A novel Quantitative Assessment of engagement in virtual reality: Task-unrelated thought is reduced compared to 2D videos.

Vishal Kiran Kuvar, Jeremy N. Bailenson, Caitlin Mills

PII: S0360-1315(23)00236-1
DOI: https://doi.org/10.1016/j.compedu.2023.104959
Reference: CAE 104959

To appear in: Computers & Education

Received Date: 3 March 2023
Revised Date: 25 October 2023
Accepted Date: 3 November 2023

Please cite this article as: Kuvar V.K., Bailenson J.N. & Mills C., A novel Quantitative Assessment of engagement in virtual reality: Task-unrelated thought is reduced compared to 2D videos., Computers & Education (2023), doi: https://doi.org/10.1016/j.compedu.2023.104959.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier Ltd.
Title:
A Novel Quantitative Assessment of Engagement in Virtual Reality: Task-unrelated Thought is Reduced Compared to 2D Videos.

Authors:
Vishal Kiran Kuvar\textsuperscript{a}, Jeremy N. Bailenson\textsuperscript{b}, Caitlin Mills\textsuperscript{a}

Affiliations:
\textsuperscript{a}Department of Educational Psychology, University of Minnesota, Minnesota, USA.
\textsuperscript{b}Department of Communication, Stanford University, California, USA.

Corresponding Author:
Vishal Kiran Kuvar, Room 193, 250 Education Sciences Bldg., 56 East River Road, Minneapolis, MN 55455. Kuvar001@umn.edu

Declaration of interest: No competing financial interests exist.
A Novel Quantitative Assessment of Engagement in Virtual Reality: Task-unrelated Thought is Reduced Compared to 2D Videos.

Abstract. Recent meta-analytic evidence suggests that students’ minds are likely to wander off-task frequently, regardless of the learning modality; yet virtual reality (VR) has been notably unexplored in this space. VR may present an opportunity to mitigate task-unrelated thought (TUT; the most common operationalization of mind wandering) because it minimizes audio-visual distractions and increases feelings of immersion. The current study tested this possibility by analyzing TUT frequency reports from 118 participants as they learned about climate change in one of two conditions: a 360° video in VR versus a traditional video on a 2D monitor. Participants answered momentary thought probes at pseudo-random intervals throughout the video and eye-gaze was recorded in both modalities. Results indicated that participants were less likely to experience TUT in the VR condition compared to non-VR (B = 0.49; p=0.02). Consistent with prior research, TUT was also negatively related to posttest performance (B=-0.05; p=0.01). Finally, TUT mediated the effect between learning modality on posttest performance, such that participants in VR experienced lower TUT and subsequently scored higher on the posttest (B = 0.19; p=0.03). We also present exploratory analyses on how gaze patterns differed across modalities as well as how gaze was related to instances of TUT.

Introduction

Imagine you are listening to an online lecture when you suddenly think of a conversation you had with a friend at dinner. You spend a few seconds considering whether you agree with her opinion, then your attention comes back to the lecture only to realize that you've missed some important information. This phenomenon, in which attention moves away from the task-at-hand, is referred to as task-unrelated thought (TUT; Smallwood & Schooler, 2015). TUT, the most common operationalization of mind wandering (Mills et al., 2018), has been a growing area of research over the last decade – especially in educational spaces (Kane et al., 2021; Mills et al., 2015; Smallwood, Fishman, et al., 2007; Wammes et al., 2019; Welhaf et al., 2022). Such interest in understanding and mitigating TUT during learning is likely due to its pervasive nature,
with students reporting being off task around 30% of the time on average (range 10-50%; Wong et al., 2022). Although TUT has some adaptative benefits, like the ability to mentally escape “the here and now” (e.g., autobiographical planning, creativity, etc.; Baird et al., 2011; Tan et al., 2015), recent meta-analytic evidence highlights a consistent negative relationship between TUT and learning outcomes (Bonifacci et al., 2022; D’Mello & Mills, 2021; Wong et al., 2022) regardless of the student’s learning modality (e.g., in-person class, online lectures, and learning with an intelligent tutor, to name a few).

Despite the purported ubiquity of TUT across learning modalities, virtual reality (VR) has been notably unexplored in this space (here we define VR as immersive virtual reality, which is typically viewed using head mounted displays). This gap is important to address given recent narratives about the future role of VR in education (Johnson et al., 2016; Ludlow, 2015; Schwarze et al., 2019). The increased affordability and availability of VR headsets (Hodgson et al., 2015), matched with a vested interest in effective and engaging remote learning options (Camilleri & Camilleri, 2021), sets up VR to become a popular learning medium in coming years. Related lines of work have primarily focused on external distractions outside of the VR environment (e.g., a cell phone ringing; Oh et al., 2019; Tao & Lopes, 2022), which are notably different from the concept of TUT. Such internally-directed thought can be triggered by or even involve current external content, but often tend to involve past- and future-oriented thought (Baird et al., 2011; Linz et al., 2021; Smallwood et al., 2009). Further, TUT does not require meta-awareness, as people are often unaware their thoughts have drifted off-task in the first place (Smallwood, McSpadden et al., 2007)—highlighting the ability to effectively leave the ‘here and now’ via our own thoughts, possibly even during an immersive VR experience.

One reason that VR may be appealing, particularly for mitigating students’ TUT, is that it creates an active and immersive perceptual experience—leading to increased feelings of presence (“a psychological state in which virtual objects are experienced as actual objects”; Lee, 2004). While completely immersed in VR, external audio-visual distractions are essentially “blocked out” (or at least reduced), an affordance that is not offered by other learning modalities. Previous work, for example, reported that the
majority of the TUTs during technology-assisted learning were about other technologies (e.g., social media; Hollis & Was, 2016); however, such TUTs may be less likely to arise while using VR, because of what is known as the place illusion (Slater, 2009), which makes the VR environment feel like real-life. Hence, we predict that TUT may therefore be less likely to arise in VR, at least when compared to watching traditional 2D videos (Hypothesis 1), which is currently the most popular form of technology-mediated learning (Ahmet et al., 2018; Fyfield et al., 2019; Hsin & Cigas, 2013).

This prediction has support from a few different angles. Based on Mayer’s cognitive theory of multimedia learning (R. E. Mayer, 2005; R. E. Mayer et al., 2023), individuals have limited cognitive capacity while they interact with computerized applications. Such cognitive capacity is thought to consist of three different forms of processing: essential (i.e., processing required information to help one’s understanding, often dependent on difficulty), generative (i.e., making connections with existing knowledge models, often dependent on motivation), and extraneous (i.e., processing of irrelevant stimuli, not supportive of learning objective). Individuals must effectively manage what type of processing they engage in for learning, and a failure to do so may leave them with an impoverished mental model – for example, focusing on extraneous information rather than essential. According to this theory, VR may increase extraneous processing due to the rich perceptual environment (Mayer et al., 2022). At the same time, motivation is proposed to be key factor that can help rebalance cognitive resources toward generative processing (Makransky & Mayer, 2022). Relevant to the current research, in both cases—whether extraneous or generative load increases during VR—both situations should effectively reduce TUT, as internal distractions should be mitigated due to increased resources being directed toward the features of the video. This line of reasoning also dovetails with perceptual load theory (Lavie, 2005): given that VR can be more demanding perceptually, with little-to-no external distractions outside the video, fewer resources should be available to process distractions. Indeed, increasing low levels of perceptual load reduced the frequency of TUT outside of VR (Faber et al., 2017; Forster & Lavie, 2009).

Finally, an affective/motivational standpoint provides similar logic as the cognitive theory perspective; there are increased feelings of presence in VR compared to
traditional monitors (Juan & Pérez, 2009; Krijn et al., 2004), and presence is positively related to interest (Parong & Mayer, 2018), and affect (Allcoat & von Mühlenen, 2018)—all of which are negatively related to TUT in prior work (Kane et al., 2021; Mills et al., 2021; Seli et al., 2015).

Although TUT has not been a major focus in VR research to date, we are certainly not the first to consider the usefulness of VR for learning. VR has typically been compared to other modalities to determine if it has any advantages with respect to learning (Barrett et al., 2022; Huang et al., 2019; Kavanagh et al., 2017; Zhao et al., 2021), and findings have been mixed. Although some studies found that VR helps learning (Ai-Lim Lee et al., 2010; Allcoat & von Mühlenen, 2018; Asad et al., 2021; Pande et al., 2021), other studies showed that it either had no significant improvement (Makransky et al., 2019; Queiroz et al., 2022; Stepan et al., 2017) or even hindered learning in some cases (Parong & Mayer, 2021b; Srivastava et al., 2019). Despite these mixed findings, recent meta-analytic evidence suggests that, on average, most studies find an improvement in learning outcomes when using VR (Di Natale et al., 2020; Wu et al., 2020). This might be true especially for memory recall tasks, where multiple studies have found higher rates of information retention in groups that use VR compared to regular monitors (Harman et al., 2017; Krokos et al., 2019; Mania et al., 2003).

These positive learning outcomes, however, are often nuanced such that they may be reliant on contextual factors within the learning environment rather than the use of VR itself. For example, learning in VR may be improved because of enhanced spatial representation afforded by VR (Dalgarno & Lee, 2010). Beyond spatial representation, feelings of presence and embodiment have also been shown to be factors that enhance learning in VR (Johnson-Glenberg et al., 2014, 2021). Other studies reporting positive learning outcomes seem to promote such presence through the use of realistic VR environments or by giving users agency through basic controls (Radianti et al., 2020). In contrast, when gestures that are used to perform actions in VR are incongruent with what happens in the VR environment (e.g., an object moving in VR environment in the opposite direction than the arm does in real life), it might not improve learning (Makransky et al., 2019). Hence, it might be the case that VR is selectively beneficial for learning when it is able to provide the factors of presence or agency.
In line with this argument, studies often report an indirect effect of VR on learning through feelings of presence (Makransky & Lilleholt, 2018; Makransky & Mayer, 2022; Moreno & Mayer, 2002). Specifically, presence (and the constructs associated with it, such as interest and motivation) effectively act as mediators between learning modality (VR) and learning outcomes. Here we predict that, if our first hypothesis is correct, TUT may play a similar mediating role given its theoretical link to presence (Hypothesis 2). This prediction also dovetails with previous work suggesting that external distractions (i.e. via cell phones) while learning in VR seem to impair recognition and recall (Oh et al., 2019), which may create similar “gaps” in our knowledge.

Overview of Current Study

Here we provide one of the first empirical studies on TUT during VR by comparing participants’ experiences while they watched a video on climate change in either an immersive VR environment (interaction via head-movement and controllers) or on a 2D interactive screen (i.e., interaction via mouse clicks on a traditional monitor). We address three key primary aims: First, how often do students experience TUT in the context of learning from VR, and is it reduced when compared to a similar 2D experience of learning? Second, what is the relationship between TUT and learning in the context of VR? Here, we are interested in exploring if the consistent negative relationship between TUT and learning outcomes replicates when using VR as a learning modality. And finally, does VR influence learning indirectly, similar to the effects of presence observed in previous studies (i.e., VR → less TUT → better learning outcomes)?

In addition to these two a priori aims, a separate set of exploratory research questions focused on understanding student behaviors—namely through eye-gaze—and how they varied across the two learning conditions. These exploratory questions were based on the idea that the immersive nature of VR may influence how students allocate their visual attention—providing a small “window into the mind” via gaze behaviors. We examined if (and how) classic gaze behaviors, such as fixations (periods with eyes are focused on one object), saccades (periods when eyes move from one object to another), etc., were different across the two learning different modalities—i.e.,
do students look longer or move their eyes around more in VR when compared to an interactive traditional 2D monitor? At the same time, gaze has also been a reliable indicator of TUT (Faber et al., 2020; Ishimaru et al., 2016; Lee et al., 2021; Vickers, 2009; Zhang et al., 2020), with a recent study extending this finding to visual attention in a classroom-based VR environment (Hasenbein et al., 2022). We therefore tested if this relationship is replicated in this study, which may be helpful for both theoretical considerations of TUT across contexts as well as future work aiming to build real-time detectors of such cognitive states (Bixler & D'Mello, 2016; Faber et al., 2018; Hutt et al., 2021) for improving learning.

Methods

Participants

Methods and procedures used in this study were approved by the IRB at [name removed for blind review]. Data was collected from 132 students who received course credit for participating in the study through the psychology subject pool. Demographic data for one of the participants was not recorded properly. Out of the remaining participants, 62% identified as female, 35% as male, and the rest as non-binary. 92.5% of the participants reported being White, 4.5% as Asian, 1.5% as Black, and 1.5% reported ‘other.’

Materials

Video and Modality

Participants were asked to learn about the effects of climate change and tourism from a video. We selected a 360-degree video titled Coral Compass (https://stanfordvr.com/coralcompass/; Queiroz et al., 2022) that details information about how coral reefs in Palau, an island nation in Oceania, were impacted due to tourist activity (see Figure 1b). The duration of the video was just over five minutes. This video was selected because it harnessed the spatial nature of VR. It was created by selecting clips from dozens of hours of footage that encouraged head movement (both pitch and yaw). Finally, it was reviewed by VR experts to ensure the preservation of its spatial qualities.
All participants saw the same video, but the modality they viewed it in differed based on the condition they were assigned to. This study had two conditions: VR and non-VR. In the VR condition, as the name implies, a participant wore a virtual reality headset (HP Reverb G2; resolution 2160x2160 pixels per eye; FOV 114 degrees; refresh rate 90Hz; see Figure 1a) to watch the video. Participants could look around in the video by moving their head. In the non-VR condition, the video was shown on a monitor (Tobii TX300). Participants could look around in this condition by dragging the screen with their mouse.

**Ecological momentary assessments**
Participants answered momentary ‘thought probes’ that assessed their experience of task-unrelated thought (TUT) throughout the duration of the video. This probing method is the gold-standard in mind wandering literature (Smallwood & Schooler, 2015; Weinstein, 2018) and has been shown to have good convergent and divergent validity across many studies in the last decade. Participants were shown the following probe – “on a scale of 7, please select a number that most appropriately reflects your attention on the current task right now. 1 being you are completely focused on task and 7 being you are not focused on the task at all.” Responses were recorded on a 7-point Likert scale, with 1 being completely on task and 7 as completely off-task. This probing procedure was performed every 55-65 seconds (Christoff et al., 2009), with a total of five probes throughout the duration of the video. Although a probe approximately every minute seems frequent, an increased time between probes might result in higher TUT reports, leading to missed reports of on-task thought (Seli et al., 2013). Additionally, probes weren’t administered after a fixed time to avoid their anticipation. The video would remain paused till the participant responded (see Figures 1c, 1d). These responses, along with their respective timestamps, were logged.

**Eye-tracking**
In both conditions, eye tracking data was collected. In the non-VR condition, the Tobii TX300 was used, which is an unobtrusive screen-based tracker, and allows for free head movement. This tracker sampled at the rate of 120Hz, with an accuracy of 0.5° -
1.5°. Participants watched the video on a 23-inch monitor, that was located two feet away, and had a resolution of 1920x1080 pixels. The HP Reverb G2 headset, with a resolution of 2160x2160 pixels per eye, was used in the VR condition. Eye trackers, which also sampled at 120 Hz, were built into the headset itself and required no additional hardware set up. The VR tracker recorded eyes with a sub degree accuracy. In both conditions, the x, y, and z coordinates of gaze were logged.

**Posttest**

Participants completed a test after the video ended to gauge fact-based knowledge retention. A total of 10 multiple choice questions were created to cover key points across the entire video. All questions were piloted on Prolific in advance of the study to avoid floor or ceiling effects (see Appendix A for the list of questions). Questions had four options, each one of them being equally plausible to keep guesses by elimination to the minimum. For example, one of the questions was “**coral reefs are home to what fraction of all marine organisms?**” While the correct answer was one-fourth, the remaining options were one-sixth, two-third, and one-third. The posttest was scored as a proportion of correct answers (i.e., summing all correct answers and dividing by 10, ranging from 0-1).
Figure 1. (a, top-left) Example of the VR headset used. (b, top-right) A frame of the video shown to the participants. In this scene, the narrator is contrasting what the view looks like with and without tourist activity. (c, bottom-left) Example of the probe that can be seen by the participant. (d, bottom-right) Example of the probe when a participant hovers the controller over one of the buttons. The buttons change color upon hover to reduce the chance of participant clicking a button accidentally.

**Procedure**

Consent forms were given to participants when they arrived to take part in the study. Participants were given the following instructions – “participating in this study will involve answering some questions about yourself, going through some educational material, and then answering simple questions afterward.” They were also asked to report if they were prone to motion sickness. This was done to prevent people who tend to be nauseous from using the virtual reality headset. Participants who reported being prone to motion sickness were assigned to the non-VR condition, while the rest were randomly assigned to either the VR or non-VR condition. After the informed consent was signed,
participants were asked to familiarize themselves with the probe they would see while watching the video, by having them go through the probe text itself as well as examples of TUT, to ensure they knew exactly what they were reporting and avoid confusion. Depending on the assigned condition, participants would be led to the appropriate room. The two rooms were identical in the layout, features, and noise levels.

**Non-VR**

In the non-VR condition, participants were asked to sit in front of a computer monitor with a built-in eye tracker. The seat was adjusted so that the participants’ eyes were visible to the tracker. A nine-point eye-tracking calibration was performed before the video started playing. The purpose of this calibration, which involves participants following a colored dot on the screen without moving their heads, is to control for the differences in shapes and geometry in people’s eyes. The video was 360° and participants were told that they could look around by dragging the screen with the mouse. Participants were also instructed to answer TUT probes using the mouse.

**VR**

The HP Reverb G2 was used for the VR condition. Before the video started, an experimenter made sure that the participant was comfortable with the way the headset rested against their eyes. Nine-point eye calibration was also administered before the video was played. Participants in the VR condition were provided with a controller that they used to register their probe responses.

**Posttest**

After the video was complete, participants answered the 10-item, multiple-choice posttest regarding the content of the video on a regular monitor, regardless of the condition. Participants were not informed of their final score. Finally, participants answered questions about demographic information. The entire procedure took 20 minutes, irrespective of the condition.

**Data processing**
A total of 64 participants were in the VR condition, and 68 were in the non-VR condition. Probe responses for five participants in the VR environment were not logged correctly, possibly due to a software error. Posttest quiz responses from one participant, who belonged to the non-VR condition, were not recorded. Gaze files from eight participants in the VR condition were corrupted, resulting in their removal from analysis. After removing these six participants with missing data, there were a total of 118 participants, with 63 in the non-VR condition, and 55 in VR.

**Extracting Gaze Measures**

**VR**
The HP Omniecept Reverb G2 was selected due to its ability to record eye movements. While recording gaze data in VR, it is crucial to consider head movement because it directly affects the gaze coordinates. To address this, HP provides a prebuilt component, named Glia, that automatically adjusts for head movement. Gaze points were recorded, at a sampling rate of 120 Hz, in three dimensions for the left and right eye. Additionally, the coordinates from both the eyes were combined to output another set of coordinates which informed about what the eyes looked at together. These nine values were saved, along with the timestamp, throughout the experiment. Individual files containing eye tracking data for the entirety of the experiment were created.

To extract measures from the raw gaze data, 15s windows were created right before the probes. The windowed approach is a tried and tested method in TUT detection (Kuvar et al., 2022; Mills et al., 2016). Duchowski’s velocity-based filter was applied to extract fixation points from the data contained in the window (Duchowski et al., 2002). However, this filter was originally based on an eye tracker with a sampling rate of 30 Hz. To match that frequency, we down sampled our data to 30Hz by only including every fourth point in our analysis. The filter returned which series of points were fixations, or times where the eyes remained relatively still. The velocity of eyes over a series of points would be calculated, and if the velocity fell below a certain threshold, that duration would be marked as fixation. Movement of the eyes from one location to the other, called saccades, were marked between consecutive fixations.
**Non-VR**

The Tobii TX300 monitor was used to capture eye gaze data from participants in the non-VR condition. Similar to the approach above, a 15s window was placed before the probe to extract fixation and saccade points. The sampling rate of the eye tracker was set to 120 Hz. However, to keep preprocessing consistent across conditions, down sampling to 30Hz was performed with the data obtained in the non-VR condition. The Duchowski algorithm was not used here since Tobii uses its velocity-based filter, called the Attention Filter, to mark fixations and saccades in the data points automatically (Olsen, 2012).

Even though we used two different eye tracking algorithms, both algorithms use the same underlying principle, as described by Salvucci and Goldberg (2000), for marking fixations. Both algorithms calculate velocity of the eyes (in degrees/sec) and if the velocity falls below a certain threshold, the points are marked as fixations, otherwise they are classified as saccades. The only difference between the two is that Duchowski includes a third dimension (the z-axis). The core formula to calculate velocity, even after inclusion of a third dimension, remains unaffected. In addition to this, window sizes, sampling rate, and velocity thresholds were kept the same across the two algorithms to ensure a fair comparison. We also compared the numbers from both algorithms to the typical values. A typical fixation usually lasts anywhere from 150 – 300 milliseconds (Albert & Tullis, 2023), which is what we see in both algorithms. This was also true for other gaze features. Thus, we are confident in claiming that the gaze measures obtained from the two algorithms are comparable.

**Extraction**

After this pre-processing step, we extracted four basic gaze-based measures in the predetermined window size (15s) that have been commonly used in previous studies focused on TUT (Hutt et al., 2016, 2017; Mills et al., 2016). Total number of fixations and saccades have a near perfect correlation between them, so we dropped the number of saccades measure to avoid redundancy. The other two measures were mean saccade duration and mean fixation duration in the 15s window. The measures of
number of fixations and mean fixation duration provide information about how often someone stopped to look at parts of a scene, and how long they tended to stay there. Mean saccade duration represents how long people spent between fixations, without necessarily focusing on anything in particular.

**Analytical Approach**

Linear mixed-effect models, implemented in lme4 (Bates et al., 2015), were used for all analyses where TUT was the dependent variable (e.g., main effect of condition) and participant as random effect in order to account for the within subject variability in TUT responses. For all other analyses, linear regression analyses were implemented. The lavaan package was used to test the indirect path between VR→TUT→learning (Rosseel, 2012).

**RESULTS**

Table 1 contains an overview of the descriptive statistics for all key variables. Overall, participants reported an average TUT of 3.20 (1.22). This indicates that students do in fact experience TUT in the context of VR, despite how engaging it is. This number is also consistent with prior work, as on-task reports are more slightly common than being off task (Mills et al., 2021; Sell et al., 2018; Wammes et al., 2019).

**Table 1. Descriptive Statistics of Key Variables**

<table>
<thead>
<tr>
<th>Measure</th>
<th>VR Mean (SD)</th>
<th>Non-VR Mean (SD)</th>
<th>Cohen’s d (VR - non VR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUT</td>
<td>2.93 (1.20)</td>
<td>3.45 (1.19)</td>
<td>-0.40</td>
</tr>
<tr>
<td>Posttest score</td>
<td>0.62 (0.16)</td>
<td>0.58 (0.16)</td>
<td>0.25</td>
</tr>
<tr>
<td>Number of fixations</td>
<td>28.49 (6.11)</td>
<td>41.51 (4.99)</td>
<td>-2.34</td>
</tr>
<tr>
<td>Mean fixation duration</td>
<td>0.14 (0.06)</td>
<td>0.28 (0.04)</td>
<td>-0.31</td>
</tr>
<tr>
<td>Mean saccade duration</td>
<td>0.33 (0.23)</td>
<td>0.07 (0.04)</td>
<td>1.60</td>
</tr>
</tbody>
</table>

**Frequency of TUT**
We first assessed whether there was a difference in the experience of TUT across the VR and non-VR conditions. In line with our predictions, a linear mixed effects regression revealed that participants experienced lower levels of TUT in VR (B = 0.49; SE = 0.21; CI = [0.08, 0.89], p=0.02). Participants in the VR condition tended to report TUT almost half a point lower on the TUT scale compared to their non-VR counterparts.

**TUT and Learning**

We constructed a linear regression to assess the relationship between Condition, TUT, and posttest score. The model included an interaction term (TUT × Condition), as well as their two corresponding main effects (Condition + TUT) in order to determine: a) whether Condition had a main effect on posttest scores; b) the relationship between TUT and posttest score; and c) if the relationship between TUT and posttest scores was moderated by learning modality. There was no evidence of a main effect of Condition (p > .10), and the interaction between condition and TUT was not significant (B=0.01; SE=0.02; p=0.60). However, we did find that TUT was negatively related to participants’ posttest scores. Specifically, we found a negative link between TUT and posttest score, similar to previous research (Wong et al., 2022), in both VR (B=-0.05; SE=0.02; p=0.01) and non-VR (B=-0.03; SE=0.02; p=0.04) conditions, as seen in Figure 3.
Given the relationship between TUT and learning, along with previous work suggesting that VR may operate through indirect effects (Makransky & Lilleholt, 2018; Makransky & Mayer, 2022), we tested whether TUT may be a possible mediator (VR $\rightarrow$ less TUT $\rightarrow$ better learning outcomes). Note that previous work has shown that the direct effect need not be significant to observe evidence of indirect effects (Faber et al., 2017; Zhao et al., 2010). A mediation analysis revealed that condition influenced posttest scores indirectly through TUT. As shown in Figure 2, the overall indirect effect was significant ($B = 0.19; CI = [0.02, 0.42], p=0.03$). Consistent with the analyses reported above (and for completeness), we also report the significant a path between learning modality and TUT, as well as the significant b path between TUT and posttest score. Simply put, people who were assigned to the VR condition tended to have lower rates of TUT, and thus scored higher on the quiz.
Gaze Behaviors

**Gaze Measures Across Modalities**

Our first set of exploratory analyses focused on how gaze behaviors differed across the two learning conditions. Similar to the analyses above that were subject to repeated measurements, we constructed mixed effects regressions; participant ID was included as random effect to account for individual differences in gaze. We found no statistical difference between VR and non-VR in terms of mean fixation duration ($p=0.11$), revealing that participants do not seem to be focusing on areas in video for different lengths of time. Number of fixations and mean saccade duration, however, were significantly different across conditions; participants in the non-VR condition had more fixations ($B = 12.82$; $SE = 1.03$; $CI = [10.80, 14.84]$; see figure 4a) and shorter saccade durations compared to VR ($B = -0.26$; $SE = 0.03$; $CI = [-0.32, -0.20]$; see figure 4b).
Figure 4. (a, left) Differences in number of fixations based on learning modality. (b, right) Differences in mean saccade duration according to learning modality.

**Gaze Measures in Relation TUT**

Our second exploratory question was how gaze patterns relate to TUT, given their past links in other contexts. Mixed effects regressions were constructed for the key gaze metrics, where participant was included as a random effect. The fixed effects included: TUT × Condition interaction term, as well as Condition and TUT individually. None of the interactions were significant, \( p > 0.05 \). Mean saccade duration was the only measure that was significant linked to TUT (\( B = -0.46; \ SE = 0.20; \ CI = [-0.84, -0.07] \); see Figure 5), suggesting that participants experienced lower levels of TUT when their eyes moved around for longer periods of time. All other measures were not significantly related, \( p > 0.1 \).

![Graph showing regression analysis between TUT and mean saccade duration](image)

Figure 5. Regression analysis between TUT and mean saccade duration

**DISCUSSION AND CONCLUSION**

**Major Findings**

To our knowledge, this is the first study to compare TUT while learning from VR and non-VR modalities. Our results suggest that TUT is quite common during VR in general, and that people tended to experience TUT at lower rates while using VR compared to a traditional 2D video screen (\( d = .40 \)). This finding is in line with previous
studies that have found that increased presence, afforded by increased immersion in content, is linked with increased attention (Kim et al., 2021; Schuemie et al., 2001). There is also a potential argument to be made about VR increasing essential and generative load while learning (Sweller et al., 2011) due to high motivation levels (Makransky & Mayer, 2022), ultimately increasing the overall cognitive load (Parong & Mayer, 2021a; Siegel et al., 2021). An increased load could lead to lowered processing of extraneous stimuli (Lavie, 2005) and thus result in lower levels of TUT. However future work is needed to more formally test the mechanisms of why TUT is reduced. Despite the fact that TUT was reduced in VR, we replicated the consistent negative relationship between TUT and learning outcomes in VR (Wong et al., 2022). That is, even though VR may reduce the sheer occurrence of going off task, it does not assuage the potential consequences of it (D'Mello & Mills, 2021; Smallwood, Fishman, et al., 2007).

Several studies have provided evidence that using VR does not improve immediate knowledge retention (Babu et al., 2018; Ekstrand et al., 2018; Makransky et al., 2019). Our results also indicate no significant differences in posttest scores between the two conditions- VR and non-VR. We note that there may have been a few reasons for this null effect. First, we designed and estimated sample size for our primary research question surrounding TUT and thus may have been unable to detect smaller effects on the posttest. Second, previous work suggests that effects on learning can be dependent on the content of the lessons and how they depend on the 3D aspect of the video to convey information. Although we chose a spatially interactive 3D video with this in mind, it nevertheless may not be a strong effect when compared highly interactive videos that rely more heavily on the 3D aspect for learning.

However, we conceptually echo previous work suggesting that presence may be driving the benefits of VR (Makransky & Lilleholt, 2018; Moreno & Mayer, 2002). Similar to the role of presence, we found that TUT mediated the relation between learning modality and learning outcomes – highlighting the importance of considering attention and engagement in the context of VR. Future work should consider testing these effects in more nuanced conditions; for example, our video was relatively short, so
understanding these effects in a longer video would be worthwhile. Beyond length, including more videos that range in terms of topic is an important next step.

In addition to our a-priori hypotheses, we also explored the relationship between gaze and modality. Results from these analyses suggested that gaze patterns were different across modalities; participants in the VR condition had a lower number of fixations and a longer mean saccade duration compared to the non-VR condition. At the same time, higher rates of TUT were also linked to longer mean saccade durations. Taken together, an interesting pattern emerges – it is possible based on this set of findings that people were more likely to be “moving” around for longer periods with respect to their gaze in VR, which may have ultimately kept them engaged (i.e., lower TUT), perhaps through more immersion and agency to explore in 3D. However, this is purely speculative and not causal, particularly given the differences in the eye-gaze hardware; more research is needed along these lines.

Limitations and Future Work

This study has a few limitations that are worth noting. First, we did not administer a pretest which would have allowed us to measure learning gains for each person after watching the video. Additionally, the posttest did not include questions about spatial nature of the videos- a critical component in VR. Even though the primary focus of the study was TUT, inclusion of such questions might have helped us understand the impact of spatial salience on TUT incidence. Future studies could add a pretest as well as questions tapping into the spatial nature to analyze the net gain in knowledge as a function of TUT.

A second limitation of the current study is that novelty of the VR may have impacted the results. Specifically, if VR was new to participants, it could have led to them to rate their TUT lower. Hence, the TUT reports could be because of the “newness” of VR, and not VR itself. To that end, longer-term studies with repeated exposure to VR would be worthwhile to explore, or perhaps with students who are familiar with VR. Thirdly, this study was the homogeneity in the sample—our sample was predominantly female and Caucasian and was exclusively college students. Thus, we cannot be sure if our findings will generalize to a broader and more inclusive
population. A study with a more diverse population needs to be performed to understand how factors such as experience with technology, gender, and race influence these results.

Another major limitation of our exploratory eye-gaze analyses is the inherent differences across modalities, such as equipment, head movement, and the need to use different velocity-based algorithms. Differences in gaze measures could also have arisen because of different movements involved in looking around across the two modalities. For example, to look around on a 2D screen, a participant needed to move the screen with the mouse whereas a participant using VR only needed to move their head. Differences in the ease of viewing might have affected the results of gaze data. A novel design that controls for such variables may help facilitate a more direct comparison of gaze behaviors across modalities.

**Conclusion**

These limitations notwithstanding, we provide novel evidence that TUT may play an important role when learning from VR, highlighting the need to study TUT in this context. Understanding how attention shifts from the content to internally generated thoughts would give us a better insight into learning and could be the first step in designing systems to keep users better engaged.

One of the largest challenges with examining VR as a medium is coming up with metrics to understand presence, engagement, and other psychological constructs. Questionnaires often fail in this endeavor, as it is difficult for people to self-report a concept as esoteric as “presence” (Bailenson et al., 2004; Slater, 2004). In the current paper we describe an innovative behavioral method to operationalize engagement in VR which can generalize to contexts such as education, storytelling, training, and entertainment.

**REFERENCES**


https://doi.org/10.1016/j.learninstruc.2016.01.007


https://doi.org/10.1089/cyber.2018.0150


https://doi.org/10.1145/3411764.3445269

textbook: Eye gaze based attention extraction on materials for learning and instruction
in physics. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and

https://www.learntechlib.org/p/171478/

learning effects comparing 2D desktop to 3D virtual reality STEM experiences. *Journal of

embodied learning in mixed reality motion-capture environments: Two science studies.
*Journal of Educational Psychology*, 106(1), 86–104. https://doi.org/10.1037/a0034008

Acrophobic Environment Viewed via HMD or CAVE. *Presence: Teleoperators and Virtual

Kane, M. J., Carruth, N. P., Lurquin, J. H., Silvia, P. J., Smeekens, B. A., von Bastian, C. C., &
021-01156-3

*Virtual Reality in education*. 35.


https://doi.org/10.1007/978-1-4419-8126-4

https://doi.org/10.1080/10400419.2015.1088290

https://doi.org/10.1145/3526113.3545682

https://doi.org/10.1016/S0079-6123(09)01322-3

https://doi.org/10.1016/j.compedu.2018.12.007


Welhaf, M. S., Phillips, N. E., Smeekens, B. A., Miyake, A., & Kane, M. J. (2022). Interpolated testing and content pretesting as interventions to reduce task-unrelated thoughts


**Appendix A.**

List of questions answered by participants to measure knowledge retention. Correct answers are in bold.

1. Emissions of what are raising the temperature and acidity of our oceans
Carbon Monoxide

**Carbon Dioxide**

Methane

Nitrogen Dioxide

2. Palau is an island nation located where?

- **Western Pacific**
- Eastern Pacific
- Northern Atlantic
- Southern Atlantic

3. Coral Reefs are home to what fraction of all known Marine Organisms

- 1/6
- 2/3
- 1/4
- 1/3

4. About how many Palauans are there

- **About 20,000**
- About 12,000
- About 35,000
About 25,000

5. To address some of the negative effects of tourism, Palauan officials have introduced initiatives that will...

- Reduced the number of people at any given reef
- Increase the cost of excursions to coral reefs
- Ban tourists from going near the reefs
- Require permits for tourists who want to swim

6. Humans affect marine ________ even if they never enter the water

Organisms

Ecosystems

Biodiversity

Life cycle

7. What can good farming practices do?

- Diverts the water back to the ocean

Reduced the amount of sediment in the coastal oceans

- Produces compostable waste that can be dealt with without having to dump it in the ocean
- Increase food production, reducing the need for fishing
8. Sediment does what?

   **Reduces the clarity of the water**

   Makes water alkaline

   Lowers the temperature of the water

   Harms aquatic flora

9. Taro farms can reduce the amount of sediment in stormwater runoff by as much as...

   75%

   80%

   85%

   90%

10. Tourists damage the coral reef by...

    **Swimming vertically and kicking the reef**

    Plucking the reef to keep as souvenir

    Killing fish that keep the coral reef clean

    Blocking sunlight necessary for the coral reef to sustain
Highlights

- Task-unrelated thought (TUT) hasn’t been examined in Virtual Reality (VR) learning
- TUT could be reduced in VR due to reduction in external audio-visual distractions
- TUT was significantly lowered in the VR condition compared to the non-VR condition
- The consistent negative relation between TUT and posttest was replicated in VR
- Participants experienced reduced TUT in VR, and thus scored higher
Vishal Kiran Kuvar: Conceptualization, Methodology, Software, Formal analysis, Writing- original draft Jeremy N. Bailenson: Validation, Resources, Writing- review & editing, Visualization. Caitlin Mills: Conceptualization, Methodology, Validation, Writing- original draft, Writing- review & editing, Visualization.