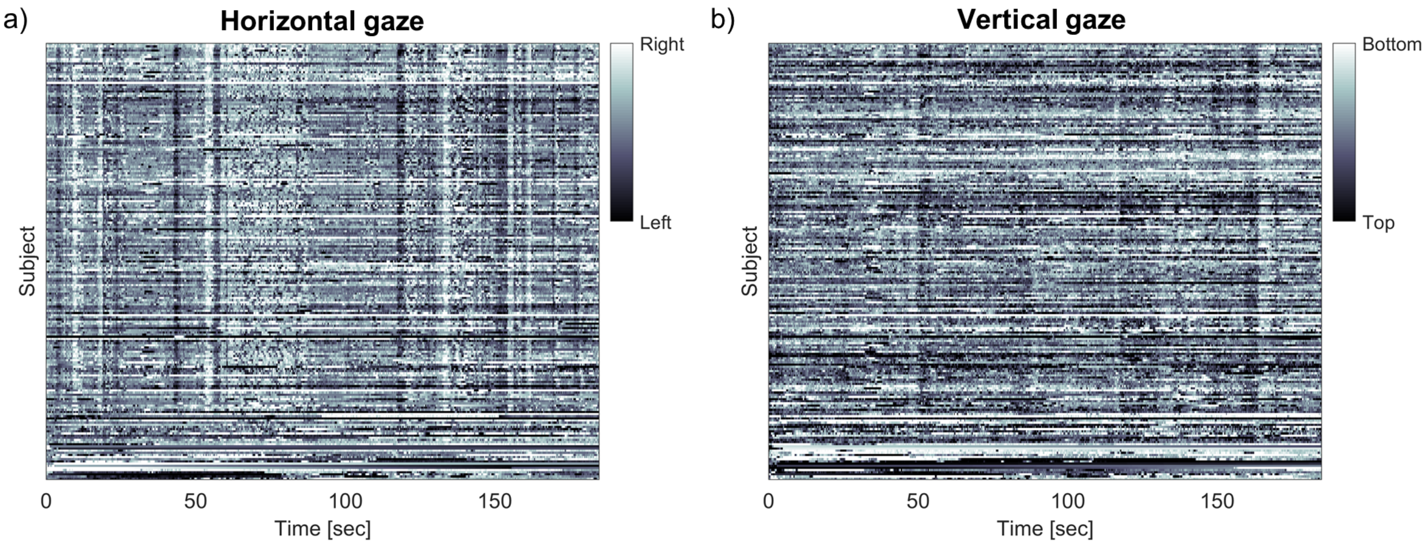

Figure S5: **Raw gaze position data:** Gaze position collected for ‘Stars’ in the at-home condition. Position is coded as brightness, and each subject is a row. Subject (N=203) are sorted by the score they obtained in the subsequent test (highest on top).

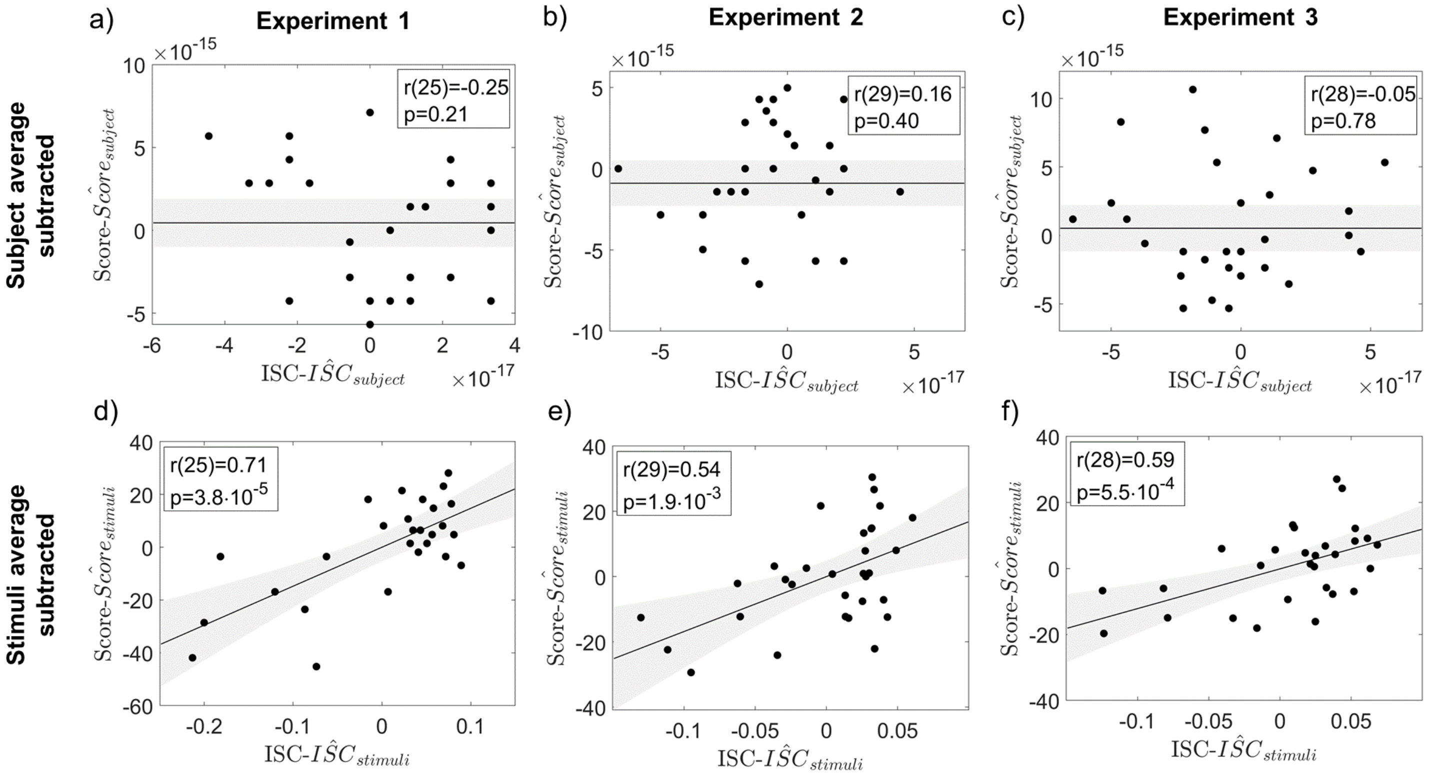

Figure S6: **Subject vs. stimuli effect: a-c)** ISC and Score with the average across subjects substracted from each (Experiment 1-3 respectively). **d-f)** ISC and Score with the average across stimuli substracted from each (Experiment 1-3 respectively).