UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Reading the criminal mind

Exploring novel methods of memory detection

Van der Cruyssen, I.A.

Publication date
2023

Document Version
Final published version

Link to publication

Citation for published version (APA):
Van der Cruyssen, I. A. (2023). *Reading the criminal mind: Exploring novel methods of memory detection*. [Thesis, fully internal, Universiteit van Amsterdam, Hebrew University of Jerusalem].

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:23 Jan 2026

READING THE CRIMINAL MIND

Exploring novel methods of memory detection

Ine Van der Cruyssen

Reading the criminal mind: Exploring novel methods of memory detection

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 29 september 2023, te 11.00 uur

door Ine Andrea Van der Cruyssen
geboren te Kortrijk

Promotiecommissie

Promotores:	dr. B.J. Verschuere	Universiteit van Amsterdam
	dr. Y. Pertzov	Hebrew University of Jerusalem
Copromotores:	prof. dr. G. Ben-Shakhar	Hebrew University of Jerusalem
Overige leden:	prof. dr. M. Kindt	Universiteit van Amsterdam
	dr. H. Larsen	Universiteit van Amsterdam
	dr. S. van Gaal	Universiteit van Amsterdam
	prof. dr. R. Frost	Hebrew University of Jerusalem
	prof. dr. M. Bar-Hillel	Hebrew University of Jerusalem

Faculteit der Maatschappij- en Gedragswetenschappen

Dit proefschrift is tot stand gekomen binnen een samenwerkingsverband tussen de Universiteit van Amsterdam en de Hebrew University of Jerusalem met als doel het behalen van een gezamenlijk doctoraat. Het proefschrift is voorbereid aan de Faculteit der Maatschappij- en Gedragswetenschappen van de Universiteit van Amsterdam en aan de Faculty of Social Sciences van de Hebrew University of Jerusalem.

This thesis was prepared within the partnership between the University of Amsterdam and the Hebrew University of Jerusalem with the purpose of obtaining a joint doctorate degree. The thesis was prepared in the Faculty of Social and Behavioural Sciences of the University of Amsterdam and in the Faculty of Social Sciences of the Hebrew University of Jerusalem.

The research for/publication of this doctoral thesis received financial assistance from the Barbara and Morton Mendel Foundation.

Table of contents

1

General introduction

CHAPTER 1

In October 2009, neuroscientist Adrian Owen received a phone call from his postdoctoral researcher Martin Monti. They had done it! They had managed to read Patient 23's mind. Five years earlier, the then 19-year-old Patient 23 had been in a terrible motorcycle accident. He had survived the crash, but had been in a vegetative state ever since. He no longer reacted to anything that happened around him until that particular night in October 2009. The researchers had put the patient in an fMRI scanner and asked him some questions. They had explained to him that when he wanted to answer "yes" he had to think about playing tennis (which would give a spike of activity in his premotor cortex) and when he wanted to answer "no" he had to think about walking around his house (which would activate his parahippocampal gyrus - an area in the temporal lobe). And it worked. After 5 years, the researchers were able to communicate with Patient 23 by reading his mind (Monti et al., 2010; Somers, 2021).

The idea of mind reading has always evoked a pervasive fascination, not only in the academic world but also in popular discussions. The idea has inspired many movies and TV series, including the most popular TV series of the past year, Stranger Things (IMDb[1], 2022). In this series, the main character Eleven could intentionally intrude into people's minds and relive their memories. Although most movies and TV series make this skill look futuristic, the above example of Patient 23 shows that these ideas might no longer be just science fiction. In addition to tapping into the mind of vegetative patients (Monti et al., 2010), it has been shown to be possible to discover what someone is dreaming while they are sleeping by measuring their nerve impulses to muscles with an electromyogram (Oldis, 2017), or to reconstruct movie trailers someone has seen from fMRI signals (Nishimoto et al., 2011). Mind reading has many opportunities and could not only be of interest for people who *can* not talk (e.g.,

[1] IMDb is the world's most popular and authoritative source for news about movies and TV series. Its annual rankings of the most popular movies and TV series is determined by the actual page views of the more than 200 million monthly visitors to IMDb.

because they are in a coma or sleeping), but also for people who do not *want* to talk – such as suspects of a crime who refuse to confess.

Reading the criminal mind

On the night of November 14, 1985, a seven-year-old girl and her eight-year-old brother went missing in New Jersey, United States. The next morning, they were both found dead in the basement of their house. The police immediately arrested Byron Halsey, who was helping to raise the children, and Clifton Hall, the neighbor of the family. Not long after the arrest, Byron Halsey was charged with murdering the two victims. Clifton Hall was released and was later called as an eyewitness in court. After spending 19 years in prison, DNA evidence revealed that not Byron Halsey, but neighbor Clifton Hall was guilty of committing the murders. In the meantime, Clifton Hall had committed three other crimes (Innocence Project, 2019).

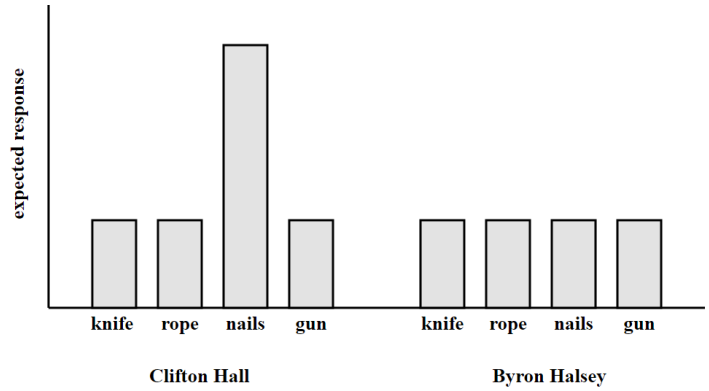
All that time, Clifton Hall of course knew what he had done. All the details of the crime had been stored in his memory. To examine whether a suspect of a crime has incriminating crime knowledge, even if they deny it, recognition of crime-related details can be detected using memory detection procedures. For example, if the investigators of the above case could have discovered that Clifton Hall was aware of several facts about the crime that only an actual perpetrator could know (e.g., the location where the bodies were found or the manner of death), while Byron Halsey did not possess any of this knowledge, a miscarriage of justice could have been avoided, and three crimes could possibly have been prevented.

Memory detection (also referred to as Guilty Knowledge Testing or Concealed Information Testing) covers a group of techniques with the same goal and methodology (Lykken, 1959, 1960; Verschuere et al., 2011). The goal is to detect through behavioral and physiological measures whether someone has critical crime knowledge. The typical

methodology involves the presentation of critical crime information within a series of similar, crime-unrelated information. If the examinee shows systematically stronger responses to the crime-related items than to the similar, crime-unrelated alternatives, concealed crime knowledge can be inferred. For example, if the investigators in the above example had tested the suspects' awareness of what the murder weapon was, responses to the real murder weapon (nails) could have been compared to several equally plausible murder weapons (e.g., knife, gun, rope…), see Figure 1.

Figure 1

Expected data pattern if a memory detection test would have been used in the Byron Halsey case



After decades of research, it is well-established that memory detection is a valid way for discovering "knowledgeable" individuals (Verschuere et al., 2011). A meta-analysis showed areas under the Receiver Operating Characteristic (ROC) curve well above chance level ranging from .735 to .848[2] (Meijer et al., 2014). Another selling feature of memory detection is that (in contrast to traditional lie detection methods such as the Control Question polygraph Test; Ben-Shakhar, 2002) it is not biased against the innocent (Lykken, 1988). While a guilty suspect responds differently to the crime-related item than to the control items,

[2] The area under the ROC curve is a measure of diagnostic accuracy of a test. It ranges from 0.5 to 1, where a value of 0.5 indicates a chance level accuracy and suggests no discrimination, while a value of 1 reflects a perfectly accurate test (Mandrekar, 2010).

an innocent suspect will not recognize the crime-related item and will therefore respond similarly to all presented items. These findings show the potential of memory detection. As such, the test has become an essential aid in criminal investigations in Japan (Osugi, 2011).

Traditionally, the possession of concealed knowledge is detected based on autonomic nervous system responses: Skin conductance, heart rate, and respiratory responses (Verschuere et al., 2011). Recently, it has been shown that in addition to the traditional measures, also reaction times (Suchotzki et al., 2017; Verschuere et al., 2014), the P300 component (i.e., a positive event-related potential that occurs after a meaningful piece of information is recognized; Fabiani et al., 1986; Rosenfeld et al., 1991), and fMRI (Langleben et al., 2002) can be used to detect concealed information. With these more recent measurements, accuracy rates rivaling those of the traditional measures can be obtained (Meijer et al., 2016). In the last decade, there is also renewed interest in eye movement measures for detecting concealed knowledge (Lancry-Dayan et al., 2022), especially since eye movements occur rapidly, uncontrollably, and sometimes even in the absence of conscious awareness (Hannula et al., 2010).

Challenges of memory detection

Although memory detection has proven to be a powerful tool for detecting the recognition of crime-related information, it is not without its obstacles. Four main challenges of the forensic application of memory detection include the leakage problem, the risk of countermeasures, item selection difficulty, and the test's limited applicability.

Leakage

One of the main challenges of memory detection is the leakage problem (Bradley et al., 2011). The assumption that only guilty individuals have knowledge about a crime might not always be valid. Innocent suspects may inadvertently acquire information about a crime

through various means such as newspapers, social media, or during the interrogation itself. Since the test detects crime knowledge, but not how this knowledge was acquired, leakage significantly increases the risk of false positive results (Bradley et al., 2011). Although in real criminal investigations, an innocent suspect could explain how they acquired this crime knowledge, the leakage problem poses a great challenge for memory detection.

Countermeasures

Another challenge of memory detection is the potential use of countermeasures (Ben-Shakhar, 2011; Peth et al., 2016). Countermeasures refer to techniques guilty suspects can deliberately use to control their responses and avoid detection (i.e., make them look innocent when they are actually guilty). By using either physical countermeasures such as inflicting pain by biting their tongue or mental countermeasures such as focusing on exciting memories, an enhanced response during the control items could be elicited, which can lead to false negative results. A review by Ben-Shakhar (2011) revealed that it is possible and unfortunately even quite easy to enhance responses to control items and bypass memory detection tests.

Item selection

The selection of proper crime-relevant and control items is crucial for a successful implementation of memory detection. Nonetheless, this process can be difficult (Krapohl, 2011). Memory detection tests must use crime-relevant details that are present in the perpetrator's memory. However, deciding which of all possible details of a crime scene are salient to the culprits and remembered by them, can be a challenging task. Especially in complex crime scenes, it can be hard for the investigators to know what the examinee encoded and still remembers days or even weeks after the crime. Furthermore, in what form should the selected item ideally be presented? In a typical memory detection test, items are presented verbally. However, memory research suggests that pictures are better remembered

and more easily recognized than their verbal counterparts (Kirkpatrick, 1894).

Limited applicability

Finally, memory detection has a limited applicability as it can only be used in narrowly defined situations. The physiological and neural measures require a laboratory setup and an experienced examiner which can be challenging in settings like airports, border control, or police interviews. With technology constantly evolving (which was further sparked by the COVID-19 pandemic forcing everyone and everything to go online), the possibilities of online methodologies are increasingly being tested (Kees et al., 2017). This includes memory detection research, which may have some significant advantages. No more lab, experienced examiner, and dedicated and expensive hardware would be needed. Large-scale memory detection programs could be implemented at critical borders, where they could identify travelers who require additional questioning (Kleinberg et al., 2019).

The current dissertation

In the past decades, research on memory detection has expanded considerably, and researchers are constantly tinkering with the different methods to develop the best possible form of it. However, some important challenges remain that prevent the test from becoming widely used. In the current dissertation, I am going to take initial steps in examining some new methods that could resolve some of these challenges. To this end, the current PhD consists of four empirical lines of research.

In the second chapter, I addressed the question of whether items are better presented pictorially or verbally in memory detection. Memory research suggests that pictures are better remembered and more easily recognized and recalled than their verbal counterparts (Kirkpatrick, 1894). However, both in practice in Japan as well as in most empirical studies items are presented verbally. I examined which modality is optimal, and tested whether

encoding and testing modalities interact in an online memory detection test based on reaction times.

In the third chapter of this dissertation, I explored the promising new development of memory detection based on eye movements. Initial memory detection studies based on eye tracking produced modest detection efficiency rates. Recently, Lancry-Dayan et al. (2018) argued that these initial studies might have been sub-optimal because they used a serial presentation of their items. They introduced a novel approach, applying a combination of both parallel and serial item presentation with a working memory task. This new paradigm resulted in areas under the ROC curve ranging from 0.89 to 0.97 in detecting familiar faces. I adopted this promising new paradigm and tested it in a mock crime setting. Additionally, I examined whether this method might provide a solution for one of the biggest challenges of memory detection: The information leakage problem (e.g., Bradley et al., 2011). Specifically, I examined whether innocents who gained knowledge about the crime by reading about it can still be distinguished from actual guilty suspects.

In the fourth chapter, I examined whether a simple webcam could be used as an eye tracker. One of the major promises of memory detection based on eye movements lies in the possibility of conducting tests remotely and without a laboratory setup. Remote memory detection has several important advantages. It has been argued that it could bypass countermeasures when the examinee is unaware of being tested (Elaad, 2011) and that it has the potential to be used for rapid detection on a large scale (Kleinberg et al., 2019). As a first step in assessing the potential of webcam-based eye tracking in memory detection, I replicated three robust eye tracking studies (the cascade effect, the novelty preference, and the visual world paradigm) online using the participant's webcam as an eye tracker. Furthermore, to be able to compare lab to web, I conducted the visual world paradigm not only online using the participant’s webcam but also in the lab using a standard eye tracker.

Lastly, in the fifth chapter, I propose an alternative way to detect crime knowledge. I examined whether suspects' familiarity with a crime scene could be detected by bringing them back to it in virtual reality while changing several salient elements of it. I tested whether gaze behavior toward those modulated parts could reveal its recognition. This new method has various advantages as compared to classic memory detection, such as the fact that it could be more generally applicable as no control items are needed. The test could be presented remotely which could protect it against countermeasures. Lastly, the risk of information leakage could be lower because leakage of an entire crime scene including the modified elements is highly unlikely.

References

Ben-Shakhar, G. (2002). A critical review of the Control Questions Test (CQT). *Handbook of Polygraph Testing, 103-126.*

Ben-Shakhar, G. (2011). Countermeasures. *Memory detection: Theory and application of the Concealed Information Test*, 200-214. doi:10.1017/cbo9780511975196.012

Bradley, M. T., Barefoot, C. A., & Arsenault, A. M. (2011). Leakage of information to innocent suspects. *Memory detection: Theory and application of the Concealed Information Test*, 187-199. doi:10.1017/cbo9780511975196.011

Elaad, E. (2011). New and old covert measures in the Concealed Information Test. *Memory detection: Theory and application of the Concealed Information Test*, 114-127. doi:10.1017/cbo9780511975196.007

Fabiani, M., Karis, D., & Donchin, E. (1986). P300 and recall in an incidental memory paradigm. *Psychophysiology, 23*(3), 298-308. doi:10.1111/j.1469-8986.1986.tb00636.x

Hannula, D. E., Althoff, R. R., Warren, D. E., Riggs, L., Cohen, N. J., & Ryan, J. D. (2010). Worth a glance: using eye movements to investigate the cognitive neuroscience of memory. *Frontiers in Human Neuroscience, 4*, 166. doi:10.3389/fnhum.2010.00166

IMDb (2022, December). *Best of 2022: Top 100 TV series of 2022.* IMDb. https://www.imdb.com/best-of/top-100-tv-series-of-2022/ls566051808/

Innocence Project (2019, July 9). *Today Marks Twelve Years of Freedom for Byron Halsey.* Innocence Project. https://innocenceproject.org/cases/byron-halsey/

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141-155. doi:10.1080/00913367.2016.1269304

Kirkpatrick, E. A. (1894). An experimental study of memory. *Psychological Review, 1*(6),

602. doi:10.1037/h0068244

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Detecting Deceptive Intentions: Possibilities for Large-Scale Applications. *The Palgrave Handbook of Deceptive Communication*, 403-427. doi:10.1007/978-3-319-96334-1_21

Krapohl, D. J. (2011). Limitations of the Concealed Information Test in criminal cases. *Memory detection: Theory and application of the Concealed Information Test,* 151-170. doi:10.1017/cbo9780511975196.009

Lancry-Dayan, O. C., Ben-Shakhar, G., & Pertzov, Y. (2022). The promise of eye-tracking in the detection of concealed memories. *Trends in Cognitive Sciences.* doi: 10.1016/j.tics.2022.08.019

Lancry-Dayan, O. C., Nahari, T., Ben-Shakhar, G., & Pertzov, Y. (2018). Do you know him? Gaze dynamics toward familiar faces on a concealed information test. *Journal of Applied Research in Memory and Cognition, 7*(2), 291-302. doi:10.1016/j.jarmac.2018.01.011

Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., O'Brien, C. P., & Childress, A. R. (2002). Brain Activity during Simulated Deception: An Event-Related Functional Magnetic Resonance Study. *NeuroImage, 15*(3), 727–732. doi:10.1006/nimg.2001.1003

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*(6), 385. doi:10.1037/h0046060

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology, 44,* 258-262. doi:10.1037/h0044413

Lykken, D. T. (1988). Detection of guilty knowledge: A comment on Forman and McCauley. *Journal of Applied Psychology, 73*(2), 303-304. doi:10.1037/0021-9010.73.2.303

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment.

Journal of Thoracic Oncology, 5(9), 1315-1316. doi:10.1097/jto.0b013e3181ec173d

Meijer, E. H., klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta-analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51*(9), 879-904. doi:10.1111/psyp.12239

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology, 53*(5), 593-604. doi:10.1111/psyp.12609

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., Owen, A. M., & Laureys, S. (2010). Willful Modulation of Brain Activity in Disorders of Consciousness. *New England Journal of Medicine, 362*(7), 579–589. doi:10.1056/nejmoa0905370

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology, 21*(19), 1641–1646. doi:10.1016/j.cub.2011.08.031

Oldis, D. (2017). Animating Dreams and Future Dream Recording. In *IASD Conference Abstracts.*

Osugi, A. (2011). Daily application of the Concealed Information Test: Japan. *Memory Detection: Theory and application of the Concealed Information Test,* 253-275. doi:10.1017/cbo9780511975196.015

Peth, J., Suchotzki, K., & Gamer, M. (2016). Influence of countermeasures on the validity of the Concealed Information Test. *Psychophysiology, 53*(9), 1429-1440. doi:10.1111/psyp.12690

Rosenfeld, J. P., Angell, A., Johnson, M., & Qian, J. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average

waveforms. *Psychophysiology, 38,* 319-335. doi:10.1111/j.1469-8986.1991.tb02202.x

Somers, J. (2021, November 29). *The Science of Mind Reading.* The New Yorker. https://www.newyorker.com/magazine/2021/12/06/the-science-of-mind-reading

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143*(4), 428e453. doi:10.1037/bul0000087.

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (Eds.). (2011). *Memory detection: Theory and application of the Concealed Information Test.* Cambridge University Press. doi:10.1017/cbo9780511975196

Verschuere, B., Suchotzki, K., & Debey, E. (2014). Detecting deception through reaction times. *Detecting deception: Current challenges and cognitive approaches,* 269-291. doi:10.1002/9781118510001.ch12

2

Modality effects in memory detection

This chapter is published as:
Van der Cruyssen, I., Regnath, F., Ben-Shakhar, G., Pertzov, Y., & Verschuere, B. (2021). Is a Picture Worth a Thousand Words? Congruency Between Encoding and Testing Improves Detection of Concealed Memories. *Journal of Applied Research in Memory and Cognition, 10*(4), 667-676. doi:10.1016/j.jarmac.2021.03.001

Abstract

The current study addressed modality effects in a web-based Concealed Information Test (CIT) by asking participants to encode, and later conceal, crime-related details. Items were encoded and tested verbally or pictorially. A pilot (N = 73) and a preregistered study (N = 158) showed a robust interaction between encoding and testing modality: Items that were encoded and tested in the same modality were associated with better detection. Moreover, recognition of verbally encoded items could not be detected in a pictorial test. Our findings support the existence of a modality-congruency effect when subjects try to conceal their knowledge. In applied scenarios, the modality of test items should be matched to the modality in which crime-related details were encoded. Furthermore, a pictorial CIT might protect informed innocents if leakage happened verbally.

Pictures are better remembered and more easily recognized than their verbal counterparts. This so-called picture superiority effect is well-established and has been repeatedly demonstrated in the memory literature (e.g., Ensor et al., 2019; Kirkpatrick, 1894; Madigan, 2014). This phenomenon has been demonstrated in both recall and recognition tests (Bevan & Steger, 1971; Shepard, 1967) and is very robust. For example, in a study by Shepard (1967), participants studied over 500 words or pictures and were respectively 88% versus 97% accurate in a subsequent recognition test. Over the years, the picture superiority effect has received considerable attention from researchers, and the phenomenon has been applied in a range of areas requiring easy and accurate recall of information such as instructional design, advertising, and marketing communications (Childers & Houston, 1984; Sansgiry et al., 1997). However, despite the many demonstrations confirming the picture superiority effect, its underlying mechanism is still a subject of discussion. Theories explaining pictures' mnemonic advantage can be roughly divided into those attributing the effect to a perceptual processing advantage of pictures and those attributing it to a conceptual processing advantage of pictures (for a review, see Stenberg, 2006).

According to perceptual theories, the picture superiority effect can be attributed to a richer perceptual encoding of pictures (Stenberg, 2006). While words are constrained to letters, phonemes, and orthographic conventions, pictures have no such constraints or limits (Nelson et al., 1979). As a result, there is more representational variability from picture to picture than from word to word, leading to enhanced distinctiveness of pictures and easier recall or recognition. Paivio's (1971) dual-code theory, one of the earliest and most influential accounts of this effect, distinguishes between verbal and nonverbal (e.g., visual, auditory, haptic) memory pathways. According to this theory, words tend to be processed only via the verbal

pathway, whereas pictures can be encoded through both routes. That is, words only evoke a verbal code, but pictures can activate both the image and its verbal label (e.g., “dog”) and are therefore more easily retrieved and recalled.

Conceptual theories, on the other hand, claim that the processing of pictures involves greater semantic elaboration than the processing of words (Stenberg, 2006). Accordingly, the transfer-appropriate processing theory assumes that cognitive operations performed during encoding are reinstated during retrieval (Rugg et al., 2015). Since a typical memory test such as a recall or recognition test requires deliberate recollection of prior experience, they have a mostly conceptual nature, which invokes subjects’ semantic knowledge. As picture encoding also invokes greater semantic elaboration, pictures have an advantage at test as compared to words, resulting in an overall better performance for pictures in most recall and recognition tests (Weldon et al., 1989).

The effect of presentation modality is particularly interesting in the context of memory detection tests. For example, in a situation where one tries to detect whether a suspect has implicating crime knowledge, it is important to know how the test items should ideally be presented (i.e., verbally or pictorially) and whether this is moderated by the way the perpetrator originally encoded the crime items. This question is especially important because in memory detection, people are deliberately suppressing or hiding their memories instead of trying to retrieve them (as in classical research about the picture superiority effect). Although the picture superiority effect is a reliable and reproducible phenomenon, it has been shown to be constrained by some limiting conditions. The test circumstances, for example, have been shown capable of abolishing or even reversing the picture superiority effect (Weldon & Roediger, 1987).

The Concealed Information Test (CIT) is a promising technique used for memory detection when subjects try to conceal their knowledge. Concealed knowledge is detected by

comparing the examinees' responses to a crime-relevant item (e.g., the murder weapon) to multiple, equally plausible control options (e.g., “A revolver? A rope? A lead pipe? A candlestick?”). A guilty suspect is assumed to have specific crime-related knowledge and is therefore expected to respond differently to the crime-related item than to the control items. An innocent examinee, on the other hand, will not recognize the crime-related item and should therefore respond similarly to all presented items. After decades of research, it is well established that the CIT provides a valid tool for the detection of "knowledgeable" individuals (e.g., Meijer et al., 2014; Suchotzki et al., 2017) and is nowadays used in criminal investigations in Japan (Osugi, 2011).

A few studies have explored modality effects in the CIT (Ambach et al., 2010; Ben-Shakhar et al., 1996; Rosenfeld et al., 2015; Seymour & Kerlin, 2008). Both Ambach et al. (2010) and Seymour and Kerlin (2008) manipulated modality at test (but not at encoding) using autonomic and brain electrical measures and Reaction Times (RT) respectively. They both found no differences between verbal versus pictorial presentation. However, since these studies did not manipulate modality at encoding, they could not examine possible effects of encoding modality and congruency between encoding and testing modalities. Rosenfeld and colleagues (2015) applied the CIT with brain electrical measures and recommended that CIT examiners should always use pictorial presentations at testing, because they found that the use of pictorial presentation was always more effective, regardless of the encoding modality. However, these suggestions were based on data from only 11 participants. Lastly, Ben-Shakhar et al. (1996) used a modified version of the CIT based on electrodermal measures and found a complete generalization across modalities. Neither encoding modality nor test modality affected electrodermal responsivity in the

CIT.

The current study is the first preregistered, highly powered study to examine modality effects in memory detection. Participants encoded several verbal and pictorial crime-related details (e.g., the getaway vehicle "bike," an image of the murder weapon). Subsequently, they completed a web-based RT CIT that incorporated a verbal block and a pictorial block. Based on the dual-code theory, we anticipated a larger CIT effect (i.e., larger RTs to the crime-related than control items) for items that were encoded as images than for items that were encoded as words (H1). In line with the transfer-appropriate processing theory, we also expected that modality congruence between encoding and testing would induce a larger CIT effect than modality incongruence (H2).

Pilot

Method

This study was approved by the ethical committee of the Social and Behavioral Sciences faculty at the University of Amsterdam and registered as number 2020-CP-11832. The protocol was carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki. The pilot was not preregistered. All materials, data, and analytic scripts have been made publicly available and can be accessed at https://osf.io/84eas/.

Participants

Ninety-six participants took part in the experiment via the online crowdsourcing platform Prolific (https://www.prolific.co). Because of possible age-related differences in lying proficiency and executive control (Debey et al., 2015), only participants between 18 and 45 years old were eligible to participate in the study. They received £2.50 for their participation and had the chance to obtain a bonus of £1 if they would beat the lie detector

test (i.e., if they would have a dCIT < 0.2; see Kleinberg & Verschuere, 2015).

Out of these 96 participants, we excluded 4 participants because they terminated the experiment before it ended, 1 participant who did not recognize any of the crime-related items during the crime recognition check (see below), 1 participant who responded incorrectly to all items during the verbal test block, and another 17 participants because they had error rates above 50% for at least one stimulus type (i.e., target, crime-related, or control item). The final sample size contained 73 participants (70% male, 29% female, 1% other) with a mean age of 24 years (*SD* = 6 years, range 18-43 years). Participants' nationality was mainly Portuguese (27%), Polish (16%), or Italian (10%), and most participants indicated Portuguese (29%), Polish (16 %), or English (16%) as their native language.

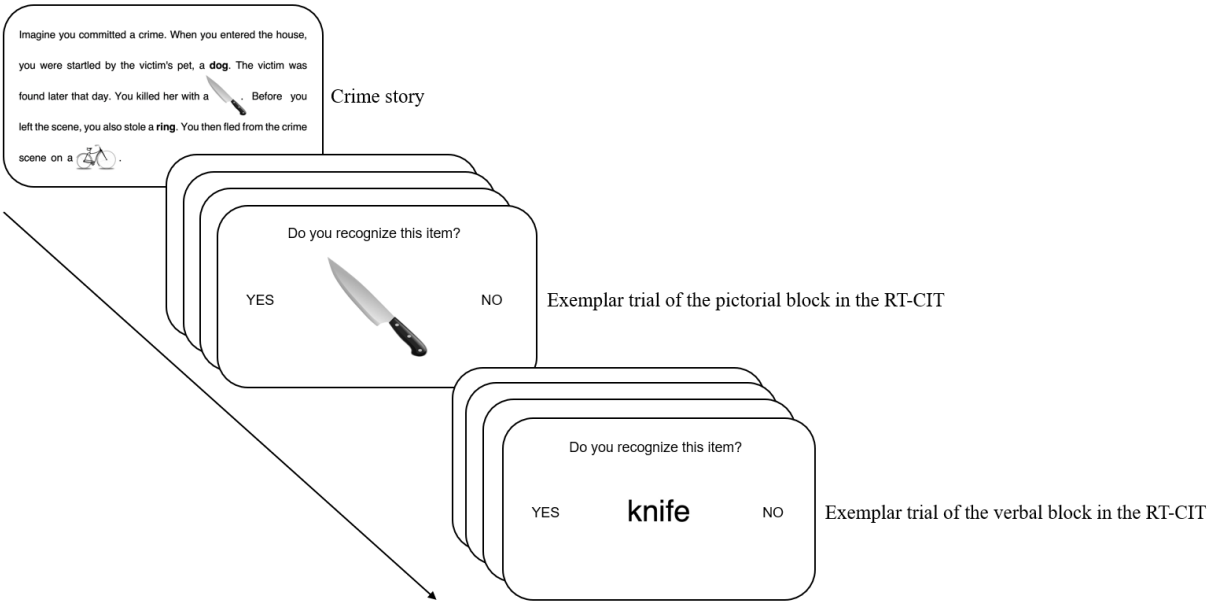
Procedure

All participants gave informed consent before taking part in the study. The task was computerized (programmed in Inquisit 5.0), completed online, and can be found at https://osf.io/sqv67/. In the first phase of the experiment (i.e., the encoding phase), participants were asked to read and imagine a crime story that contained two verbal and two pictorial crime-related items (see below). Subsequently, all participants completed a memory game to ensure that all crime-related items were sufficiently encoded. In this memory game, participants had to flip virtual cards containing the crime-related items from the crime story in order to find the four matching pairs. In the second phase of the experiment (i.e., the test phase), participants had to try to hide their crime knowledge in an RT-CIT task (see below). This task consisted of two blocks: A verbal block in which the items were presented as written words and a pictorial block in which the items were presented as pictures. An illustration of a typical encoding and test phase can be found in Figure 1. After

completing both blocks, participants were informed that the lie detection test had ended and were asked to respond honestly to all subsequent questions. To guarantee that participants still remembered the actual crime-related items, recognition was assessed by asking the participants to select each item they had seen in the crime story from a set of five items (the actual crime-related item and four control items) from the same category (e.g., all murder weapons), all presented in the modality in which the crime-related item was encoded. At the end of the experiment, participants provided demographic information about their age, gender, home country, and native language. The entire study was conducted in English.

Figure 1

Illustration of the experimental procedure consisting of an encoding phase (crime story) and a test phase (including a pictorial and a verbal block)



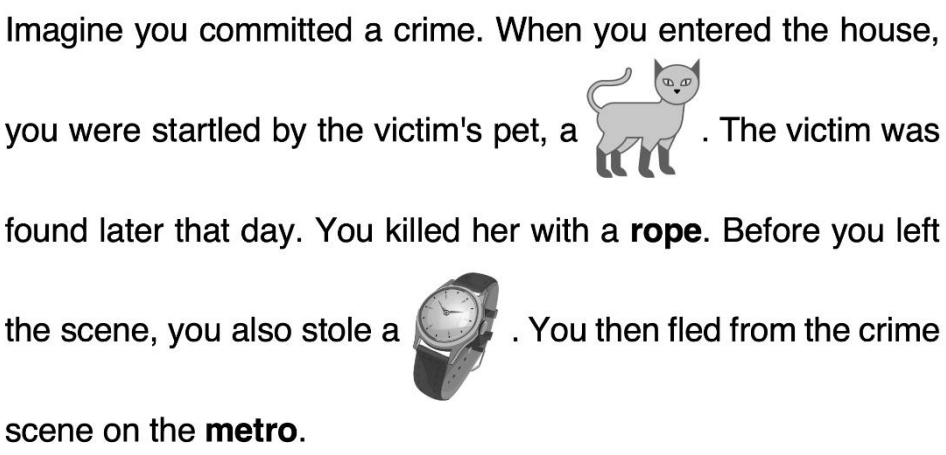
Materials

Crime Stories. To avoid stimulus-specific effects, we created four different profiles. These profiles determined which items were crime-related, target, or control items and which items were presented verbally or pictorially during the encoding phase for each participant (i.e., items that were crime-related/presented verbally for one participant were control

items/presented pictorially for another). Participants were counterbalanced over these four profiles and received their crime story accordingly. The four crime stories included information about four crime-related categories: Murder weapon, stolen item, pet of the victim, and mode of transport. Of these crime-related items, two were presented as written words and two as images (see Figure 2 for an example; the four crime stories can be found at https://osf.io/bwc5d/). Participants were asked to read the story carefully. Afterward, there was a 20 seconds time interval in which participants were prompted to recall the four crime-related items. Subsequently, participants read the crime story again and filled possible memory gaps.

Figure 2

Illustration of one of the four crime stories



RT-CIT. Before the beginning of the RT-CIT, participants were instructed to conceal their knowledge of the crime-related items presented in the crime story. We informed them that if they would succeed to do this, they would obtain a monetary bonus. Then, a new set of four target items from the same categories as the crime-related items (i.e., one murder weapon, one stolen item, one pet, and one mode of transport) was introduced. Participants were instructed to memorize these targets.

During the RT-CIT task, crime-related items and targets were presented pseudo-randomly (see below) within a series of control items. Participants were prompted with the question "Do you recognize this item?" and instructed to respond 'yes' to target items and 'no' to all other stimuli (i.e., crime-related items and control items) by pressing the corresponding buttons on their keyboard as quickly as possible (i.e., 'q' and 'p' respectively). This procedure introduces a so-called stimulus-response incompatibility: Participants have to override their automatically activated response to the crime-related items (i.e., report crime-related items as recognized) and initiate the correct response instead (i.e., treat crime-related items as unrecognized, control items), a process that is known to be time-consuming and error-prone (see Verschuere & De Houwer, 2011). However, note that labeling an item as (un)recognized might not be a crucial aspect in the RT-CIT (Lukács & Ansorge, 2019). The targets are used to ensure attention to the stimuli and ensure that participants would not answer 'no' to all items.

As in Kleinberg and Verschuere (2015), all items were presented for 1500ms on the screen with randomly varying inter-stimulus intervals of either 250ms, 500ms, or 750ms to reduce anticipation and prepared responding. If no button was pressed 800ms after an item was displayed, the words "TOO SLOW" appeared for 200ms above the stimulus. If the participant pressed the wrong key (e.g., indicating 'yes' for a crime-related item), the words "WRONG" appeared for 200ms below the displayed item.

The test part of the RT-CIT was divided into two blocks, a verbal block, and a pictorial block. Each block contained instructions, training, and testing in one modality. The order of the blocks (first the verbal or first the pictorial) was counterbalanced across participants. Each block started with a training phase to familiarize participants with the RT-CIT task. This training phase consisted of three stages of 12 trials that became increasingly similar to the testing phase in terms of response limitations and feedback (see Table 1). In

order to pass each stage of the training phase, participants had to meet certain accuracy and/or RT criteria. In stage 1 of the training phase, participants had to reach an overall target accuracy of 50%. In stage 2, we also required responses to be slower than 150ms on at least 20% of the trials to prevent participants from pressing the same button continuously. In stage 3, participants also needed to obtain an average RT that was faster than 800ms before they could move on to the testing phase. Participants had one re-attempt per stage if they failed to meet the criteria in their first attempt. After two unsuccessful attempts within the same training stage of the first modality section, the experiment was automatically terminated, since this indicated bad compliance with the task.

Table 1

Overview of Properties in the Training and Test Phase of the RT CIT

Stage	Trials	Stimulus duration	Accuracy feedback	Response time feedback
Training 1	12	Self-paced	"WRONG"	No feedback
Training 2	12	1500ms	"WRONG"	No feedback
Training 3	12	1500ms	"WRONG"	"TOO SLOW"
Test	480	1500ms	"WRONG"	"TOO SLOW"

After the training phase, the test trials started. Forty-eight unique stimuli were used (i.e., 24 written words and 24 images), of which eight were crime-related (i.e., four verbal, four pictorial), eight were targets (four verbal, four pictorial), and 32 were control items (16 verbal, 16 pictorial). Participants completed 240 trials in each of the two modality blocks. That is, all 24 items of the respective modality were presented in a random order and this process was repeated ten times. An overview of the test stimuli per crime story can be found at https://osf.io/p3xyf/.

General Overview of Data Analyses

In all data analyses, we only included trials with a correct response (i.e., 'no' for crime-related and control items) with RTs between 150ms and 800ms. We employed a within-subjects design: Each participant encoded verbal and pictorial stimuli and encountered both modalities in the test phase. That is, item modality (i.e., verbal versus pictorial) at encoding and at testing were the independent variables. The dependent variable was the standardized difference in RT (in milliseconds) between the crime-related and control items per participant (i.e., the CIT effect; $dCIT = [M_{RT(crime\text{-}related)} - M_{RT(control)}]/SD_{RT(control)}$; see also Kleinberg & Verschuere, 2015).

We tested our hypotheses by performing a 2 by 2 Bayesian repeated-measures (RM) ANOVA on the dCIT scores, with Encoding modality (verbal versus pictorial) and Test modality (verbal versus pictorial) as the two factors. In a (Bayesian) RM ANOVA, the same set of participants is tested multiple times under different conditions, after which is assessed whether there is a difference in means between the conditions. The Bayesian approach of an RM ANOVA compares the predictive performance of different models. The models under consideration are the null model, where there are no differences between the conditions (i.e., no effect), and the alternative models, where there are differences between the conditions (i.e., there is an effect). There are four alternative models in our analysis: The model with only a main effect of encoding, the model with only a main effect of testing, the model with two main effects (encoding and testing), and the model with two main effects (encoding and testing) and their interaction. For each model, a Bayes Factor (BF) is computed. This BF reflects how likely the data are under that model compared to the null model. BF_{01} shows evidence in favor of the null model and BF_{10} shows evidence in favor of an alternative model. Subsequently, we performed follow-up post hoc comparisons (Bayesian t-tests controlled for multiplicity). After comparing the different models, we compared the different effects. For

each effect (i.e., both main effects and the interaction effect), we computed a $BF_{inclusion}$ reflecting a comparison of all models containing a particular effect to those without the effect. In other words, $BF_{inclusion}$ can be interpreted as the evidence in the data for including an effect or interaction, similar to BF_{10} in the case of simple comparisons (see also van den Bergh et al., 2020). Lastly, we checked whether there was a CIT effect present in each modality condition. Therefore, we performed both a one-tailed paired samples t-test and a Bayesian one-tailed paired samples t-test on the crime-related-control item difference (in ms). For all analyses, we used JASP's default prior *r* scale of 0.5 for fixed effects in the Bayesian RM ANOVA and the default Cauchy scale of 0.707 for all follow-up Bayesian paired sample *t*-tests.

In the RT-CIT task, error rates tend to be less reliable and valid than RTs due to floor effects (i.e., often <5% errors; see Kleinberg & Verschuere, 2015). For the sake of completion, results for error rates can be found at https://osf.io/7gh89/ for the pilot and https://osf.io/ru8eb/ for study 1.

Results

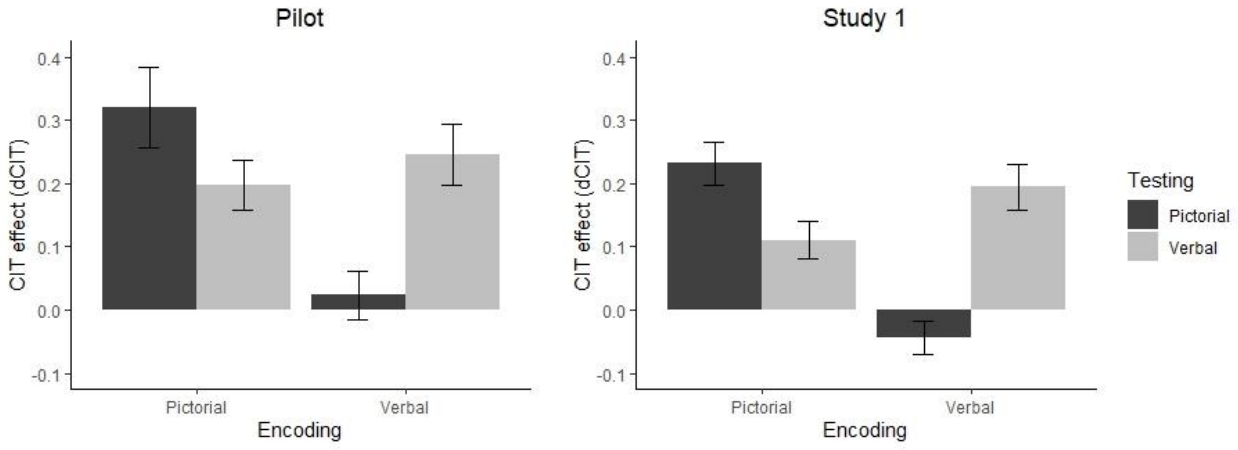
Modality Effects

Using a Bayesian RM ANOVA, the alternative model that included the two main effects Encoding and Testing and their interaction turned out to be the best performing model. The BF indicated that the data are 82.27 times more likely under that model than under the null model. The data showed substantial evidence for the main effect of Encoding ($BF_{inclusion}$ = 3.65), and also our post hoc comparison showed anecdotal evidence for a difference between verbally and pictorially encoded items (BF_{10} = 2.08), such that items encoded as pictures (M_{dCIT} = 0.26; SD_{dCIT} = 0.45) yielded larger dCIT scores than items encoded as words (M_{dCIT} = 0.13; SD_{dCIT} = 0.39). However, a Bayesian comparison between the Pictorial-Pictorial and the Verbal-

Verbal condition showed substantial evidence for the absence of a difference (BF_{01} = 4.83). We found decisive evidence for the presence of an interaction between Encoding and Testing ($BF_{inclusion}$ = 104.82). dCIT scores were larger when there was a match between encoding and test modality (M_{dCIT} = 0.28; SD_{dCIT} = 0.48) than when there was a mismatch in modalities (M_{dCIT} = 0.11; SD_{dCIT} = 0.34). Lastly, we found substantial evidence for the absence of an effect of Testing ($BF_{exclusion}$ = 4.55), which was also confirmed in our post hoc comparison (BF_{01} = 6.75). The results are summarized in Figure 3.

Figure 3

Standardized differences between response times to crime-related and control stimuli (dCIT scores) for verbal versus pictorial encoding and testing in the pilot and in study 1.



Note. Error bars represent standard errors.

CIT Effect for each Modality Combination

We examined whether RT-based CIT effects were present in each modality condition and which modality condition yielded the largest CIT effect (i.e., Pictorial-Pictorial, Verbal-Verbal, Verbal-Pictorial, or Pictorial-Verbal). Descriptive statistics (means and standard deviations of RTs of crime-related items and control items, the corresponding effect size, its 95% confidence interval (CI), and the corresponding BF) can be found in Table 2. Our results indicate the presence of a CIT effect in the Pictorial-Pictorial condition (M_{dCIT} = 0.32; SD_{dCIT}

= 0.54), the Verbal-Verbal condition (M_{dCIT} = 0.25; SD_{dCIT} = 0.41), and the Pictorial-Verbal condition (M_{dCIT} = 0.20; SD_{dCIT} = 0.33), but not in the Verbal-Pictorial condition (M_{dCIT} = 0.02; SD_{dCIT} = 0.33), suggesting that participants did not convert verbal information to pictorial information.

Table 2

Overview of Reaction Time-Based Crime-related-Control item Differences for each Modality Combination in the Pilot Study (N=73)

	RTs		Differences between crime-related and control items	
Condition	Crime-related	Control	*d* [95% CI]	BF_{10}
	M (*SD*)	*M (SD)*		
Pictorial – Pictorial	482 (59)	457 (37)	0.60 [0.38, ∞]	11 210.80
Verbal – Verbal	504 (65)	480 (49)	0.60 [0.39, ∞]	11 922,80
Pictorial – Verbal	499 (65)	480 (49)	0.59 [0.38, ∞]	8 802.46
Verbal – Pictorial	459 (49)	457 (37)	0.08 [-0.11, ∞]	4.17[a]
Across modalities	486 (62)	468 (45)	0.48 [0.37, ∞]	7.96×10^{11}

Note. Reaction times are given in milliseconds. Cohen's d expresses the size of the RT difference between crime-related items and control items. A Cohen's d of 0.20 can be interpreted as a small effect, a Cohen's d of 0.50 as a medium effect, and a Cohen's d of 0.80 as a large effect (Cohen, 1988). The BF_{10} expresses how much more likely the alternative model is (participants were slower for crime-related items than for control items) than the null model. [a]Reported number is BF_{01} which shows evidence for the null.

Study 1

The pilot showed that congruency in modality between encoding and testing was associated with better detection than incongruency and that it was easier to detect recognition of items that were encoded as pictures than items that were encoded as written words. Importantly, our results suggest that participants did not convert verbal information to pictorial information: Participants did not show recognition when items were tested as pictures but encoded as words. To test the robustness of these findings and to rule out the possibility that the pilot's results were due to the use of a mostly non-native English-speaking sample, we recruited a new sample that consisted of 50%

native English speakers and 50% non-native English speakers in study 1.

Method

The method of study 1 was identical to that of the pilot with the only exception that we now collected a sample comprised of 50% native English speakers and 50% non-English speakers. The study was preregistered (https://osf.io/vrkfq). All materials, data, and analytic scripts have been made publicly available and can be accessed at https://osf.io/84eas/.

Participants

As described in our preregistration, we first opened up the experiment to 100 participants on Prolific. For 50 of these, we required their native language to be English, for the other 50, we allowed every native language except English. Eligibility was restricted to participants between the ages of 18 and 45. Then, after applying our preregistered exclusion criteria (see below), we ran a 2 (encoding: verbal versus pictorial) by 2 (testing: verbal versus pictorial) Bayesian RM ANOVA on the dCIT scores in JASP. The decision to stop collecting data was based on the BF of the main effect of encoding and the BF of the interaction effect (by adding the main effects of encoding and testing as nuisance variables to the null model). We planned that once we would reach substantial evidence for either the alternative hypothesis (i.e., BF_{10} larger than 5) or the null hypothesis (i.e., BF_{10} smaller than 1/5) for both models (i.e., the main effect of encoding and the interaction between encoding and testing), we would stop testing, otherwise we would open up the experiment for another 100 participants (Schönbrodt et al., 2017). We intended to repeat this procedure until we would reach substantial evidence for both models for either the null or the alternative hypothesis or until we would have tested N = 500. We reached substantial evidence favoring the alternative hypotheses for both models after opening up the experiment two times for 100 participants. In total, we recruited 196 participants. Participants received £2 as compensation and an additional £1 bonus payment when they would beat the lie detector test (i.e., dCIT < 0.2; see

Kleinberg & Verschuere, 2015).

Based on our preregistered exclusion criteria, we excluded 6 participants because they terminated the experiment, 4 participants because they did not recognize any crime-related items by the end of the experiment, 2 participants due to high (>50%) error rates on at least one experimental test (i.e., verbal or pictorial, both participants had less than 50% accuracy on the verbal block), and another 26 participants due to high (>50%) error rates on at least one item type (i.e., target, crime-related, or control item; additionally, 1 participant who had less than 50% accuracy on the verbal block also had less than 50% accuracy for an item type). Our final sample consisted of 158 participants (49% men, 48% women, 2% other gender, and 1% preferred not to say) with a mean age of 26 years (*SD* = 7 years, range 18-45 years). Most participants indicated to be of British (32 %), Portuguese (13%), or Polish (13%) nationality, and most reported English (47%), Portuguese (13%), or Polish (13%) as their native language.

Results

Confirmatory Analyses

The BF indicated that the data were 2.98 x 10^7 times more likely under the alternative model that included both main effects of Encoding and Testing and their interaction than under the null model. The $BF_{inclusion}$ indicated strong evidence for an effect of Encoding ($BF_{inclusion}$ = 10.77) and the post hoc comparison indicated substantial evidence for a difference between verbal and pictorial encoding (BF_{10} = 8.01). Items that were encoded as pictures yielded a higher CIT effect (M_{dCIT} = 0.17; SD_{dCIT} = 0.40) than items that were encoded as words (M_{dCIT} = 0.08; SD_{dCIT} = 0.41). However, a Bayesian comparison between the Pictorial-Pictorial and the Verbal-Verbal condition showed substantial evidence for the absence of a difference (BF_{01} =

8.71). In addition, the analysis indicated decisive evidence for the model with the interaction effect between Encoding and Testing ($BF_{inclusion}$ = 5.27 × 10^6). That is, modality congruency between encoding and testing items (i.e., the Pictorial-Pictorial and Verbal-Verbal conditions) resulted in larger dCIT scores (M_{dCIT} = 0.21; SD_{dCIT} = 0.43) than incongruent modality combinations (i.e., the Pictorial-Verbal and Verbal-Pictorial conditions; M_{dCIT} = 0.03; SD_{dCIT} = 0.36). Lastly, we found anecdotal evidence for the absence of an effect of Testing ($BF_{exclusion}$ = 1.89), which was also confirmed in our post hoc comparison (BF_{01} = 2.23). Our findings are summarized in Figure 3.

Exploratory Analyses

CIT Effect for each Modality Combination. We also tested whether each item modality combination induced a CIT effect and which combination yielded the largest CIT effect. Descriptive statistics (means and standard deviations of RTs of crime-related items and control items, the corresponding effect size, its 95% CI, and the corresponding BF) can be found in Table 3. The results indicate the presence of a CIT effect in the Pictorial-Pictorial condition (M_{dCIT} = 0.23; SD_{dCIT} = 0.42), the Verbal-Verbal condition (M_{dCIT} = 0.19; SD_{dCIT} = 0.45), and the Pictorial-Verbal condition (M_{dCIT} = 0.11; SD_{dCIT} = 0.38), but not in the Verbal-Pictorial condition (M_{dCIT} = -0.04; SD_{dCIT} = 0.32), suggesting that participants did not convert verbal to pictorial information.

Table 3

Overview of Reaction Time-Based Crime-related-Control item Differences for each Modality Combination in Study 1 (N=158)

	RTs		Differences between crime-related and control items	
Condition	Crime-related	Control	*d* [95% CI]	BF_{10}
	M (SD)	*M (SD)*		
Pictorial – Pictorial	476 (56)	457 (42)	0.52 [0.38, ∞]	1.97×10^7
Verbal – Verbal	485 (66)	469 (46)	0.40 [0.27, ∞]	19 261.05
Pictorial – Verbal	478 (60)	469 (46)	0.26 [0.13, ∞]	28.71
Verbal – Pictorial	453 (52)	457 (42)	-0.14 [-0.27, ∞]	33.33[a]
Across modalities	473 (60)	463 (44)	0.28 [0.21, ∞]	8.53×10^8

Note. Reaction times are given in milliseconds. Cohen's d expresses the size of the RT difference between crime-related items and control items. The BF_{10} expresses how much more likely the alternative model is (participants were slower for crime-related items than control items) than the null model. [a] Reported number is BF_{01} which shows evidence for the null.

Modality Effects and Language. To examine possible language effects, we performed a 2 (Encoding: verbal versus pictorial, within-subjects) by 2 (Testing: verbal versus pictorial, within-subjects) by 2 (Language: native versus foreign, between-subjects) Bayesian RM ANOVA on the dCIT scores. We found substantial evidence for the absence of a three-way interaction between Encoding, Testing, and Language ($BF_{inclusion}$ = 4.55), suggesting native versus non-native language did not moderate the effect. A more detailed discussion of the results of this exploratory analysis can be found at https://osf.io/zwfx8/.

Discussion

The goal of this study was to examine modality effects in concealed memory detection. Our first aim was to investigate whether item modality at encoding would influence the CIT effect. Next, we examined whether the CIT effect is affected by (in)congruency between item modalities at encoding and testing. Our results were qualified by a robust interaction between encoding and test modality. That is,

congruent item modalities at encoding and testing yielded larger CIT effects, while incongruent modality combinations resulted in a comparatively lower (for Pictorial-Verbal) or even absent (for Verbal-Pictorial) CIT effect.

Perceptual Processing

Although the presence of the interaction limits its generalizability, items that were encoded pictorially were easier to detect than items that were encoded verbally, especially when testing was done using pictures. Moreover, no conversion across modalities was observed when items were encoded as words. These findings are in line with the dual code theory (Paivio, 1971), which suggests that pictures are processed via both the verbal and nonverbal memory pathways and are therefore stored as an image (e.g., picture of a dog) and its corresponding label (e.g., "dog"). In comparison, words only invoke a verbal code (e.g., "dog"). Therefore, items that were encoded as pictures were more accessible at testing, resulting in a larger CIT effect for encoded pictures than encoded words. Furthermore, if participants only store a word as its verbal code, but are then prompted with the corresponding picture, retrieval is hampered, resulting in the lack of a CIT effect in the Verbal-Pictorial condition. Similarly, the dual code theory also explains the CIT effect in the incongruent Pictorial-Verbal condition. Encoded pictures invoke both a visual and a verbal code: Even if participants are prompted with the corresponding word during testing, they can still retrieve the accompanying verbal code from memory.

Conceptual Processing

The finding that congruent item modality at encoding and testing resulted in a larger CIT effect than incongruent item modality is consistent with the transfer-appropriate processing account. This theory assumes that the cognitive processes performed during encoding are recapitulated at retrieval (see Rugg et al., 2015). Therefore, this theory predicts better memory performance when the type of processing is shared between encoding and

retrieval. In the context of the current experiment, congruent item modalities at encoding and testing (Pictorial-Pictorial or Verbal-Verbal) resulted in higher CIT effects. Previous research using the RT-CIT have found similar results: Items that were encoded on a categorical (e.g., car) or exemplar level (e.g., Citroën) incurred the largest CIT effect if the test modality was congruent (i.e., category-category or exemplar-exemplar; see Geven et al., 2019). Also in memory literature, it has been shown that similarities between study and testing are important determinants of memory performance (e.g., Stenberg et al., 1995; Weldon & Roediger, 1987). The current results are inconsistent with the results reported by Ben-Shakhar et al. (1996) that revealed a complete generalization of the skin conductance CIT effect from the verbal to the pictorial mode and vice versa. This inconsistency may be related to the use of very different dependent measures (skin conductance versus response time) which could be driven by different underlying mechanisms (e.g., klein Selle et al., 2016, 2017). Future studies will be needed to resolve this issue.

Practical Implications

Our results suggest that examiners should consider in which modality a guilty suspect likely encoded crime-related information and present items in a congruent modality during the test. This way, the detection efficiency of the RT-CIT could be maximized. Since it can be challenging to find out how a suspect encoded crime information, future research could examine the possibility of presenting pictorial and verbal items simultaneously at testing (e.g., a picture of the victim's dog with the word "dog" printed below or above the picture).

The current findings also bear significance to one of the biggest challenges in CIT: The leakage problem. The notion that only guilty suspects have knowledge about a crime is, unfortunately, not always the case. Through newspapers, social

media, or even during interrogation itself, innocent suspects may be exposed to critical details of the crime. Since the CIT only detects whether a suspect has knowledge of particular aspects of a crime, but not how this knowledge was acquired, leakage of information can induce an enhanced response to the crime-related items in informed innocent participants (Bradley et al., 2011). Based on our results, it could be assumed that if the perpetrator encoded crime-related aspects pictorially (i.e., visually) at the crime scene, while an innocent suspect only read or heard about the crime in a verbal form, then a pictorial CIT could protect the innocent suspect from being falsely accused. It can be argued that leakage is not always limited to a purely verbal form and encoding of a crime is not always limited to a pictorial form. For instance, leakage can also occur through pictures or videos of the crime, and certain crime-related information could be encoded verbally by the perpetrator (e.g., what was said in an extortion or threat, or the name of a city). However, since it is much more likely that leakage occurs verbally, and, arguably, almost all crime-related details are encoded pictorially, the promising lack of conversion from verbal to pictorial information found in the current study could possibly protect informed innocents in a pictorial CIT. We argue that the better the match between test and encoding, the better the CIT will be able to detect the guilty and protect the innocent. Related, in a study by Norman et al. (2020), crime scenes that were revisited in virtual reality resulted in even better concealed recognition than 2D images of these crime scenes.

Limitations

A possible limitation of the current study could be the fact that the stimuli were not split into crime-related, target, or control item in a fully randomized manner. We prevented stimulus-specific effects by preparing four profiles so that items that were crime-related for one participant were control items for another. However, it is still theoretically possible that the items that were crime-related in each profile were simply more salient or took more

processing time for some reason than the other items. Furthermore, as images have many varying properties, it is difficult to properly compare them to verbal equivalents in a manner that can be safely applied to other image types (e.g. from drawings in the current study to photographs or even virtual reality), limiting the generalizability of the reported results.

Conclusion

The current study demonstrated that successful detection of crime knowledge with the RT-CIT depends on the way items are encoded and tested. We found evidence for a robust effect of modality congruence. That is, our results suggest that the CIT effect can be maximized if the testing modality is congruent to the encoding modality. Interestingly, we also found no conversion from verbal encoding to pictorial testing. When participants encoded the crime-related items verbally and were tested with pictorial representation of the items, crime-knowledge was not detected. These results are in line with the principles of the dual code and transfer-appropriate processing account and highlight the involvement of both perceptual, but in particular conceptual memory processing in the RT-CIT.

References

Ambach, W., Bursch, S., Stark, R., & Vaitl, D. (2010). A concealed information test with multimodal measurement. *International Journal of Psychophysiology, 75*(3), 258-267. doi:10.1016/j.ijpsycho.2009.12.007

Ben-Shakhar, G., Frost, R., Gati, I., & Kresh, Y. (1996). Is an apple a fruit? Semantic relatedness as reflected by psychophysiological responsivity. *Psychophysiology, 33*(6), 671-679. doi:10.1111/j.1469-8986.1996.tb02363.x

Bevan, W., & Steger, J. A. (1971). Free recall and abstractness of stimuli. *Science, 172*(3983), 597-599. doi:10.1126/science.172.3983.597

Bradley, M. T., Barefoot, C. A., & Arsenault, A. M. (2011). Leakage of information to innocent suspects. *Memory detection: Theory and application of the Concealed Information Test*, 187. doi:10.1017/cbo9780511975196.011

Childers, T. L., & Houston, M. J. (1984). Conditions for a picture-superiority effect on consumer memory. *Journal of Consumer Research, 11*(2), 643-654. doi:10.1086/209001

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences.* New York, NY: Routledge Academic.

Debey, E., De Schryver, M., Logan, G. D., Suchotzki, K., & Verschuere, B. (2015). From junior to senior Pinocchio: A cross-sectional lifespan investigation of deception. *Acta psychologica, 160*, 58-68. doi:10.1016/j.actpsy.2015.06.007

Ensor, T. M., Surprenant, A. M., & Neath, I. (2019). Increasing word distinctiveness eliminates the picture superiority effect in recognition: Evidence for the physical-distinctiveness account. *Memory & Cognition, 47*(1), 182-193. doi:10.3758/s13421-018-0858-9

Geven, L. M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2019). It's a match!? Appropriate item selection in the Concealed Information Test. *Cognitive Research: Principles and Implications, 4*(1), 1-11. doi:10.1186/s41235-019-0161-8

JASP Team (2020). *JASP* [Computer software]. https://jasp-stats.org/

Kirkpatrick, E. A. (1894). An experimental study of memory. *Psychological Review, 1*(6), 602. doi:10.1037/h0068244

Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test. *PLOS ONE, 10*(4), e0118715. doi:10.1371/journal.pone.0118715

klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2016). Orienting versus inhibition in the Concealed Information Test: Different cognitive processes drive different physiological measures: Orienting and inhibition processes in the CIT. *Psychophysiology, 53*(4), 579-590. doi:10.1111/psyp.12583

klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2017). Unraveling the roles of orienting and inhibition in the Concealed Information Test: Orienting and inhibition in the CIT. *Psychophysiology, 54*(4), 628-639. doi:10.1111/psyp.12825

Lukács, G., & Ansorge, U. (2019). Information leakage in the response time-based Concealed Information Test. *Applied Cognitive Psychology, 33*(6), 1178-1196. doi:10.1002/acp.3565

Madigan, S. (2014). Picture memory. *Imagery, Memory and Cognition*, 65-89.

Meijer, E. H., klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51*(9), 879-904. doi:10.1111/psyp.12239

Nelson, D. L., Cermak, L., & Craik, F. (1979). Remembering pictures and words: Appearance, significance and name. *Levels of processing in human memory*, 45-76. doi:10.4324/9781315796192-11

Norman, D. G., Wade, K. A., Williams, M. A., & Watson, D. G. (2020). Caught virtually lying-Crime scenes in virtual reality help to expose suspects' concealed recognition. *Journal of Applied Research in Memory and Cognition.* doi:10.1016/j.jarmac.2019.12.008

Osugi, A. (2011). Daily application of the Concealed Information Test: Japan. *Memory detection: Theory and application of the Concealed Information Test*, 253-275. doi:10.1017/CBO9780511975196.015

Paivio, A. (1971). *Imagery and Verbal Processes* (1st ed.)., New York: Holt, Rinehart and Winston.

Rosenfeld, J. P., Ward, A., Thai, M., & Labkovsky, E. (2015). Superiority of pictorial versus verbal presentation and initial exposure in the P300-based, complex trial protocol for concealed memory detection. *Applied Psychophysiology and Biofeedback*, *40*(2), 61-73. doi:10.1007/s10484-015-9275-z

Rugg, M. D., Johnson, J. D., & Uncapher, M. R. (2015). Encoding and retrieval in episodic memory. *The Wiley Handbook on the Cognitive Neuroscience of Memory*, 84-107. doi:10.1002/9781118332634.ch5

Sansgiry, S. S., Cady, P. S., & Adamcik, B. A. (1997). Consumer comprehension of information on over-the-counter medication labels: effects of picture superiority and individual differences based on age. *Journal of Pharmaceutical Marketing & Management*, *11*(3), 63-76. doi:10.3109/j058v11n03_05

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological methods, 22*(2), 322. doi:10.1037/met0000061

Seymour, T. L., & Kerlin, J. R. (2008). Successful detection of verbal and visual concealed knowledge using an RT-based paradigm. *Applied Cognitive Psychology: The Official*

Journal of the Society for Applied Research in Memory and Cognition, 22(4), 475-490. doi:10.1002/acp.1375

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior, 6*(1), 156-163. doi:10.1016/s0022-5371(67)80067-7

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology, 18*(6), 813-847. doi:10.1080/09541440500412361

Stenberg, G., Radeborg, K., & Hedman, L. R. (1995). The picture superiority effect in a cross-modality recognition task. *Memory & Cognition, 23*(4), 425-441. doi:10.3758/bf03197244

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143*(4), 428-453. doi:10.1037/bul0000087

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E., Derks, K., Dablander, F., Gronau, Q. F., Kucharský, Š., Gupta, A. N., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique, 120*(1), 73-96. doi:10.3917/anpsy1.201.0073

Verschuere, B., & De Houwer, J. (2011). Detecting concealed information in less than a second: Response latency-based measures. *Memory detection: Theory and application of the Concealed Information Test,* 46-62. doi:10.1017/CBO9780511975196.004

Weldon, M. S., & Roediger, H. L. (1987). Altering retrieval demands reverses the picture superiority effect. *Memory & Cognition, 15*(4), 269-280. doi:10.3758/bf03197030

Weldon, M. S., Roediger, H. L., & Challis, B. H. (1989). The properties of retrieval cues constrain the picture superiority effect. *Memory & Cognition, 17*(1), 95-105. doi:10.3758/bf03199561

3

Memory detection based on eye movements

This chapter is under review at Journal of Applied Research in Memory and Cognition as:
Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., & Verschuere, B. (2022). *Detecting concealed familiarity using eye movements: The effect of leakage of mock crime details to innocents.*

Abstract

The current study examined the eye tracking Concealed Information Test (CIT) in a mock crime scenario. Participants were instructed to either commit a mock crime on campus (guilty participants; *n* = 42), read an article about this mock crime (informed innocents; *n* = 45), or read an unrelated article (naïve innocent participants; *n* = 46). Afterward, all participants were presented with an eye tracking CIT task. Based on preregistered analyses of participants' gaze behavior, we were able to distinguish the guilty participants from the naïve innocents (AUC = .71). Interestingly, we were also able to distinguish the guilty participants from the informed innocent ones (AUC = .65). Although these results are promising, the observed detection efficiency was lower than both previous eye tracking CIT studies that used highly familiar stimuli, as well as mock crime CIT studies relying on physiological measures.

The fact that physiological responses can be used to reveal hidden memories is of interest to crime investigations. Ever since David Lykken introduced the test in 1959 (Lykken, 1959), the CIT has proven to be a valid method for determining whether an examinee recognizes crime-relevant details (for a review, see Verschuere et al., 2011). The CIT relies upon the assumption that only the actual perpetrator has specific knowledge about a crime. By comparing the examinees' physiological responses to crime-relevant items known only to the actual perpetrator (e.g., the murder weapon), to equally plausible alternatives (e.g., gun, knife, rope), concealed knowledge can be detected. After decades of research, it is well established that the CIT provides a valid, theory-driven, tool for the detection of "knowledgeable" individuals (Meijer et al., 2014) and is routinely used in criminal investigations in Japan (Osugi, 2011).

Traditionally, the possession of concealed knowledge is detected based on autonomic nervous system activity (i.e., electrodermal, respiratory, and heart rate responses; Gamer, 2011). However, capturing physiological measures could be challenging in certain ecological settings like airports, border control, or police interviews, since they require a laboratory setup and direct contact with the examinee. Recently, the relative ease and rapidity of video-based eye tracking have sparked an interest within the scientific community to use this tool for the detection of concealed knowledge (Gamer & Pertzov, 2018; Lancry-Dayan et al., 2022).

Several studies have already shown that eye movements are not only influenced by features of the stimuli, but also by the observer’s memory, which appears to have a substantial impact on visual exploratory behavior. In a series of studies, Neil Cohen and colleagues showed that familiar faces are scanned with a smaller number of fixations and fewer regions sampled as compared to unfamiliar faces (e.g., Althoff & Cohen, 1999). This memory-guided gaze dynamic also occurs for recently learned faces (Heisz & Shore, 2008), as well as complex scenes (Ryan et al., 2000). This link between memory and eye movements is of

immediate relevance for the detection of concealed information.

Indeed, initial CIT studies based on eye tracking revealed that familiar stimuli elicit fewer and longer fixations which are directed less toward the inner regions of the stimulus as compared to unfamiliar stimuli (Millen & Hancock, 2019; Millen et al., 2017, 2020; Peth et al., 2013, 2016). However, these studies typically resulted in a lower detection efficiency (measured as the area under the Receiver Operating Characteristic [ROC] curve [AUC]) than physiological measures. For example, Peth et al. (2016) directly compared CIT detection efficiency based on physiological to efficiency based on various eye movement measures and reported an AUC of 0.88 for physiological measures and AUCs ranging between 0.59 and 0.83 for eye movement measures. Only the number of fixations showed a relatively good detection efficiency with an AUC of 0.83. This relatively low detection efficiency for eye movements may be accounted for by the sequential presentation of the familiar and unfamiliar items. A parallel presentation of both familiar and unfamiliar items can provide additional important information as participants can choose when and for how long they look at each stimulus, which could possibly unfold various patterns of visual attention related to recognition.

Lancry-Dayan et al. (2018) introduced a novel approach to memory detection based on eye tracking, which included both a parallel and a single stimulus display. Specifically, a display of four stimuli was presented, followed by a presentation of a single stimulus. Participants had to perform a short term memory task and decide whether the single item appeared in the display of four items presented a few seconds earlier. The addition of this task is particularly important because a successful performance requires participants to focus more on the unfamiliar items and less on the familiar ones (Nahari et al., 2019). Lancry-Dayan et al. (2018) indeed found that when one of the four faces was personally familiar, participants' gaze was initially directed toward it, followed by a strong preference towards the other,

unfamiliar faces. This pattern was also found when participants were explicitly asked to conceal recognition of familiar faces or when they were instructed to look equally at all faces. Using the gaze deployment pattern, a signal detection analysis revealed high detection efficiency estimates (ranging from 0.89 to 0.97), which are even larger than estimates derived from physiological measures (see Meijer et al., 2014), suggesting the potential applicability of detecting "knowledgeable" individuals based on their eye movements.

However, Lancry-Dayan et al. (2018) used only highly familiar stimuli (faces of close friends), which can be used in restricted cases (e.g., identifying members of a terror group), but are not very common in realistic criminal investigations. Moreover, previous studies have found that the degree of familiarity has an influence on behavioral and neuronal responses (Buttle & Raymond, 2003; Leveroni et al., 2000). Specifically, it has been shown that newly learned stimuli -such as crime related items- can limit the scope of memory detection based on eye tracking (Millen et al., 2017, 2020). This is also the case for the paradigm developed by Lancry-Dayan et al. (2018). A recent eye tracking CIT study that used this paradigm found lower detection efficiency estimates for newly learned compared to highly familiar items (Lancry-Dayan et al., 2021). Thus, the primary goal of the present study is to assess the detection efficiency of eye movement measures in the paradigm proposed by Lancry-Dayan et al. (2018) in a mock crime context.

Furthermore, the notion that only guilty suspects have knowledge about a crime is, unfortunately, not always valid. Through newspapers, social media, or even during interrogation itself, innocent suspects may be exposed to critical details of the crime and could therefore also show enhanced responses to the crime related items (Bradley et al., 2011). Several CIT studies based on autonomic nervous system measures have already shown that such leakage can indeed result in false positive outcomes (Bradley et al., 2011). Various attempts have been made to overcome this problem (such as presenting the items on an

exemplar level, rather than a categorical level, which is less susceptible to leakage; Geven et al., 2019, 2022), but these attempts were only partially effective and so far no ideal solution was offered to the leakage problem. It is yet unclear how information leakage may affect the eye tracking measures. Thus, the second goal of the present study is to examine whether and to what extent eye movement measures are affected by information leakage.

In the current study, we apply the novel eye tracking paradigm (Lancry-Dayan et al., 2018) in a more ecologically valid 'mock crime' situation. To that end, participants were randomly assigned to either the guilty condition (commit a mock crime), the informed innocent condition (read an article about the mock crime), or the naïve innocent condition (read an unrelated article). Next, participants were informed that they were suspected of committing this crime. Our primary research question was whether guilty participants can be distinguished from naïve innocents based on their eye movements when performing a short term memory task similar to the one used in Lancry-Dayan et al. (2018). Our secondary research question was whether informed innocents can be differentiated from guilty and naïve innocent participants.

Method

After running a preliminary study (preregistration: https://osf.io/eh9dv) which yielded no significant results, we modified the design and method, which resulted in the current study. There were two main problems with this preliminary study. First, we had some empty cells in our mixed design, preventing us from running all our planned ANOVAs (i.e., we had a two [condition: guilty vs innocent, between subjects] by two [item: crime related vs foil, within subjects] by two [leakage: leaked vs non-leaked, within subjects] mixed design, but we did not leak foils, resulting in some empty cells). Second, our study had insufficient statistical power (there were only 14 guilty participants and 21 innocent participants and a substantial

exclusion rate of 27%). The key changes applied in the current study were the between subjects manipulation of leakage (instead of within subjects) and a strong increase in power. To be transparent, the data and materials of this preliminary study have been made publicly available and can be accessed at https://osf.io/se6nd/.

The current study was preregistered (https://osf.io/sa6yb), and we made no deviations to the preregistration. All materials, data, and analytic scripts of the study can be accessed at https://osf.io/p562n/. The study was approved by the ethical committee of the Social and Behavioral Sciences faculty at the University of Amsterdam and registered as number 2019-CP-11391. The protocol was carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

Participants

We determined our sample size based on a power analysis. Specifically, we aimed for 90% power to detect an AUC of at least .70. AUC is a standard measure used in CIT studies to assess how well guilty participants can be discriminated from naïve innocent participants. Meta-analyses of CIT studies revealed that the average AUCs range from .74 for heart rate to .94 for fMRI (Meijer et al., 2014, 2016). To provide a more conservative estimate that is practically meaningful we aimed to be able to detect an AUC of at least .70. Using MedCalc as statistical software, this resulted in a required sample size of 41 participants per group, resulting in a total sample size of minimum N = 123. We continued testing until we had 41 inclusions in each group, which resulted in a total sample size of 155 participants.

Participants with normal or corrected to normal vision were recruited through the online recruitment system of the University of Amsterdam and received 1 credit or 10 euros for participation. Out of these 155 participants, following our preregistered criteria, we excluded 3 participants because they had less than 70% valid trials (i.e., no pupil detected, due to closed eyes or large head movements), 8 participants because they had

calibration/validation problems, 7 participants because they underperformed on the short term memory task (i.e., because they had less than 60% correct trials), 2 participants of the guilty condition who did not perform the mock crime properly (i.e., who did not return with the two stolen items), and 2 participants of the guilty condition who did not remember at least 5 out of 8 crime related items in the post-study memory test (see below).

The final sample size contained 133 participants, of which 42 were assigned to the guilty condition, 46 to the naïve innocent condition, and 45 to the informed innocent condition. 77% of the sample was female, 22% was male, and 1% defined themselves as other. They had a mean age of 21 years (SD = 4 years) and were from the Netherlands (27%), Germany (11%), or one of 43 other countries.

Procedure

At the beginning of the experiment, participants gave written informed consent and were asked to shut down their phones.

Mock crime

Participants were randomly assigned to either the guilty condition, the informed innocent condition, or the naïve innocent condition. Participants received instructions from the experimenter accordingly. Guilty participants were instructed to commit a mock crime on campus in which they had to steal a *wallet* and a *phone* from another *researcher*'s lab. The phone was hidden in a jacket; the wallet was in a *safe* covered by a *scarf.* The key to the safe was hidden in a *purse*. The participants received *gloves* and a *backpack* from the experimenter to carry their loot (the eight crime related items are italicized in the previous sentences; the exact instructions can be found at https://osf.io/5grs7). Informed innocents were instructed to leave the lab and read an article about the crime in which pictures describing the theft were included. Naïve innocents were instructed to leave the lab to read an unrelated article.

Upon return, all participants were informed by the experimenter that there had been a

crime on campus of which they were suspected. Then, participants did a short term memory task similar to the one used in Lancry-Dayan et al. (2018). They were told that they should try to appear innocent and were promised an extra reward if they would be classified as innocent by the computer.

Short term memory task

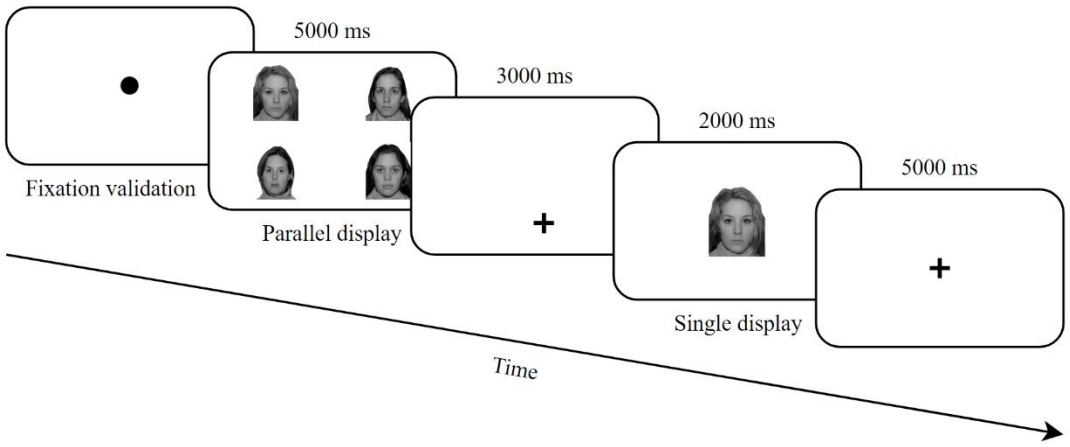
The task was displayed on a 23 in. Syncmaster monitor, with a 120 Hz refresh rate, and a 1024 × 768 screen resolution. Monocular gaze position was tracked at 1000 Hz with an Eyelink 1000+ (SR Research Ltd., Mississauga, Ontario, Canada). Participants' head was stabilized using a chinrest, situated 60 cm from the screen. The experiment started with the standard nine-point calibration and validation procedure provided with the eye tracker. Afterward, all participants did five practice trials, with a set of stimuli not used during the actual test, to get familiar with the task. Participants had to complete at least three correct practice trials to be able to continue with the real experiment. Participants who failed in more than two out of the five trials did another session of five practice trials. Participants repeated this procedure until they reached the minimal threshold of three correct practice trials.

Each trial of the main task started with a fixation validation process. During this process, participants had to fixate on a fixation point in the middle of the screen and press the space bar. When the visual angle between the predicted gaze position and the center of the fixation point was less than one degree, participants were able to continue. Larger deviations were accompanied by an error beep and the opportunity to repeat the calibration process. After the fixation validation process, participants saw a parallel display of four stimuli (5000 ms), followed by a fixation point (3000 ms), a single stimulus display (2000 ms), and a blank screen with a central fixation point (5000 ms; see Figure 1). The fixation point before the single display was displayed below the stimulus to avoid a biased gaze position to any specific location on the stimulus (Arizpe et al., 2012; Peterson & Eckstein, 2013). During the

single stimulus display, participants were required to press "p" if the single stimulus also appeared in the previous display with four stimuli, and "q" if it did not.

Figure 1

Schematic overview of the course of a trial of the short term memory task



The short term memory task consisted of 64 trials in total. For each participant, there were 8 unique crime related items (see above) and 56 unique foils. Half of the trials contained one crime related item and 3 foils in the parallel display. The other half contained only foils. Each mock crime item was presented 4 times, once in each location (top left, top right, bottom left, and bottom right). In the single display, 25% of the displays contained a crime related item, and the other 75% contained a foil. Half of the trials in the single display contained an item that also appeared in the parallel display and required the answer "p". The other half contained an item that did not appear in the parallel display and required the answer "q". Thus, a crime related item could be displayed only in the parallel display, only in the single display, in both displays, or none.

After the short term memory task, participants were informed that the deception detection study was over and were asked to honestly enter their demographics, report the items that were stolen during the mock crime (or guess, if they were innocent), and rate the

significance level of all crime related items and 8 foils. This information was used to examine whether the effect was driven primarily by recognition and/or by item significance. Finally, guilty participants were asked to return the two stolen items.

Results

Eye movement data were parsed into saccades and fixations using Eyelink's standard parser configuration (i.e., samples of which the deviation of consecutive samples exceeded 30°/s velocity or 8000°/s² acceleration were defined as saccades, intervals in between saccades were defined as fixations).

Confirmatory primary analysis

AUC for guilty vs. naïve innocent participants

The main research question of the current study was whether guilty participants can be distinguished from naïve innocent participants based on their gaze behavior. To examine the detection efficiency of the eye tracking measures in differentiating between guilty and naïve innocent participants, a ROC curve was constructed. For this purpose, a combined index of all 6 relevant indices from Lancry-Dayan et al. (2018; i.e., number of visits, fixation count, dwell time during the first phase [1-1000 ms], dwell time during the second phase [1000-5000 ms] of the parallel display, mean fixation duration and reaction time during the single display) was computed for crime items, for each participant. To this end, we first standardized all indices within each participant across all trials, using a Z transformation. The direction of the Z score for each measure was defined according to the results reported by Lancry-Dayan et al. (2018; e.g., the number of visits was expected to be lower for guilty participants, so these z scores were multiplied by -1 such that higher indices reflected more recognition). Then we computed the detection score as the average of these 6 Z scores of the crime items, using equal weights. Finally, the detection-score distribution of the guilty participants was

compared to the detection score distribution of naïve innocent participants and the AUC was computed.

This analysis showed that we were able to distinguish guilty participants from naïve innocent ones significantly above chance with an AUC of .71, 95% CI [.60; .82], $p = .001$.

Confirmatory secondary analyses

AUC for other conditions

The secondary research question of the current study was whether guilty participants can be distinguished from informed innocents and whether informed innocent participants can be distinguished from naïve innocents. Similar to our primary analysis, we computed the AUCs for guilty vs. informed innocents and informed innocents vs. naïve innocents. These analyses showed that while informed innocents could be significantly distinguished from guilty participants (AUC = .65, 95% CI [.53; .77], $p = .028$), they could not be significantly distinguished from naïve innocents (AUC = .61, 95% CI [.49; .73], $p = .082$).

Analysis of variance (ANOVA)

Lastly, we ran two orthogonal contrasts on the combined index: 1. Comparing all knowledgeable participants (i.e., guilty + informed innocent participants) to unknowledgeable participants (i.e., naïve innocent participants) where the effect of interest was the interaction between stimulus type (crime related item vs. foil) and group, and 2. Comparing guilty to informed innocents, where once again the factor of interest was the interaction.

In the first ANOVA, the interaction was significant, $F(1, 131) = 8.75$, $p = .004$, $\eta^2 = 0.04$, indicating that the difference between crime related items and foils was larger for knowledgeable participants than for unknowledgeable participants. In the second ANOVA the interaction was also significant, $F(1, 85) = 7.80$, $p = .006$, $\eta^2 = 0.04$, indicating that the difference between crime related items and foils was larger for informed innocents, compared to naïve ones.

Exploratory analyses

Individual measurements

To test the contribution of the individual measures to the combined index, we ran the analyses again with each separate measure, comparing the guilty participants to the naïve innocents. The results of these analyses can be found in Table 1. Results comparing guilty to informed innocents and comparing informed innocents to naïve innocents can be found on osf: https://osf.io/56qv7.

Table 1

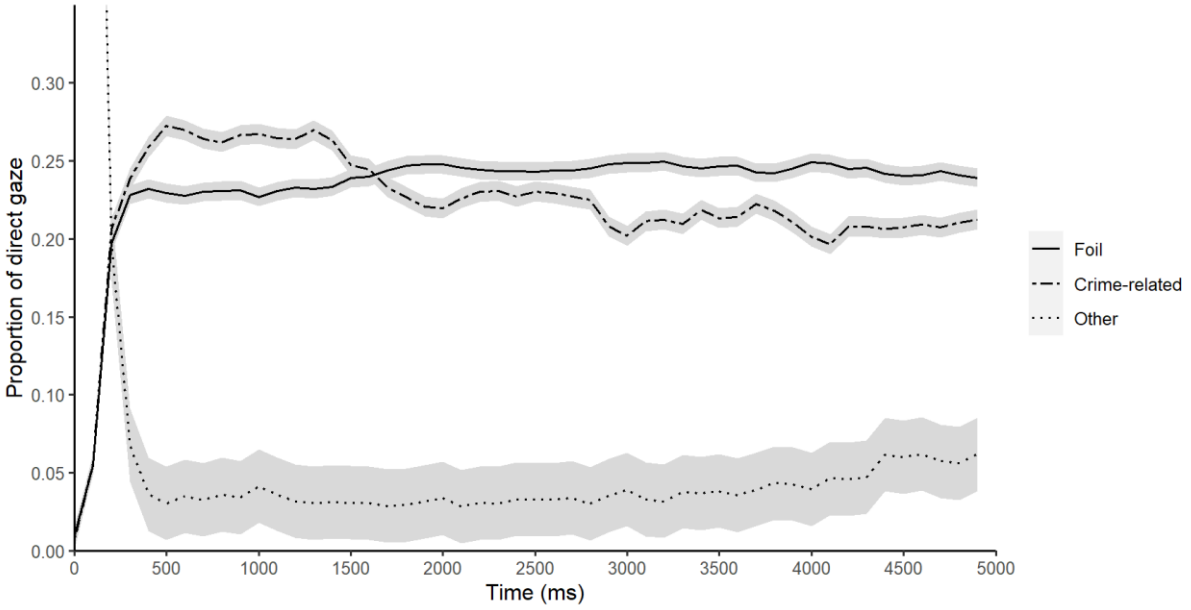
AUC, 95% CI, and p-values of the different measurement that were part of the combined index

Display	Measure	AUC [95% CI]	*p*
Parallel display	Dwell time first phase	.67 [.56; .79]	.005
	Dwell time second phase	.63 [.51; .75]	.041
	Number of visits	.70 [.59; .82]	.001
	Fixation count	.60 [.48; .72]	.114
Single display	Mean duration of fixation	.60 [.48; .72]	.102
	Reaction time	.57 [.44; .69]	.301

Although the differences between these measures are small, Table 1 suggests that the variables of the parallel display seemed to contribute more than the variables of the single display. As compared to naïve innocent participants, guilty participants visited the crime related items less, dwelled more on them during the first phase of the parallel display, and less during the second phase of the parallel display. This pattern of an initial preference followed by an avoidance of crime related items is similar to the pattern observed by Lancry-Dayan et al. (2018). This particular pattern in guilty participants is visualized in Figure 2.

Figure 2

Time course of gaze position of guilty participants towards crime related items, foils, or other areas.



Memory test

At the end of the experiment, participants were asked to try to recognize (or guess) the crime-related items in a multiple choice question (chance level: 12.5%). The results for each crime related item for the guilty, informed innocent, and naïve innocent participants can be found in Table 2. A one-way ANOVA showed that there was a significant difference between the three conditions, $F(2, 130) = 189.25$, $p < .001$, $\eta^2 = 0.74$. Post hoc comparisons with Tukey's correction revealed that guilty participants had a higher accuracy (81%) than informed innocent participants (46%), $t(85) = 11.34$, $p < .001$, $d = 2.43$, and than naïve innocent participants (21%), $t(86) = 19.40$, $p < .001$, $d = 4.14$. Informed innocent participants also scored significantly higher than naïve innocent participants, $t(89) = 8.14$, $p < .001$, $d = 1.71$. Across conditions, there was a significant positive correlation between participants' memory test results and their combined index, $r = 0.26$, $p = .002$. More accurate results on the memory test were related to higher combined indexes.

Table 2

Memory test results for each condition

	Guilty	Informed innocent	Naïve innocent
Phone	71%	53%	24%
Wallet	79%	62%	26%
Victim	98%	96%	26%
Safe	93%	40%	24%
Scarf	93%	38%	22%
Purse	98%	4%	20%
Gloves	95%	42%	9%
Backpack	24%	33%	22%
Average	**81%**	**46%**	**21%**

Significance ratings

At the end of the experiment, participants also had to indicate how significant they found each crime related item on a 9 point Likert scale (where 1 was not significant at all, and 9 was very significant). The results for each crime related item in each condition can be found in Table 3. A one-way ANOVA showed that there was a significant difference between the three groups, $F(2, 130) = 4.55$, $p = .012$, $\eta^2 = 0.07$. Post hoc comparisons with Tukey's correction showed that significance ratings were higher for guilty participants (4.62) compared to naïve innocents (3.51), $t(86) = 2.88$, $p = .013$, $d = 0.61$, but not in comparison to informed innocents (3.74), $t(85) = 2.27$, $p = .064$, $d = 0.49$. Informed innocent participants did not rate the items significantly higher than naïve innocent participants, $t(89) = 0.61$, $p = .814$, $d = 0.13$. Across conditions, the correlation between participants' significance ratings and their combined index was not significant, $r = 0.12$, $p = .177$.

Table 3

Significance ratings of the crime items for each condition

	Guilty	Informed innocent	Naïve innocent
Phone	4.55	4.42	4.07
Wallet	4.07	3.91	3.46
Victim	5.81	5.11	3.72
Safe	4.64	3.51	3.13
Scarf	4.55	3.36	3.13
Purse	5.45	3.27	3.67
Gloves	5.00	3.16	3.20
Backpack	2.86	3.20	3.72
Average	**4.62**	**3.74**	**3.51**

Discussion

Can eye movements reveal crime enactment? As expected, we found that it is possible to distinguish guilty participants from naïve innocents based on their gaze behavior. Additionally, we found that it is possible to distinguish guilty participants from innocent participants who have crime knowledge because they read about the crime.

Mock crime

This was the first study that tested the short term memory CIT task in a mock crime design. Although we found that it is possible to distinguish between guilty and innocent in a crime-like scenario, the detection efficiency observed in the current study (AUC = .71) was considerably lower compared to the results reported by Lancry-Dayan et al. (2018; AUC = .89). One explanation for this difference could be that the original study used highly familiar faces while our study used newly learned stimuli. Indeed, previous research showed that highly familiar items are detected more efficiently than newly learned items (Lancry-Dayan et al., 2021; Millen et al., 2020).

The lower detection efficiency for newly learned items as compared to highly familiar

items could be explained by a poorer memory for those items (i.e., not all guilty participants may have seen and remembered all relevant information). The recognition test at the end of the study indeed showed suboptimal recognition (81%). One crime related item was not even recognized by a majority of the guilty group and had an accuracy rate of only 24%. The fact that poor recognition can lower the accuracy of the CIT is highly relevant for forensic applications and highlights the importance of using only very salient crime-related items (e.g., Gamer et al., 2010; Nahari & Ben-Shakhar, 2011). This is particularly important when the CIT is administered a long time after the crime.

The detection efficiency estimates in the current study are also lower than what is usually observed with physiological measures that are typically used to detect concealed information. The mean detection efficiency of autonomic nervous system activity is 0.81 (Lancry-Dayan et al., 2022). While awaiting a direct comparison, these estimates seem substantially better than the AUC found in the current study. On the other hand, eye tracking does come with some important advantages as compared to physiological measures. For example, no lab, expensive equipment, or direct contact with the examinee is needed. Detection can be achieved fast and possibly even without the suspect's knowledge. Furthermore, the current method might be better resistant to leakage and countermeasures (Lancry-Dayan et al., 2022). Because of these advantages, future studies should examine the added value of the eye tracking CIT by testing it in combination with physiological measures.

Leakage

This is the first study to explore the effect of leakage in the short term memory CIT task. Leakage is one of the biggest challenges for the application of the CIT. The idea behind the CIT is that people with crime knowledge respond to crime related information, regardless of how this crime knowledge was acquired. This is also why David Lykken, one of the pioneers in CIT research, suggested that if there was any concern that information about the crime had

been leaked, the test should be avoided (1998). Although leakage is a major concern, several studies have demonstrated that in some cases it is possible to distinguish guilty participants from informed innocents (e.g., Geven et al., 2019, 2022). The observation that guilty participants show a larger CIT effect than informed innocents suggests that knowledge is not the only factor that determines the CIT effect. Factors such as motivation to avoid detection and the significance of the CIT items also affect the responses elicited by the critical items (e.g., Ben-Shakhar & Elaad, 2003). Informed innocents may be less engaged in processing the leaked items, resulting in poorer recognition and lower detection scores. This is particularly evident when the test is delayed. For example, Nahari and Ben-Shakhar (2011) demonstrated that guilty and informed innocents show similar recognition and detection efficiency when the CIT is administered immediately after the mock crime. However, when introducing delays, both measures attenuate much more in informed innocents. In the current study, better memory contributed to higher detection scores, but significance did not correlate with detection scores. This suggests that the CIT effect is driven primarily by memory of the critical items.

The current study used a form of leakage that may be unlikely in real life. In real life, leakage often happens only verbally (e.g., by reading or hearing about the crime), while in the current study, leakage was introduced both verbally and pictorially (the article which described the crime verbally, also contained pictures of all crime details). This procedure enhances the effect of leakage because leakage of verbal items is only a partial description of the item (e.g., if a participant was told that a phone has been stolen, it would be impossible for him to show an effect if four different phones are presented in the CIT). Although this strict form of leakage might be less likely in real life, it does show the effects in its most drastic form. This could mean that in real life, it might be easier to distinguish between guilty and informed innocent suspects. Moreover, research has shown that a pictorial test could lead to

better detection as it is more likely to result in a better match between encoding and testing, and it could be resistible to verbal leakage (see Chapter 2; Van der Cruyssen et al., 2021). In sum, the current study might underestimate the potential of the eye tracking CIT because pictorial leakage is rare in real life and because a pictorial test might be more accurate than a verbal test which is still primarily used in practice today (Osugi, 2011).

However, it should be noted that even though the comparison between guilty and informed innocent participants indicates that the two groups can be differentiated, this does not imply that the current method is fully protected against leakage. The current study showed that leakage did reduce detection efficiency (AUCs decreased from .71 to .65). Although a carefully chosen threshold could protect most informed innocents, false-positive outcomes could still occur.

Limitations

In the control condition of naïve innocents, participants read an unrelated article, rather than executing an irrelevant activity. However, the critical factor in the attempt to simulate innocent suspects who are unaware of the crime scene, is to insure that the control participants will not be exposed to the critical items. From this respect, reading an unrelated article or executing an unrelated activity are similarly valid procedures. In the computation of the combined index, we used a conservative approach and assigned equal weights to the 6 components. This was done to avoid the risk of capitalization on chance. Equal weights are recommended when no independent validation of optimal weights based on a large sample is available (Dawes, 1979; Wainer, 1976). Clearly, once a proper estimation of optimal weights is available, detection efficiency based on the eye movement indices can be increased. From this respect, the present detection efficiency estimate should be viewed as an underestimate.

Conclusion

In conclusion, the current study found that it is possible to distinguish guilty participants

from naïve innocent participants based on their eye movements in a mock crime scenario. We found that it is also possible to distinguish guilty participants from innocent participants who have crime knowledge because they read about the crime. Although the observed detection efficiency in the current study is lower than what was found in previous eye tracking CIT studies and than CIT based on autonomic nervous system measures, the advantages of eye tracking imply that this approach may still be relevant for forensic applications. Future research focusing on eye tracking based CIT should examine whether and how detection efficiency with this method could be improved (Lancry-Dayan et al., 2022). Moreover, future CIT studies should use both eye movement and physiological measures as a combination of the measures is likely to yield improved detection efficiency.

References

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: a reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 997–1010. doi:10.1037/0278-7393.25.4.997

Arizpe, J., Kravitz, D. J., Yovel, G., & Baker, C. I. (2012). Start position strongly influences fixation patterns during face processing: Difficulties with eye movements as a measure of information use. *PloS one, 7*(2), e31106. doi:10.1371/journal.pone.0031106

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology, 88*(1), 131–151. doi:10.1037/0021-9010.88.1.131

Bradley, M. T., Barefoot, C. A., & Arsenault, A. M. (2011). Leakage of information to innocent suspects. *Memory detection: Theory and application of the Concealed Information Test,* 187 –199. doi:10.1017/cbo9780511975196.011

Buttle, H., & Raymond, J. E. (2003). High familiarity enhances visual change detection for face stimuli. *Perception & psychophysics, 65*(8), 1296-1306. doi: 10.3758/bf03194853

Dawes, R. M. (1979). The Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist, 34*(7), 571-582. doi:10.1037/0003-066x.34.7.571

Gamer, M. (2011). Detecting concealed information using autonomic measures. *Memory detection: Theory and application of the Concealed Information Test,* 27–45. doi:10.1017/cbo9780511975196.003

Gamer, M., Kosiol, D., & Vossel, G. (2010). Strength of memory encoding affects physiological responses in the Guilty Actions Test. *Biological psychology, 83*(2), 101-107. doi: 10.1016/j.biopsycho.2009.11.005

Gamer, M., & Pertzov, Y. (2018). Detecting Concealed Knowledge From Ocular Responses. *Detecting Concealed Information and Deception,* 169–186. doi:10.1016/b978-0-12-

812729-2.00008-2

Geven, L. M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2019). It's a match!? Appropriate item selection in the Concealed Information Test. *Cognitive research: principles and implications, 4*(1), 1-11. doi: 10.1186/s41235-019-0161-8

Geven, L. M., Verschuere, B., Kindt, M., Vaknine, S., & Ben-Shakhar, G. (2022). Countering information leakage in the Concealed Information Test: The effects of item detailedness. *Psychophysiology, 59*(2), e13957. doi: 10.1111/psyp.13957

Heisz, J. J., & Shore, D. I. (2008). More efficient scanning for familiar faces. *Journal of Vision, 7*(9), 24-24. doi:10.1167/8.1.9

Lancry-Dayan, O. C., Ben-Shakhar, G., & Pertzov, Y. (2022). The promise of eye-tracking in the detection of concealed memories. *Trends in Cognitive Sciences.* doi: 10.1016/j.tics.2022.08.019

Lancry-Dayan, O. C., Nahari, T., Ben-Shakhar, G., & Pertzov, Y. (2018). Do you know him? Gaze dynamics toward familiar faces on a concealed information test. *Journal of Applied Research in Memory and Cognition, 7*(2), 291-302. doi:10.1016/j.jarmac.2018.01.011

Lancry-Dayan, O. C., Nahari, T., Ben-Shakhar, G., & Pertzov, Y. (2021). Keep an eye on your belongings: Gaze dynamics toward familiar and unfamiliar objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* doi:10.1037/xlm0001086

Leveroni, C. L., Seidenberg, M., Mayer, A. R., Mead, L. A., Binder, J. R., & Rao, S. M. (2000). Neural systems underlying the recognition of familiar and newly learned faces. *Journal of Neuroscience, 20*(2), 878-886. doi: 10.1523/jneurosci.20-02-00878.2000

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*(6), 385. doi:10.1037/h0046060

Lykken, D. T. (1998). *A tremor in the blood: Uses and abuses of the lie detector.* Plenum Press.

Meijer, E. H., klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta-analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51*(9), 879–904. doi:10.1111/psyp.12239

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology, 53*(5), 593-604. doi:10.1111/psyp.12609

Millen, A. E., & Hancock, P. J. (2019). Eye see through you! Eye tracking unmasks concealed face recognition despite countermeasures. *Cognitive research: principles and implications, 4*(1), 1-14. doi: 10.1186/s41235-019-0169-0

Millen, A. E., Hope, L., & Hillstrom, A. P. (2020). Eye spy a liar: assessing the utility of eye fixations and confidence judgments for detecting concealed recognition of faces, scenes and objects. *Cognitive Research: Principles and Implications, 5*(1), 1-18. doi:10.1186/s41235-020-00227-4

Millen, A. E., Hope, L., Hillstrom, A. P., & Vrij, A. (2017). Tracking the truth: the effect of face familiarity on eye fixations during deception. *Quarterly Journal of Experimental Psychology, 70*(5), 930-943. doi:10.1080/17470218. 2016.1172093.

Nahari, G., & Ben-Shakhar, G. (2011). Psychophysiological and behavioral measures for detecting concealed information: The role of memory for crime details. *Psychophysiology, 48*(6), 733-744. doi: 10.1111/j.1469-8986.2010.01148.x

Nahari, T., Lancry-Dayan, O., Ben-Shakhar, G., & Pertzov, Y. (2019). Detecting concealed familiarity using eye movements: the role of task demands. *Cognitive research: principles and implications, 4*(1), 1-16. doi: 10.1186/s41235-019-0162-7

Osugi, A. (2011). Daily application of the Concealed Information Test: Japan. *Memory detection: Theory and application of the Concealed Information Test,* 253. doi:10.1017/cbo9780511975196.015

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological science, 24*(7), 1216-1225. doi:10.1177/0956797612471684

Peth, J., Kim, J. S., & Gamer, M. (2013). Fixations and eye-blinks allow for detecting concealed crime related memories. *International Journal of Psychophysiology, 88*(1), 96-103. doi: 10.1016/j.ijpsycho.2013.03.003

Peth, J., Suchotzki, K., & Gamer, M. (2016). Influence of countermeasures on the validity of the Concealed Information Test. *Psychophysiology, 53*(9), 1429-1440. doi:10.1111/psyp.12690

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological Science, 11*(6), 454–461. doi:10.1111/1467-9280.00288

Van der Cruyssen, I., Regnath, F., Ben-Shakhar, G., Pertzov, Y., & Verschuere, B. (2021). Is a Picture Worth a Thousand Words? Congruency Between Encoding and Testing Improves Detection of Concealed Memories. *Journal of Applied Research in Memory and Cognition, 10*(4), 667-676. doi:10.1016/j.jarmac.2021.03.001

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (Eds.). (2011). *Memory detection: Theory and application of the Concealed Information Test.* Cambridge University Press. doi:10.1017/cbo9780511975196

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*(2), 213-217. doi:10.1037/0033-2909.83.2.213

4

The validation of webcam-based eye tracking

This chapter is under review at Behavior Research Methods as:
Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., Guy, N., Cabooter, Q., Gunschera, L. J., & Verschuere, B. (2022). *The validation of online webcam-based eye-tracking: The replication of the cascade effect, the novelty preference, and the visual world paradigm. Manuscript submitted for publication.*

Abstract

The many benefits of online research and the recent emergence of open-source eye-tracking libraries have sparked the interest in transferring time-consuming and expensive eye-tracking studies from lab to web. In the current study, we validate online webcam-based eye-tracking by replicating three robust eye-tracking studies (the cascade effect, n = 134, the novelty preference, n = 45, and the visual world paradigm, n = 32) online using the participant's webcam as eye-tracker with the WebGazer.js library. We successfully replicated all three effects, although the effect sizes of all three studies shrank by 20 to 27%. The visual world paradigm was not only conducted online but also in the lab, using the same participants and a standard laboratory eye-tracker. This showed that replication per se could not fully account for the effect size shrinkage, but that the shrinkage is also due to the use of online webcam-based eye-tracking, which is noisier. In conclusion, we argue that eye-tracking studies with relatively large effects that do not require extremely high precision (e.g., studies with 4 or fewer large regions of interest) can be done online using the participant's webcam. We also make recommendations for how the quality of online webcam-based eye-tracking could be improved.

Humans are the only species on earth with visible sclera (i.e., the white of the eye; Kobayashi & Kohshima, 1997). Even great apes, which are extremely close to humans in evolution, do not have visible sclera. It has been claimed that the white surrounding human's darker-colored iris evolved with evolution to make it easier for humans to follow the gaze direction of their conspecifics (Kobayashi & Kohshima, 2001). Following others' gaze is a valuable feature because gaze direction is an indicator of human visual attention (Just & Carpenter, 2018), and events that capture attention for one person could also be relevant for its conspecifics.

What a person is looking at is not only interesting for non-verbal communication in social interactions, but it is also useful to explore more general questions concerning human attention. Since the late twentieth century, video-based eye-trackers have been used to track the eyes in real-time (Singh & Singh, 2012) by measuring the position of an infrared light reflection on the cornea (i.e., the transparent layer forming the front of the eye), relative to the pupil (Carter & Luke, 2020). This method allows researchers to track gaze behavior and identify what guides visual attention. In the last 20 years, eye-tracking research gained much popularity and became a common measurement tool in many areas of science (Carter & Luke, 2020).

However, even though eye-tracking research has led to very interesting insights in recent years, this research method has some important limitations. The need for a lab, an expensive eye-tracker device, an experienced researcher who is familiar with the method, and a required calibration procedure make eye-tracking research a rather elaborate, expensive, and time-consuming method. Furthermore, these limitations do not allow for field research in natural environments. These restrictions recently sparked an interest in the use of common webcams to infer the eye-gaze locations of participants (e.g., Bott et al., 2017; Semmelmann & Weigelt, 2018). Moreover, the recent social and economic pressures of the COVID-19

pandemic reinforced this existing interest in webcam-based eye-tracking as it would allow studies to move online.

The use of a webcam as eye-tracker would make the research quicker, easier, and cheaper as no lab, experimenter, or dedicated hardware is needed. Moving from lab to web could also allow for reaching a larger and more diverse participant pool more quickly, or a hard-to-reach sample (e.g., patients with dementia; or a US-based researcher wishing to compare US to Chinese participants). The collection of data would no longer be limited by time or location as individuals could participate whenever they want from the comfort of their homes. Importantly, in other fields, research has already shown that the benefits of online research do not necessarily come at a price. Data quality has been shown to be similar to that of lab research (Kees et al., 2017; Walter et al., 2019), and several effects from other fields have already been replicated in online settings (e.g., Dodou & de Winter, 2014; Gosling et al., 2004; Klein et al., 2014; Semmelmann & Weigelt, 2018).

The movement toward online webcam-based eye-tracking research has been facilitated by recent advances in eye-tracking scripts. Open-source eye-tracking libraries such as WebGazer and TurkerGaze enable researchers to use participants' webcams to infer their gaze position in real-time (Papoutsaki et al., 2016; Xu et al., 2015). To do this, the eye-tracking modules build a mapping between the characteristics of the eye (e.g., pupil position) and gaze positions on the screen. The libraries can be easily integrated into any script or experiment with only a few lines of code.

Some studies have already successfully implemented eye-tracking libraries in their online experiments (Semmelmann & Weigelt, 2018; Slim & Hartsuiker, 2021; Yang & Krajbich, 2021). Semmelmann and Weigelt (2018), for example, demonstrated some basic gaze properties (i.e., a fixation task, a pursuit task, and a free viewing task) with online webcam-based eye-tracking. Slim and Hartsuiker (2021), and Yang and Krajbich (2021) both

successfully replicated a behavioral eye-tracking experiment (a visual world experiment and a food choice task respectively), although in both studies most participants did not pass the initial calibration/validation phase, and were excluded from the study (73% and 61%). Moreover, the latter two studies did not directly compare online webcam-based eye-tracking to lab-based eye-tracking. In conclusion, it remains to be established to what extent online webcam-based eye-tracking could be a valid replacement for lab-based eye-tracking and what the cost in terms of capturing cognitive effects on gaze behavior would be.

To validate online webcam-based eye-tracking, we replicated three classic, robust eye-tracking studies online using the participant's webcam as an eye-tracker. Furthermore, in the third study, we directly compared the data of participants undergoing both an online webcam-based and a classic lab-based eye-tracking study. Based on Semmelmann and Weigelt (2018), Slim and Hartsuiker (2021), and Yang and Krajbich (2021), we predict to be able to replicate these effects in a web-based setting. Even though some loss of accuracy can be expected, online eye-tracking studies create unprecedented opportunities as it makes research easier, quicker, and cheaper. This would create great possibilities for studies that require large or hard-to-reach samples or have limited funding. It would also enable the progress of research during pandemic lockdowns.

Study 1: Cascade effect

The first effect we aimed to replicate was the cascade effect, originally shown by Shimojo et al. (2003). The cascade effect refers to the phenomenon that when people choose which of two presented faces they find most attractive, their gaze is initially distributed evenly between the faces, but then gradually prioritizes the face that they eventually choose. Here, we define the cascade effect as the likelihood of looking at the face that people eventually select during the 100 ms before reporting the decision.

Method

This study was preregistered (https://osf.io/ykd25). All materials, data, and analytic scripts have been made publicly available and can be accessed at https://osf.io/p3xac/.

Participants

An a priori power analysis revealed that for a one-sided one-sample binomial test with an alpha of 0.05, the minimum required sample size was 119 participants to reach 90% power to detect a small effect (*g = 0.13*; this was the observed size of the cascade effect in our pilot study of $N = 20$)[3]. Anticipating exclusions, 152 participants were recruited via the online crowdsourcing platform Prolific (https://www.prolific.co). Eligibility was restricted to English-speaking participants with a computer connected to a functioning webcam who did not wear glasses at the time of the experiment and did not participate in the pilot study.

Based on our preregistered exclusion criteria, we excluded 5 participants for showing no variation in estimated eye gaze across all trials, 4 for showing no variation in the selected responses across all trials, and 9 for having more than 50% invalid samples (i.e., missing measurements or measurements falling outside the screen). Our final sample contained 134 participants (87% of the original sample; 69% male, 31% female, and 0% other) with a mean age of 25 years (SD = 7 years, range 18-48 years). They originated from 28 different countries, of which Portugal had the largest share (24%).

Procedure

All participants gave informed consent before taking part in the study. The task was computerized and completed online. In the first part of the experiment, participants provided some demographic information and we double-checked whether they had a working webcam and whether it was placed in a correct manner on their computer. Next, participants saw an

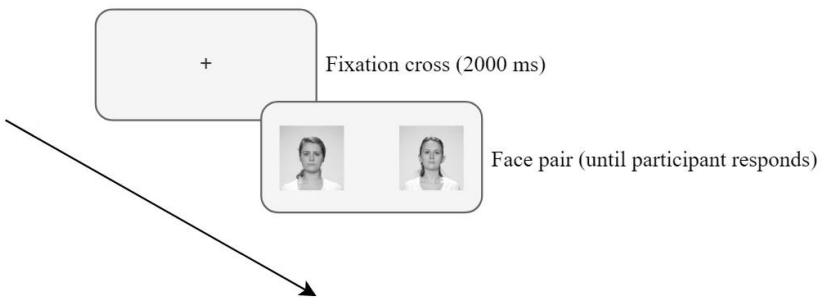
[3] Afterwards, we decided that a one-sided one-sample t-test would be a more appropriate test. To achieve 90% power to detect a medium effect of d = 0.61 (i.e., the effect size of the pilot), we only needed to test 25 participants, which we greatly exceeded in our study.

instruction screen detailing optimal conditions for webcam-based eye-tracking (see Semmelmann & Weigelt, 2018). Once participants indicated that they had successfully set up according to the instructions, they proceeded to an eye-tracking calibration phase (i.e., participants were instructed that they would see a series of white squares and that they had to look at them and click on them), followed by the main task.

Each trial of the main task started with a fixation cross (2000 ms), followed by a pair of faces (see Figure 1). They were instructed to select the one they deemed most attractive by pressing the corresponding key on their keyboard ("F" for the left face, "J" for the right face). They could take as long as they needed to make a decision. There were 18 trials in total. The faces were selected from the London Face Research database (DeBruine & Jones, 2017) which contains images of 102 adult faces with accompanying attractiveness ratings from 2513 individuals. Face pairs were combined based on minimal differences in average attractiveness ratings to replicate the face attractiveness difficult condition of Shimojo et al. (2003). The selected face pairs were matched for gender and ethnicity and were limited to a maximum age difference of four years. The order and composition of the face pairs were fixed, meaning that the presentation and the location of each face were consistent across all participants. Faces were presented on a light gray background and vertically centered. The images had a size of 173 x 173 px and had a distance between each other of 295 px.

Figure 2

Schematic overview of the course of a trial of the face attractiveness task



After completing 18 trials, participants received a short debriefing and were thanked for their participation. The demographics and webcam check were programmed in and hosted on Qualtrics (https://www.qualtrics.com), and the eye-tracking part was programmed in PsychoJS and hosted on Pavlovia (https://pavlovia.org/). For the eye-tracking part, we made use of the open-source eye-tracking library WebGazer (Papoutsaki et al., 2016). The entire study was conducted in English.

Results

Preregistered analyses

As previous studies that used online webcam-based eye-tracking lost many participants because they did not pass the initial validation phase, we used an alternative approach in which we manipulated the data after they had been collected, rather than excluding participants who had a too large offset. We extracted gaze position during the tailing 80% of each central fixation cross at the beginning of each trial and estimated the measurement bias for that given trial. To account for the offset, the estimated bias was then added to the midline and area of interest (AOI) bounds of that trial. For instance, if we found an offset of 50 px to the right of the fixation cross, we would shift the midline and bounds of the AOIs 50 px to the right. As this experiment was limited to a left-right distinction, we applied this correction only for x-values. On average, the midline was corrected for 119 px to either direction (median = 97 px, min = 0 px, max = 480 px). Our confirmatory analyses are based on the midline corrected data. For a comparison between the raw data and the midline corrected data, see our non-preregistered analyses.

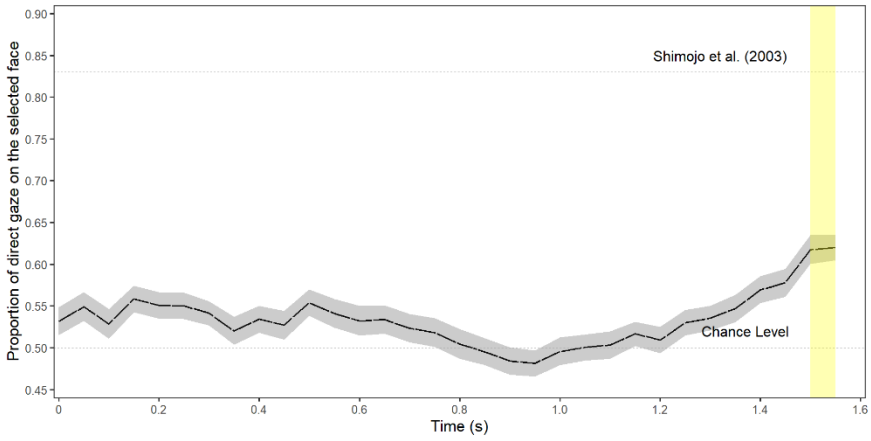
Based on our preregistered exclusion criteria, no trials were excluded due to a deviation of the corrected midline of more than 25% of the total screen width from the true midline; no trials were excluded because the standard deviation of the corrected midline was larger than 25% of the screen width; 12% of the measurement points were excluded because

they fell outside any of the specified AOIs, and 16 % of the trials because the response time was below 0.5 s or above 30 s.

The results revealed that the likelihood of looking at the face that was eventually selected as more attractive by the participant during the 100 ms leading up to the decision was 63% (as compared to the 50% chance level). A one-sided one-sample t-test revealed that this rate was significantly larger than chance, $t(132) = 7.39$, $p < .001$, $d = 0.63$, 95% CI [0.28; ∞][4]. Additionally, a Bayesian one-sample t-test with the default Cauchy scale of 0.707 showed that the data were 1.10×10^9 times more likely under the alternative model that participants looked more at the eventually chosen face than under the null model of no difference in viewing proportion. This cascade effect is also shown in Figure 2, which demonstrates a steady increase in viewing proportion over time, which resembles the overall trend reported by Shimojo et al. (2003), although smaller in magnitude (the original study reported an 83% likelihood of looking at the selected face in the 100 ms leading up to the decision).

Figure 2

The likelihood that an observer's gaze was directed toward the chosen stimulus against time until the decision. In yellow is the period on which we based our analyses



[4] Contrary to what we described in the preregistration, we decided that a t-test was ultimately a more appropriate test for this study than a binomial test. We did also run all analyses as described in the preregistration, from which the same conclusions were drawn. These analyses can be found at https://osf.io/kwnxc.

Non-preregistered analyses

Midline correction. In the raw data, 7% of the measurement points fell outside of any of the AOIs. In the midline corrected data, 5% of the measurement points fell outside of any of the AOIs. This indicates that we captured slightly more measurement points when adjusting for the offset. When we reran the analyses without the midline correction, we found a likelihood of looking at the face that was eventually selected of 61% (compared to 63% with midline correction). This proportion was also significantly higher than chance level, $t(132) = 28.33$, $p < .001$, $d = 0.63$, 95% CI [0.28; ∞]. A Bayesian one-sided one-sample t-test showed that the data were 2.64×10^3 times more likely under the alternative model than under the null model.

Study 2: Novelty preference

The second effect we replicated was the novelty preference effect. This effect refers to the finding that people are more likely to attend to new stimuli than to stimuli they have already seen before. This effect is typically demonstrated with the Visual Paired Comparison Task and has been shown, among others, by Crutcher et al. (2009).

Method

The preregistration, materials, data, and analytic scripts of this study are available at https://osf.io/eqg2n/.

Participants

An a priori power analysis revealed that for a one-sided one-sample t-test (with $\alpha = 0.05$), a minimum sample size of only 6 participants was required to reach 90% power to detect a large effect ($d = 1.47$; this effect size was based on the observed effect size in our pilot study). To make sure we did have a sufficiently large sample size (as the pilot sample size may not suffice to reliably estimate the true effect size; Brysbaert, 2019), and to account

for exclusions, we decided to use a Bayesian stopping rule (Schönbrodt et al., 2017). As described in the preregistration we first opened the experiment to 50 participants on Prolific. Eligibility was restricted to English-speaking participants with a computer connected to a functioning webcam who did not wear glasses at the time of the experiment and did not participate in the pilot study. Then, after applying the preregistered exclusion criteria (see below), we ran a Bayesian one-sided one-sample t-test. The decision to stop collecting data was based on that Bayesian Factor (BF). We planned that once we reached substantial evidence for either the alternative hypothesis (i.e., BF_{10} larger than 5; participants look more at novel stimuli than what could be expected by chance) or the null hypothesis (i.e., BF_{10} smaller than 1/5; participants do not look more at the novel stimuli), we would stop testing, otherwise we would open up the experiment for another 50 participants. After our first batch of 50 participants (which ended up being only 49 participants because one participant did not consent to the use of their data), we reached a BF_{10} of 18.16, so data collection was stopped.

Based on the preregistered exclusion criteria, 1 participant was excluded for showing no variation in estimated gaze position across all trials, and 3 participants were excluded for having a tracking loss percentage (i.e., measurements falling outside any AOI) higher than 50%. The final sample consisted of 45 participants (92% of the original sample; 62% female, 38% male, and 0% other) with a mean age of 27 years (SD = 6 years, range 19-41 years). Participants originated from 14 countries, with the majority having a South African nationality (45%).

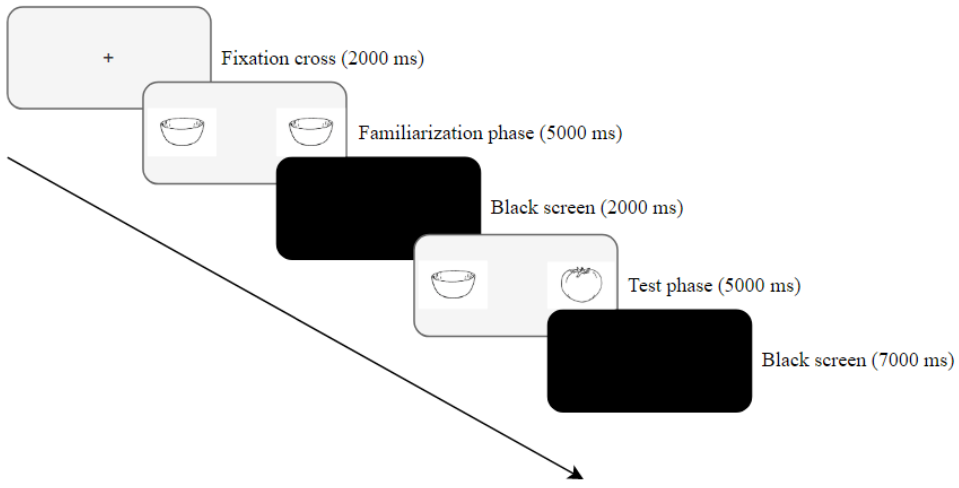
Procedure

The first part of the procedure was the same as in study 1 (i.e., informed consent, demographics and camera check, instructions about optimal conditions for webcam-based eye-tracking followed by a calibration phase). After this first part, participants proceeded to the main novelty preference task.

Each trial of the main task started with a fixation cross (2000 ms), followed by a familiarization phase which consisted of two identical images on the left and right side of the screen (5000 ms). After the familiarization phase, participants saw a black screen (2000 ms), followed by the test phase (5000 ms), in which participants saw two images: One that was the same as the one presented during the familiarization phase and another one that was novel. The left or right positioning of the novel stimulus was randomized across trials. Each experimental trial ended with a black screen (7000 ms; Figure 3). There were 10 trials in total. Stimuli were black and white, horizontally oriented clipart images selected from the Snodgrass and Vanderwart (1980) database. They were presented on a light grey background, centered vertically, had a size of 472 x 331 px, and had a distance between each other of 295 px.

Figure 3

Schematic overview of the course of a trial of the novelty preference task



After the main task, an attention check was performed in which participants were asked to select which of three image pairs they recognized from the experiment. Afterward, they were asked to report how reliable they estimated their own data on a 5-point Likert scale (with 1 as “unreliable, do not use my data”, and 5 as “very reliable, use my data”). The

demographics and webcam check were programmed in and hosted on Qualtrics (https://www.qualtrics.com), the eye-tracking part was programmed in PsychoJS and hosted on Pavlovia (https://pavlovia.org/). For the eye-tracking part, we made use of the open-source eye-tracking library WebGazer (Papoutsaki et al., 2016). The entire study was conducted in English.

Results

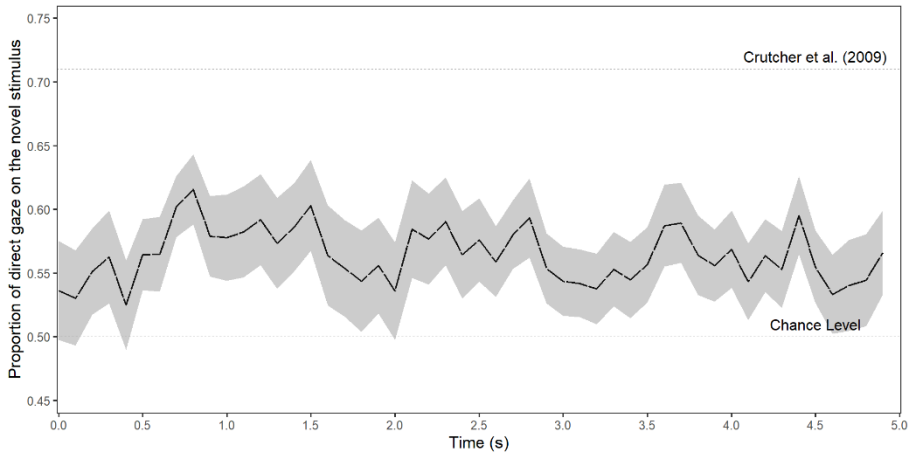
Preregistered analyses

In the current study, we corrected the midline on average for 128 px to either direction (median = 122 px, min = 36 px, max = 256 px). Our confirmatory analyses are based on this midline corrected data. Based on our preregistered exclusion criteria, we excluded 14% of the measurement points because they fell outside any of the specified AOIs, no trials were excluded because the corrected midline deviated more than 25% of the total screen width from the true midline, and 8% of the trials were excluded because the standard deviation of the corrected midline was larger than 25% of the screen width.

On average, participants looked at the novel stimulus 57% of the time (compared to the 50% chance level). A one-sided one-sample t-test revealed that this viewing proportion toward the novel stimulus was significantly higher than chance level (.5), $t(44) = 3.06$, $p = .002$, $d = 0.46$, 95% CI [0.15; ∞]. Additionally, a Bayesian one-sided one-sample t-test with the default Cauchy scale of 0.707 showed that the data were 18.16 times more likely under the alternative model that participants looked more than 50 % at the novel stimulus than under the null model of no difference in viewing proportion. The effect is visualized in Figure 4. The results are similar to those reported in previous studies about the novelty preference such as the paper by Crutcher et al. (2009), but with a smaller effect size. For example, in the original study of Crutcher et al. (2009), participants looked at the novel stimulus 71% of the time.

Figure 4

The likelihood that participants looked at the novel stimulus during the test phase of the novelty preference task



Non-preregistered analyses

Midline correction. In the raw data, 15% of the measurement points fell outside of any of the AOIs. In the midline corrected data, 14% of the measurement points fell outside of any of the AOIs. When we reran the analyses without performing the midline correction, we found a likelihood of 54% for looking at the novel stimulus (compared to 57% with midline correction). This proportion was still significantly higher than chance level, $t(44) = 2.05$, $p = .023$, $d = 0.31$, 95% CI [0.01; ∞]. A Bayesian one-sided one-sample t-test showed that the data were 2.12 times more likely under the alternative model than under the null model.

Sensitivity Analysis: Can we improve the quality of the data? In this non-preregistered analysis, we examined whether alternative ways of analyzing the data could improve the results. This could indicate which choices are most consequential for the outcomes. To that end, we sequentially added more and more exclusion criteria so that our data became more and more strict. First, we excluded participants who failed more than one attention check at the end of the study. Second, we excluded participants who clicked through the instruction screen too fast (i.e., < 7.5 s). Third, we excluded participants who indicated that their data was unreliable at the end of the experiment. Last, we excluded measurement

points in which the live webcam feed (indicating that the eye-tracker lost the participant's eyes) appeared. The results can be found in Table 1. Note that these exclusion criteria led to no or very few exclusions, making the results in the different rows very similar.

Table 1

Non-preregistered Sensitivity Analysis

	N	Proportion direct gaze	*p*	Cohen's d [95% CI]	BF
Confirmatory analyses	45	57%	.002*	0.46 [0.15; ∞]	18.16
+ Attention checks	45	57%	.002*	0.46 [0.15; ∞]	18.16
+ Instruction screen	44	57%	.001*	0.50 [0.18; ∞]	33.85
+ Unreliable data	44	57%	.001*	0.50 [0.18; ∞]	33.85
+ Live feed	44	57%	.001*	0.48 [0.17; ∞]	24.77

Note. The table displays the statistical results from the one-sided one-sample (Bayesian) t-test for the different exclusion criteria. The exclusion criteria are presented in hierarchical order and include the previous criteria.
*Bonferroni corrected, statistically significant with $p < .01$.

Study 3: Visual world paradigm

The first two studies of this validation project indicate that it is possible to replicate robust eye-tracking effects with the participant's webcam while retaining 87-92% of the original sample. However, in both studies, the effect was noticeably smaller compared to the original studies. This could be due to the fact that webcam-based eye-tracking data are noisier, but also to the fact that these studies are replications, as it has been demonstrated that the effect sizes of replications are on average 50% smaller than the original effect sizes (Camerer et al., 2018). To estimate how much of the effect size reduction could be attributed to webcam-based eye-tracking, we set up a third study in which each participant conducted the visual world paradigm both online as well as in the lab with a standard state-of-the-art laboratory eye-tracker (Eyelink 1000+; SR Research Ltd., Mississauga, Ontario, Canada).

We replicated the visual world paradigm effect demonstrating that when people hear

utterances while looking at a visual display showing common objects, some of which mentioned in the sentences, they tend to look more at the images of the words that they hear in the utterances. Among others, this effect has been shown by Huettig and Altmann (2005).

Method

The preregistration, materials, data, and analytic scripts of this study are available at https://osf.io/jucge/.

Participants

As described in our preregistration, we first opened the experiment to 50 participants on the participant recruitment site of the University of Amsterdam. Eligibility was restricted to English-speaking participants with a laptop or computer with a functioning webcam and audio device who were not wearing glasses at the time of the experiment. Then, after applying our preregistered exclusion criteria (see below), we ran 2 Bayesian t-tests using JASP's default Cauchy scale of 0.707. First, we ran a Bayesian one-sided one-sample t-test on the data of the online version comparing the mean proportion viewing time at the target word to chance level (25%). Second, we compared the mean proportion of viewing time towards the target between the two conditions (lab vs. online), using a Bayesian one-sided paired-sample t-test. The decision to stop collecting data was based on the BFs of both t-tests. We planned that once we would reach substantial evidence for either the alternative hypothesis (i.e., BF_{10} larger than 5) or the null hypothesis (i.e., BF_{10} smaller than 1/5), for both tests, we would stop testing, if not, we would test another 25 participants. We intended to repeat this procedure until we would reach substantial evidence in both t-tests for either the null or the alternative model or until we would have tested N = 150. We reached substantial evidence favoring the alternative hypotheses for both models after testing 50 participants.

Based on the preregistered exclusion criteria, no participants were excluded for showing no variation in estimated gaze position across all trials, no participants were

excluded for having a tracking loss percentage (measurements falling outside the AOI) higher than 50%, and 18 participants were excluded for having more than 50% missing data (often due to only finishing either the online version or the lab version of the experiment). The final sample consisted of 32 participants (64% of the original sample) that performed both tasks. The sample consisted of 72% females, 28% males, and 0% others and had a mean age of 20 years (SD = 2 years, range 18-27 years). Most participants indicated to be of German nationality (32%).

Procedure

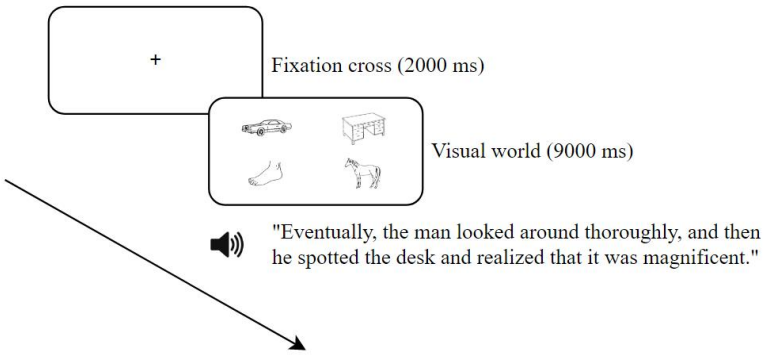
This study had a within-subjects design, in which participants completed the experiment both online and in the lab. The order (first online or first in the lab) was counterbalanced. The online part started with a similar procedure as in studies 1 and 2 (i.e., they started with a webcam and audio check, received instructions about optimal conditions for webcam-based eye-tracking, and finished with a calibration phase). In the lab part, we immediately started with the calibration phase as eye-tracking conditions are already close to optimal in the lab. After the calibration phase, the main task (the visual world paradigm) started, which was identical for the online and lab version but used different stimuli. The set of stimuli that were used online or in the lab was counterbalanced across participants.

Each trial of the main task started with a fixation cross (2000 ms), followed by a screen with four images (e.g., a *desk*, a car, a foot, and a horse), one in each quadrant of the screen (9000 ms). Each screen with images was paired with a sentence such as ‘Eventually, the man looked around thoroughly, and then he spotted *the desk* and realized that it was magnificent’. One of the four images of each scene was a target object that was mentioned in the sentence (in the above example this was the *desk*) and the other three images (car, foot, and horse) were unrelated distractors (see Figure 5). Participants were instructed to listen to the sentences carefully and were told that they could look wherever they wanted (they were

not asked to perform any task). There was a 1000 ms preview of the display before the onset of the sentence and the trial was automatically terminated after 9000 ms, which is typically 2000 ms after the end of each sentence. The target word typically appeared 4000 ms after the onset of the sentence. These materials (sentences and scenes) were recreated based on the materials of the experiment of Huettig and Altmann (2005). The images were presented on a white background, were centered on each of the four quadrants of the screen, and had a size of 265 x 189 px. There were 12 trials in each version of the experiment, so 24 trials in total per participant.

Figure 5

Schematic overview of the course of a trial of the visual world task



After the main task, participants were asked to report how reliable they estimated their own data on a 5-point Likert scale (with 1 as “unreliable, do not use my data” and 5 as “very reliable, use my data”). For the online part, the webcam and audio check were programmed in and hosted on Qualtrics (https://www.qualtrics.com), and the eye-tracking part was programmed in PsychoJS and hosted on Pavlovia (https://pavlovia.org/) in which we made use of the open-source eye-tracking library WebGazer (Papoutsaki et al., 2016). For the lab part, the entire experiment was programmed in Experiment Builder (SR Research Ltd., Mississauga, Ontario, Canada). The study was conducted in English.

Results

Preregistered analyses

In the online part, we corrected the horizontal midline on average for 133 px to either direction (median = 131 px, min = 21 px, max = 265 px), and the vertical midline on average for 134 px to either direction (median = 134 px, min = 48 px, max = 355 px). Our confirmatory analyses are based on the midline corrected data. Based on our preregistered exclusion criteria, no measurement points were excluded because the live webcam feed indicated that the eye-tracker lost the participant's eyes, 15% of the measurement points were excluded because they fell outside any of the specified AOIs, 23% of the trials were excluded because the corrected midline deviated more than 25% (for the vertical midline) or more than 37.50% (for the horizontal midline) of the total screen width from the true midline, and no trials were excluded because the standard deviation of the corrected midline was larger than 25% of the screen width.

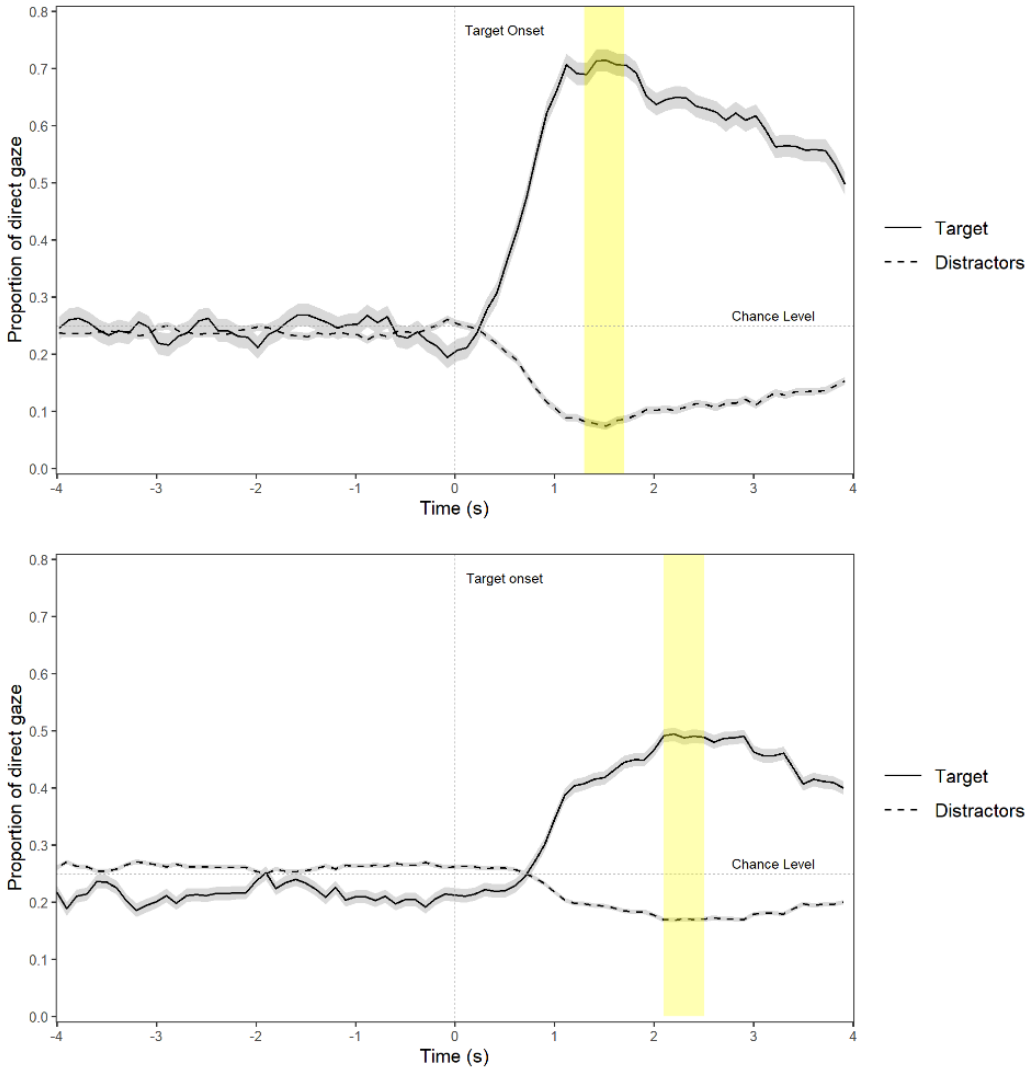
Replication of visual world paradigm online with participants' webcam. The participants looked more at the target items (52%) than the control items (17%) in the online version of the experiment. The proportion of direct fixations on the target items was significantly higher than what could be expected by chance (25%), $t(31) = 6.39$, $p < .001$, $d = 1.13$, 95% CI [0.68; ∞]. Additionally, a Bayesian one-sided one-sample t-test showed that the data were 81 572.97 times more likely under the alternative model that participants looked more at the target item than under the null model of no difference in viewing proportion.

Comparison between lab data and online data. The proportion of fixations on the target item was higher in the lab version (71%) than in the online version (52%). This difference was significant, $t(31) = 3.58$, $p < .001$, $d = 0.63$, 95% CI [0.25; ∞]. Also, a Bayesian one-sided paired-samples t-test showed that the data were 56.75 times more likely under the alternative model (a larger effect size in the lab compared to the online version)

than under the null model (no difference between lab and web). A visualization of the effect in both the lab version and the web version can be found in Figure 6.

Figure 6

The likelihood that participants looked at the target versus distractors in the lab version (top) versus online version (bottom) of the experiment. In yellow is the 400 ms time interval on which we based our analyses



Non-preregistered analyses

Midline Correction. In the raw data, 15% of the measurement points fell outside of any of the AOIs. In the midline corrected data, 16% of the measurement points fell outside of any of the AOIs. When we reran the analyses without doing the midline correction, we found a likelihood of looking at the target of 38% (compared to 52% with midline correction). This

proportion was also significantly higher than the 25% chance level, $t(32) = 4.11$, $p < .001$, $d = 0.72$, 95% CI [0.33; ∞]. A Bayesian one-sided one-sample t-test showed that the data were 215.70 times more likely under the alternative model than under the null model.

Sensitivity Analysis: Can we make the online data more similar to the lab data? We examined whether alternative ways of analyzing the data could converge the online data to the lab data. This could give an indication of which choices are most consequential for the results. To do this, we added an extra exclusion criterion to the online data making it more strict. We excluded participants who indicated that their data were unreliable at the end of the experiment. The results can be found in Table 2.

Table 2

Non-preregistered Sensitivity Analysis

	N	Proportion direct gaze lab	Proportion direct gaze online	*p*	Cohen's d [95% CI]	BF
Confirmatory analyses	32	71%	52%	<.001	0.63 [0.25; ∞]	56.75
+ Unreliable data	31	72%	52%	<.001	0.63 [0.24, ∞]	47.52

Note. The table displays the statistical results from the one-sided one-sample (Bayesian) t-test for the confirmatory analyses, and when the exclusion criterium unreliable data is added.

Calibration Score. Inspired by the paper of Slim and Hartsuiker (2021), we calculated the proportion of the estimated gaze position that fell on the center of the screen during a three-second fixation cross right after the calibration phase. Similar to Slim and Hartsuiker (2021), we found a mean calibration score of 41% (SD = 26%, range 7-93%) in the online webcam-based data. This was considerably lower than the lab-based data, where the mean calibration score was 99% (SD = 4%, range 75-100%). To test whether higher calibration scores could lead to larger effect sizes in the online webcam-based data, we calculated the Pearson correlation between the calibration scores and effect sizes of the online webcam-based data. However, this analysis showed no relation between the two ($r = -0.04$, $p = .843$).

Discussion

Can we validly run eye-tracking studies with the participants' own webcam? We replicated three robust eye-tracking studies (the cascade effect, the novelty preference, and the visual world paradigm) online with the participant's webcam as eye-tracker. All three studies were replicated successfully, and the research was conducted considerably faster, easier, and cheaper than comparable studies conducted in the lab. To illustrate, data collection for studies 1 and 2 was completed within 1 day, and participants from 14 to 28 different countries took part. Moreover, contrary to previous online webcam-based eye-tracking studies, we could retain most of our sample (64-92%). This indicates the potential of online webcam-based eye-tracking procedure that captures gaze directed to left\right or four quadrants of the screen. However, even though the overall effects were replicated, the effect sizes of all three studies shrank by 20 to 27% as compared to lab-based eye-tracking.

Smaller effect sizes

There are several reasons that could explain why the effect sizes of the online webcam-based studies were smaller than those of lab-based eye-tracking studies. For example, it is possible that webcam-based eye-tracking leads to an underestimation of the true effect size. This could be due to smaller numerators (i.e., if the mean differences between the test values of online webcam-based eye-tracking are smaller) or larger denominators (i.e., if the standard deviations of online webcam-based eye-tracking are larger). The third study of the current paper revealed that both the numerator of online webcam-based eye-tracking was 41% smaller than that of lab-based eye-tracking and the denominator was 32% larger.

At the same time, original effect sizes often overestimate the true effect sizes. It has been shown that the effect sizes of replications are typically 50% smaller than the effect sizes of original studies (Camerer et al., 2018). It is argued that this is caused by exaggerated effect sizes in the existing literature due to a combination of publication bias and questionable

research practices (e.g., Simmons et al., 2011; Sterling, 1959). In the third study of the current paper, we found the effect size of the lab-based replication indeed to be lower than that of the original study. However, the effect size of the online webcam-based replication was even lower than that of the lab-based replication, so replication per se could not fully account for the effect size shrinkage in online webcam-based eye-tracking. This indicates that the decreased effect sizes of online webcam-based eye-tracking are probably caused by a combination of both factors.

As in other conceptual replications (i.e., replications where there are changes to the original procedures), the decrease in effect size could also be related to several methodological differences between the original studies and our replications (Zwaan et al., 2018). First, while all original studies were taken in the lab, which is a very controlled environment, all webcam-based replications were done online. Second, as we did not have the original materials for studies 1 and 2, we considerately reconstructed them ourselves. Third, we used different measurement tools. The original studies used standard laboratory eye-trackers, we used webcams. It is possible that we did not measure the exact same construct as the original studies because of these changes, which could result in different effect sizes.

Lower data quality

In the current study, the data quality for online webcam-based eye-tracking was lower than for lab-based eye-tracking. This was probably due to the less controlled environment of online testing and the lower quality of webcam hardware as compared to cameras used in standard eye-trackers. We tested whether this quality could be improved in 2 ways: midline correction and stricter exclusion criteria.

Midline correction

The use of validation trials is a common method of correcting for deterioration of gaze estimation accuracy. Although effective, they can take up large proportions of the

experimental duration and it has been debated how much validation is required to ensure high gaze estimation accuracy (e.g., Semmelmann & Weigelt, 2018). For example, in previous studies that used online eye-tracking, the validation phase took up to 40 to 50% of the experimental duration (Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021). Moreover, only 27- 39% passed the initial validation phase in the studies of Slim and Hartsuiker (2021) and Yang and Krajbich (2021). In the current study, we tested an alternative approach that involves manipulating the data after it has been collected, rather than monitoring the offset throughout the experiment and excluding participants based on this offset (Hornof & Halverson, 2002). This approach has some important benefits such as cutting back on computational demands and experimental duration, and losing fewer participants during the validation phase. Based on the systematic measurement error estimates from the fixation period preceding each trial, we shifted the midline and AOI bounds to the respective direction. Our non-preregistered analyses showed that the effect sizes of the midline corrected data were higher than those of the raw data, suggesting a successful improvement in data quality.

Stricter exclusion criteria

It has been debated that exclusion criteria and attention checks can greatly improve the quality of the data. In our non-preregistered analyses, we tested whether applying stricter exclusion criteria could improve data quality and decrease the differences between online webcam-based eye-tracking and lab-based eye-tracking. However, attention checks did not seem to make large differences in our study. Participants were largely attentive, which led to few or no exclusions based on the criteria. This is in line with the literature indicating online participants to be of similar quality to those acquired via university pools (e.g., Goodman & Paolacci, 2017; Gould et al., 2015). However, our exclusion criteria were merely focused on inattention and the live webcam feed (indicating that the eye-tracker lost the participant's

eyes). Checks for poor environmental circumstances (e.g., too much light, too little light, poor webcam quality, slow internet,…) might have been more successful in improving data quality.

In sum, in all three studies, we were able to slightly improve the data quality. However, in all three studies, the conclusions remained exactly the same with or without midline correction, and with or without stricter exclusion criteria. Furthermore, higher calibration scores did not seem to predict higher effect sizes. This shows that extensive validation phases, data manipulation, and extreme exclusion criteria might not be necessary to reach valid results. Excluding more participants and measurement points might improve the estimated effect size, but might also lead to higher costs and a longer process. Furthermore, hard-to-reach samples are likely excluded in lab-based experiments. The ideal trade-off between data quality and data exclusion/manipulation probably depends on the specific study characteristics and study goal.

Limitations of the current study and moving forward

The experiments in the current paper were quite simple; they only took five to seven minutes and only had two to four AOIs. It would be interesting to investigate how far the effects could be pushed and how many distinct AOIs could be effectively used online. Based on WebGazer's spatial precision, Yang and Krajbich (2021) suggested that up to six AOIs could be used without any degradation in data quality. Their study also took 30 min and with several re-calibrations they were able to maintain appropriate data quality. For longer studies or studies having more than six AOIs, we hypothesize that more re-calibrations might be needed.

Once the groundwork has been established, conceivable applications of online webcam-based eye-tracking are numerous. Among others, the possibility to reach hardly accessible populations is a major advantage for online compared to lab-based research. For instance, eye-tracking has been found valuable for predicting the onset of Alzheimer's disease

(Bourgin et al., 2018; Crawford et al., 2015; Crutcher et al., 2009). Potentially conducting this or similar assessments from the comfort of one's home would not only save resources and reduce interpersonal contact but would allow for reaching thousands of (at-risk) patients, broaden the availability of eye-tracking methodology (even during pandemic lockdowns), and change the way we conduct eye-tracking research in general.

Conclusion

Our study provides evidence for the applicability of online webcam-based eye-tracking. In three replications of robust eye-tracking studies, we demonstrated that the same findings could be obtained when doing the study online with the participant's webcam as eye-tracker. While the conclusions are limited by the simplicity of our task, these replications collaboratively serve to inform gold standards for the application of online webcam-based eye-tracking. Although the data quality was considerably lower, the speed and convenience of online research could be worth the switch for studies with large effect sizes and relatively few AOIs. Moreover, with the ever-improving quality of computer webcams, the future of online eye-tracking looks promising.

References

Bott, N. T., Lange, A., Rentz, D., Buffalo, E., Clopton, P., & Zola, S. (2017). Web camera based eye tracking to assess visual memory on a visual paired comparison task. *Frontiers in neuroscience, 11,* 370. https://doi.org/10.3389/fnins.2017.00370

Bourgin, J., Guyader, N., Chauvin, A., Juphard, A., Sauvée, M., Moreaud, O., Silvert, L., & Hot, P. (2018). Early emotional attention is impacted in Alzheimer's disease: An eye-tracking study. *Journal of Alzheimer's Disease, 63*(4), 1445–1458. https://doi.org/10.3233/JAD-180170

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of cognition.* https://doi.org/10.5334/joc.72

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology, 155,* 49-62. https://doi.org/10.1016/j.ijpsycho.2020.05.010

Crawford, T. J., Devereaux, A., Higham, S., & Kelly, C. (2015). The disengagement of visual attention in Alzheimer's disease: A longitudinal eye-tracking study. *Frontiers in Aging Neuroscience, 7*(118), 1–10. https://doi.org/10.3389/fnagi.2015.00118

Crutcher, M. D., Calhoun-Haney, R., Manzanares, C. M., Lah, J. J., Levey, A. I., & Zola, S. M. (2009). Eye tracking during a visual paired comparison task as a predictor of early

dementia. *American Journal of Alzheimer's Disease & Other Dementias, 24*(3), 258-266. https://doi.org/10.1177/1533317509332093

DeBruine, L., & Jones, B. (2017). Face Research Lab London Set. Figshare. https://doi.org/10.6084/m9.figshare.5047666.v5

Dodou, D., & de Winter, J. C. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior, 36*, 487-495. https://doi.org/10.1016/j.chb.2014.04.005

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research, 44*(1), 196-210. https://doi.org/10.1093/jcr/ucx047

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American psychologist, 59*(2), 93. https://doi.org/10.1037/0003-066x.59.2.93

Gould, S. J. J., Cox, A. L., Brumby, D. P., & Wiseman, S. (2015). Home is Where the Lab is: A Comparison of Online and Lab Data From a Time-sensitive Study of Interruption. *Human Computation, 2*(1), 45–67. https://doi.org/10.15346/hc.v2i1.4

Hornof, A. J., & Halverson, T. (2002). Cleaning up systematic error in eye-tracking data by using required fixation locations. Behavior Research Methods. *Instruments & Computers, 34*(4), 592–604. https://doi.org/10.3758/bf03195487

Huettig, F., & Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96*(1), B23-B32. https://doi.org/10.1016/j.cognition.2004.10.003

Just, M. A., & Carpenter, P. A. (2018). Using eye fixations to study reading comprehension. In *New methods in reading comprehension research* (pp. 151-182). Routledge. https://doi.org/10.4324/9780429505379-8

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality:

Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141-155. https://doi.org/10.1080/00913367.2016.1269304

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Kobayashi, H., & Kohshima, S. (1997). Unique morphology of the human eye. *Nature, 387*(6635), 767-768. https://doi.org/10.1038/42842

Kobayashi, H., & Kohshima, S. (2001). Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. *Journal of Human Evolution, 40*(5), 419-435. https://doi.org/10.1006/jhev.2001.0468

Papoutsaki, A., Daskalova, N., Sangkloy, P., Huang, J., Laskey, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. *International Joint Conference on Artificial Intelligence*, 3839–3845. https://webgazer.cs.brown.edu

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322. https://doi.org/10.1037/met0000061

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods, 50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience, 6*(12), 1317–1322. https://doi.org/10.1038/nn1150

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. https://doi.org/10.1037/e519702015-014

Singh, H., & Singh, J. (2012). Human eye tracking and related issues: A review. *International Journal of Scientific and Research Publications, 2*(9), 1-9.

Slim, M. S., & Hartsuiker, R. (2021). Visual world eyetracking using WebGazer. js. https://doi.org/10.31234/osf.io/5adgf

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

SR Research Experiment Builder 2.1.140 [Computer software]. (2017). Mississauga, Ontario, Canada: SR Research Ltd.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54*(285), 30–34. http://doi.org/10.2307/2282137

Walter, S. L., Seibert, S. E., Goering, D., & O'Boyle, E. H. (2019). A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology, 34*(4), 425-452. https://doi.org/10.1007/s10869-018-9552-y

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755. 91*(12), 5. http://arxiv.org/abs/1504.06755

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making*, *16*(6), 1486. https://doi.org/10.31234/osf.io/qhme6

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.31234/osf.io/4tg9c

5

Memory detection in virtual reality: The effect of scene modifications

The study presented in this chapter is currently being replicated using a larger participant sample. The results of that study will be written up and submitted along with the results outlined in the present chapter.

Abstract

Can your eyes give away your deepest secrets? In the current preregistered study, we examined whether eye movements toward modified areas of a crime scene allow to detect crime knowledge. After committing either a mock crime or an unrelated task, guilty (n = 12) and innocent (n = 21) participants viewed the crime scene in virtual reality to which several salient modifications were made. Guilty participants looked more (d = 0.88) and earlier (d = 1.00) at these modifications as compared to innocent participants. It was possible to significantly distinguish guilty participants from innocent ones based on their eye movements toward the modified areas (AUC = 0.75). While awaiting replication, we argue that this approach has several important advantages in comparison to traditional memory detection.

When we visually observe the world around us, our eyes are constantly moving. We make rapid, jerky movements called "saccades" interspersed with brief pauses called "fixations" (Yarbus, 1967). During these fixations, our eyes take in information, which is then processed in our brains to form a coherent visual representation of what we are looking at. This way, our eye movements allow us to learn more about the world. Interestingly, this can also be applied the other way around. *When* a person is looking *where* can say a lot about what is going on in that person's mind. This way, eye movements also allow us to learn more about a person.

Several studies have already shown that eye movements are not only influenced by features of the stimuli, but also by features of the observers themselves. Even when the visual stimulus is exactly the same, different individuals can generate completely different scanning patterns (Lancry-Dayan et al., 2018; Mehoudar et al., 2014; Tatler et al., 2011). In particular, memory has been shown to have a strong influence on oculomotor behavior. In a series of studies, Neil Cohen and colleagues showed that familiar faces are scanned with a smaller number of fixations and fewer regions sampled as compared to faces that are unfamiliar to the observer (e.g., Althoff & Cohen, 1999). This memory-guided gaze dynamic also occurs for recently learned faces (Heisz & Shore, 2008), as well as complex scenes (Ryan et al., 2000). The effects can occur very rapidly, obligatorily, and even in the absence of conscious awareness (Hannula et al., 2010).

This link between memory and eye movements is highly relevant for memory detection. Memory detection is a technique to detect whether a suspect of a crime has critical crime knowledge that only the perpetrator should know (e.g., what the murder weapon was, or where it was hidden; Lykken, 1959; for a review see Verschuere et al., 2011). To examine whether an examinee recognizes crime-related information, critical crime information is presented within a series of similar, crime-unrelated information. If the examinee responds

differently to the crime information than to the crime-unrelated information, recognition is inferred. Measured responses are usually skin conductance, heart rate, and breathing rate but other responses like reaction times and brain activity have also proven to be effective (Meijer et al., 2016). More recently, it has been shown that under specific conditions, eye movements can also be used to detect concealed knowledge (Lancry-Dayan et al., 2018, 2021, 2022). However, this method requires multiple crime-related items, and for each crime-related item, several equally plausible alternatives (for example, in the studies by Lancry-Dayan et al., 2018, 2021, for each crime-related item, 7 control items were used). Finding sufficient crime-related items and formulating enough appropriate control items is often stated as a major challenge of memory detection (Iacono, 2011). As a result, this method may often be difficult to apply in practice.

Previous studies showed that when people observe an image they have seen before, they tend to look more at regions in the image where a modification occurred (e.g., Ryan et al., 2000). The idea that recognition can also be detected by examining eye movements toward a modified version of a known item was applied to a forensic setting by Derrick et al. (2011). In their study, 11 participants had to try to pass a screening station. Before the screening interview, half of them had assembled a realistic bomb. In the screening interview, the participants had to answer some questions and were shown three images while their eye movements were tracked. Critically, one of the pictures was a modified version of the bomb. The authors found that the bomb-making group fixated more on the manipulated part of the bomb image than the innocent group. One advantage of this method is that only a picture of the crime related item is needed and no control items have to be formulated. While the exploratory findings of Derrick are promising, they await confirmation. Moreover, recent technological advances make it possible to enhance recognition by presenting the crime scene in virtual reality (Norman et al., 2020).

In the current study, we will examine whether crime knowledge can be detected based on eye movements toward modified scene parts, presented in virtual reality. To that end, participants were randomly assigned to either the guilty condition (who had to commit a mock crime) or the innocent condition (who had to do an unrelated task). Upon return, all participants were informed that there had been a crime on campus and that they were suspected of committing this crime. Then, all participants viewed the crime scene in virtual reality to which we made several salient modifications as compared to the original crime scene. Based on Derrick et al. (2011), we hypothesized that the eyes of guilty participants would be drawn toward these modifications, while innocent participants would deploy fewer fixations to the modified parts of the crime scene.

Method

This study was approved by the ethical committee of the Faculty of Social Sciences of the Hebrew University of Jerusalem and registered as number 2022-02132. The study was preregistered (https://osf.io/ybg4q), and no deviations were made to the preregistration. Materials and data can be found at https://osf.io/pr79y/.

Participants

Participants were recruited through the online recruitment system of the Hebrew University of Jerusalem and received 2 research credits or 20 shekels for participation. As described in our preregistration, we first opened up the experiment to 40 participants. Then, after applying our preregistered exclusion criteria (see below), we ran a Bayesian one-sided independent samples *t*-test in JASP (version 0.16.2; using JASP's default Cauchy scale of 0.707) with the gaze duration proportion as the dependent variable and condition (guilty vs. innocent) as the independent variable. The decision to stop collecting data was based on the Bayes Factor (BF_{10}) of this test. BF_{10} expresses how likely the data are under the alternative

hypothesis (larger proportion of gaze duration in the guilty compared to the innocent condition on the modified items) as compared to the null hypothesis (there is no difference between the two groups). We stated that if we would reach substantial evidence for either the alternative hypothesis (i.e., BF_{10} larger than 5) or the null hypothesis (i.e., BF_{10} smaller than 1/5), we would stop testing, if not, we would open up the experiment for another 20 participants. We would repeat this procedure until we would have reached substantial evidence (i.e., $BF_{10} > 5$ or $< 1/5$) or until we would have tested N = 100. After our first batch of 40 participants, we reached a BF_{10} of 5.67, so data collection was stopped.

Based on the preregistered exclusion criteria, 7 guilty participants were excluded because they did not commit the crime properly (i.e., they could not give the jewels and credit card back at the end of the experiment). The final sample consisted of 33 participants, of which 12 were guilty and 21 innocent. Sixty-seven percent of the sample was female, 30% was male, and 3% defined themselves as another gender or non-binary. They had a mean age of 24 years (*SD* = 2 years, range 19-29 years).

Procedure

The study consisted of 3 parts: A mock crime/control task, a virtual reality task, and a questionnaire.

In the first part of the experiment, participants were randomly assigned to either the guilty condition or the innocent condition. To this end, the experimenter gave the first of a stack of 10 shuffled envelopes (5 envelopes for guilty and 5 for innocent) which contained condition-specific instructions. This way, the experimenter could remain blind to the condition of the participant. Participants who received guilty instructions were instructed to steal some jewels and a credit card from a professor's lab on campus. Participants who received innocent instructions were instructed to read some posters on campus for 10 min.

Instructions can be found in full at https://osf.io/gpkdb. Upon return, all participants were informed that there had been a crime on campus and that they were suspects of this crime.

In the second part of the experiment, participants did a virtual reality task. Participants were told that they would see two rooms in virtual reality and that they had to try to remember as much as possible from each room because there would be a memory task afterward. Critically, one of the two rooms was a modified version of the crime scene (the professor's lab; see Figure 1). Six modifications were made to the crime scene: two omissions (a folder and a dartboard were removed), two additions (a lamp and a gift were added), and two changes (an envelope and a backpack were changed to a different envelope and backpack). From each category (omission, addition, and change), one modification was central to the crime, and one was peripheral. Participants saw the crime scene and control room for 30 seconds each. The order of the rooms (first the crime scene or first the control room) was counterbalanced among participants.

Figure 1

Pictures of the crime scene during the mock crime (left) and the two rooms presented during the virtual reality task (top right is the control room, bottom right is the modified crime scene).

Mock Crime **Virtual reality task**

In the third part of the experiment, participants had to complete a short questionnaire. Participants were asked to answer eight multiple-choice recognition questions about the rooms they visited in virtual reality, provide their demographics (age and gender) and provide some feedback on their experience with the experiment.

Results

Confirmatory analyses

For our main analyses, we first computed the average proportion of time where gaze was directed at the modified areas, out of the whole viewing time. Afterward, we tested whether this proportion was higher for guilty participants than for innocent participants by running a one-sided independent sample *t*-test (both frequentists and Bayesian) with the gaze duration proportions as dependent variable and condition (guilty vs. innocent) as independent variable. These analyses showed that guilty participants looked significantly more at the modified areas than innocent participants, $t(31) = 2.43$, $p = .011$, $d = 0.88$, 95% CI [0.249; ∞], $BF_{10} = 5.67$

Exploratory analyses

The area under the ROC curve

To test whether guilty individuals could be distinguished from innocent individuals based on the proportions of gaze duration on the modifications, we calculated the area under the ROC curve. The area under the ROC curve is a measure of the diagnostic accuracy of a test. It ranges from 0.5 to 1, where a value of 0.5 indicates a chance level accuracy and suggests no discrimination, while a value of 1 reflects a perfectly accurate test (Mandrekar, 2010). This analysis showed that it was possible to distinguish guilty participants from innocent ones significantly above chance with an area under the ROC curve of 0.75, 95% CI [.57; .94], p = .018. To define the optimal cutoff point, the Youden index was calculated. The maximum Youden index was obtained for cutoff point 0.039, J_{max} = 0.452. Although this cutoff point corresponded to a high specificity (only 5% false positives), it did also correspond to a low sensitivity (50% false negatives).

Different types of modifications

To explore the influence of the different types of modifications, we ran a 3 (modification type: omission, addition vs. change; within-subjects) by 2 (crime-relatedness: central to the crime vs. peripheral; within-subjects) by 2 (condition: guilty vs. innocent; between-subjects) mixed (Bayesian) ANOVA with gaze duration proportions as dependent variable. The results of interest were the interaction between the condition and modification type and the interaction between condition and crime-relatedness. The interaction between condition and modification type was not significant, $F(2, 62) = 0.74$, p = .480, $\eta^2 = 0.01$. The interaction between condition and crime-relatedness was also not significant, $F(1, 31) = 0.01$, p = .926, $\eta^2 = 0.00$. The Bayesian ANOVA indicated that the data are 7.68 times more likely under the model that included the two main effects modification type and condition and the interaction between modification type and crime-relatedness, as compared to the null model.

Crime-related areas without modifications

There were two crime-related areas in the virtual reality crime scene to which we made no modifications. There was no evidence that guilty participants looked more at these areas than innocent participants, $t(31) = -0.56$, $p = .711$, $d = -0.20$, 95% CI [-0.80; ∞], $BF_{10} = 0.24$.

Recognition questions

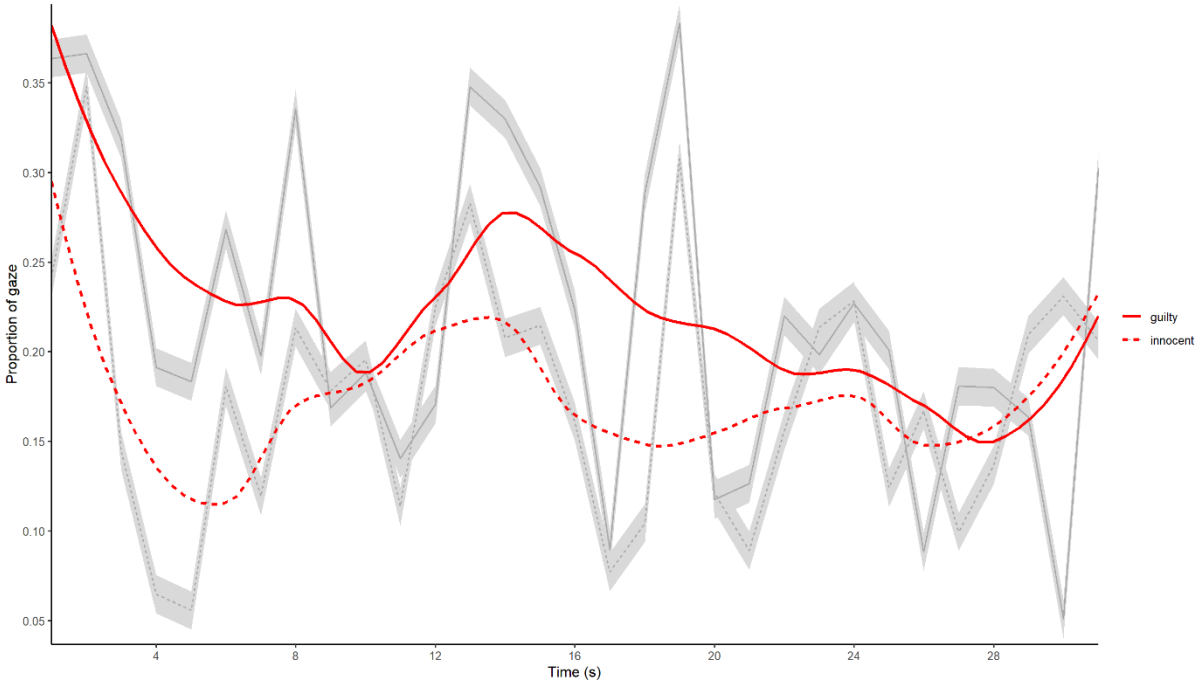
Participants answered on average 6.61 ($SD = 0.86$) out of 8 recognition questions correctly. Guilty participants ($M = 6.67$, $SD = 0.49$) scored slightly better than innocent participants ($M = 6.57$, $SD = 1.03$), but this difference was not significant, $t(31) = 0.30$, $p = .766$, $d = 0.11$, 95% CI [-0.60; 0.82], $BF_{10} = 0.35$.

Other dependent variables

Time to first fixation. We tested whether the time until participants looked at modifications was shorter for guilty participants than for innocent participants. To test this, we ran a (Bayesian) one-sided independent sample t-test with time to first fixation as dependent variable and condition (guilty vs. innocent) as independent variable. This analysis showed that guilty participants looked significantly earlier at the modifications than innocent participants, $t(31) = -2.77$, $p = .005$, $d = -1.00$, 95% CI [-∞; -0.36], $BF_{10} = 10.46$. The time course of when guilty versus innocent participants looked at the modifications is visualized in Figure 1. This figure illustrates that the majority of the increased attention toward modifications occurs within the initial 10 seconds, and then gradually decreases during the next 10 seconds.

Figure 1

Time course of gaze towards modifications by guilty participants and innocent participants.



Note. In grey, the gaze proportions in each time bin with their standard errors are depicted. In red, smoothed trend lines are added for clarity purposes.

Fixation duration. We tested whether the fixation durations on the modifications were longer for guilty than for innocent participants through a (Bayesian) one-sided independent sample t-test with fixation duration as dependent variable and condition (guilty vs. innocent) as independent variable. Guilty participants did not exhibit longer fixations at the modifications than innocent participants, $t(30) = 0.51$, $p = .306$, $d = 0.19$, 95% CI [-0.42; ∞], $BF_{10} = 0.51$.

Discussion

In the current study, we tested whether eye movements toward modified areas of a crime scene presented in virtual reality allow to detect crime knowledge. As predicted, we found that guilty participants looked more at the modified areas than innocent participants.

Furthermore, exploratory analyses showed that it is possible to distinguish guilty participants from innocent ones based on their eye gaze behavior and that guilty participants also looked earlier at the modified areas.

Eye movements serve as a means of gathering information from the environment to form a coherent visual representation of what we are looking at. In doing so, people look more at objects that are the targets of a search, but also dwell on objects that differ from expectations. These expectations can be based on general knowledge about the context in which they are embedded (e.g., an octopus on a farm; Loftus & Mackworth, 1978), or, as in the current study, on previous viewing history. These expectation-based biases of attention toward areas of interest facilitate efficient processing of the environment (Hannula et al., 2010). They allow us to quickly compare the current input with past experiences in order to update the visual representation, guide decision-making, and navigate through the world (Ryan & Cohen, 2004).

Advantages over classical methods of memory detection

The current results not only demonstrate the potential of memory detection based on eye movements towards modifications, but they also imply that this approach may have some important advantages over classical memory detection methods.

Although the classical method of memory detection has already been validated in many studies and obtains high accuracy rates (Meijer et al., 2014; Verschuere et al., 2011), it has been rarely used in practice (with Japan being the only exception; Hira & Furumitsu, 2002; Osugi, 2011). Some of the most important reasons for its limited use are the information leakage problem and the risk of countermeasures. In the next paragraphs, we argue how the current method might provide a solution to these problems.

First, this method could provide a solution to the leakage problem. The leakage problem refers to the fact that innocent suspects could gain crime knowledge through means

other than committing the crime, leading to false positive results (Bradley et al., 2011). If a suspect heard or read about certain items of a crime scene, then using these items could result in false positive outcomes, so they become useless. For example, if it is known to the public that a crime happened in the living room of the suspect, then this item can no longer be used in a classic memory detection test anymore. However, if the investigators would make certain modifications to the particular living room, then innocent suspects who only heard where the crime took place, but have never seen the living room, would not notice the changes while guilty suspects would. So even if information has been leaked, the researchers are in control because they can decide which changes should be made to the critical item that might have been leaked. Furthermore, as in the current study, the test could be used to detect recognition of a crime scene, and leakage of an entire crime scene is highly unlikely in practice.

Second, this method could bypass countermeasures. Countermeasures are techniques that a guilty suspect could use to control their responses in order to beat the test (i.e., to look innocent), leading to false negative results (Ben-Shakhar, 2011). One of the main advantages of using eye movements as a measurement is that it does not give away what is being measured. If participants are not aware of what is being tested, there would be no need for faking. Furthermore, faking under the proposed method is particularly unlikely because suspects may not be explicitly aware of the specific changes in the crime scene. Even if participants would know what is being tested, research has shown that eye movements could be attracted to changes even without awareness of the change (Ryan et al., 2000).

Limitations and future directions

While we lack a direct comparison with established measures, the detection efficiency of the current method seems lower than what is usually found in previous research. A recent review by Lancry-Dayan et al. (2022) showed that autonomic nervous system responses (skin conductance, heart rate, and breathing rate) result on average in areas under the ROC curve of

0.81, neural activity leads to a mean detection efficiency of 0.91, and eye tracking paradigms that include and a simultaneous display elicit a mean detection efficiency of 0.86. These results are all well above the detection accuracy of 0.75 that was found in the current study. However, previous studies primarily used highly familiar items, whereas the present study used newly learned items, which have been shown to have lower detection efficiency (Lancry-Dayan et al., 2021; Millen et al., 2020). Moreover, unlike previous studies, the current study employed only one crime-related item (the crime scene). Future research could investigate whether utilizing multiple 3D pictures could enhance the detection efficiency of the current method.

There are also several questions that remain unanswered. For example, as guilty participants in the current study did the virtual reality task immediately after they performed the mock crime, their memories were strong. Future studies could examine whether the effect lasts over time when information is forgotten. Moreover, it is unknown how well the 3D pictures should match the crime scene to register appropriate responses. For example, in real-life applications, it could be possible that the 3D picture would be taken only weeks or even months after the crime took place and the crime scene may have changed in the meantime. The 3D picture could also be taken from a different point of view from the one the criminal had while committing the crime. Future studies could assess whether the effects are robust to these kinds of variations, and if so, to what extent.

Conclusion

In conclusion, when visiting a modified crime scene in virtual reality, guilty participants looked at the modifications earlier and longer, compared to innocent participants. Based on their eye gaze behavior, it was possible to significantly distinguish the guilty from the innocent participants. This approach has several important advantages in comparison to traditional memory detection, although its detection efficiency seems to be lower.

References

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: a reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 997–1010. doi:10.1037/0278-7393.25.4.997

Ben-Shakhar, G. (2011). Countermeasures. *Memory detection: Theory and application of the Concealed Information Test*, 200-214. doi:10.1017/cbo9780511975196.012

Bradley, M. T., Barefoot, C. A., & Arsenault, A. M. (2011). Leakage of information to innocent suspects. *Memory detection: Theory and application of the Concealed Information Test*, 187-199. doi:10.1017/cbo9780511975196.011

Derrick, D. C., Moffitt, K., & Nunamaker, J. F., Jr. (2011). Eye gaze behavior as a guilty knowledge test: Initial exploration for use in automated, kiosk-based screening. In *Hawaii International Conference on System Sciences* (pp. 4-7).

Hannula, D. E., Althoff, R. R., Warren, D. E., Riggs, L., Cohen, N. J., & Ryan, J. D. (2010). Worth a glance: using eye movements to investigate the cognitive neuroscience of memory. *Frontiers in human neuroscience, 4*, 166. doi:10.3389/fnhum.2010.00166

Heisz, J. J., & Shore, D. I. (2008). More efficient scanning for familiar faces. *Journal of Vision, 7*(9), 24. doi:10.1167/8.1.9

Hira, S., & Furumitsu, I. (2002). Polygraphic examinations in Japan: Application of the guilty knowledge test in forensic investigations. *International Journal of Police Science & Management, 4*(1), 16-27. doi:10.1177/146135570200400103

Iacono, W. G. (2011). Encouraging the use of the Guilty Knowledge Test (GKT): what the GKT has to offer law enforcement. *Memory detection: Theory and application of the Concealed Information Test*, 12-23. doi:10.1017/cbo9780511975196.002

Lancry-Dayan, O. C., Ben-Shakhar, G., & Pertzov, Y. (2022). The promise of eye-tracking in the detection of concealed memories. *Trends in Cognitive Sciences.* doi:

10.1016/j.tics.2022.08.019

Lancry-Dayan, O. C., Nahari, T., Ben-Shakhar, G., & Pertzov, Y. (2018). Do you know him? Gaze dynamics toward familiar faces on a concealed information test. *Journal of Applied Research in Memory and Cognition, 7*(2), 291-302. doi:10.1016/j.jarmac.2018.01.011

Lancry-Dayan, O. C., Nahari, T., Ben-Shakhar, G., & Pertzov, Y. (2021). Keep an eye on your belongings: Gaze dynamics toward familiar and unfamiliar objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(11), 1888-1901. doi:10.1037/xlm0001086.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human perception and performance, 4*(4), 565. doi:10.1037/0096-1523.4.4.565

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*(6), 385. doi:10.1037/h0046060

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology, 5*(9), 1315-1316. doi:10.1097/jto.0b013e3181ec173d

Mehoudar, E., Arizpe, J., Baker, C. I., & Yovel, G. (2014). Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. *Journal of vision, 14*(7), 6. doi:10.1167/14.7.6

Meijer, E. H., klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta-analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51*(9), 879-904. doi:10.1111/psyp.12239

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology, 53*(5), 593-

604. doi:10.1111/psyp.12609

Millen, A. E., Hope, L., & Hillstrom, A. P. (2020). Eye spy a liar: assessing the utility of eye fixations and confidence judgments for detecting concealed recognition of faces, scenes and objects. *Cognitive Research: Principles and Implications, 5*(1), 1-18. doi:10.1186/s41235-020-00227-4

Norman, D. G., Wade, K. A., Williams, M. A., & Watson, D. G. (2020). Caught virtually lying—Crime scenes in virtual reality help to expose suspects' concealed recognition. *Journal of Applied Research in Memory and Cognition, 9*(1), 118-127. doi:10.1016/j.jarmac.2019.12.008

Osugi, A. (2011). Daily application of the Concealed Information Test: Japan. *Memory detection: Theory and application of the Concealed Information Test,* 253–275. doi:10.1017/cbo9780511975196.015

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological Science, 11*(6), 454–461. doi:10.1111/1467-9280.00288

Ryan, J. D., & Cohen, N. J. (2004). The nature of change detection and online representations of scenes. *Journal of Experimental Psychology: Human Perception and Performance, 30*(5), 988. doi:10.1037/0096-1523.30.5.988

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision, 11*(5), 5. doi:10.1167/11.5.5

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (Eds.). (2011). *Memory detection: Theory and application of the Concealed Information Test.* Cambridge University Press. doi:10.1017/cbo9780511975196

Yarbus, A. L. (1967). *Eye movements and vision.* doi:10.1007/978-1-4899-5379-7

6

General discussion

Memory detection is a valuable tool for uncovering whether a criminal suspect possesses crucial knowledge about a crime, even if they deny it. Despite extensive research over the past few decades, the leakage problem, the countermeasures problem, the difficulty in selecting suitable items, and the limited applicability of the test have hindered the widespread adoption of this technique. The studies of the current dissertation explored the use of different modalities, measurements, and techniques to detect crime-related knowledge in an effort to address these challenges and improve the effectiveness of the method.

Summary of the main findings

In Chapter 2, the question of whether items are better presented pictorially or verbally in memory detection was addressed. To investigate this, participants were asked to encode verbal (e.g., the word “knife”) and pictorial (e.g., an image of a knife) crime-related items and hide their knowledge in a memory detection test consisting of a verbal and a pictorial part. The results showed that detection was more successful when items were encoded and tested in the same modality. Furthermore, recognition of verbally encoded items could not be detected in a pictorial test. These results highlight the importance of matching the modality of test items to the modality in which crime-related information was encoded to improve detection efficiency. Moreover, using a pictorial test could help prevent false accusations against suspects who only heard or read about the crime but did not enact it. This way, the leakage problem could be circumvented.

Chapter 3 delved into the new development of a memory detection technique based on eye movements, testing it with a mock crime design. Additionally, the study was the first to address whether the promising eye-tracking method could provide a solution for the leakage problem (e.g., Bradley et al., 2011). Participants were instructed to either commit a mock crime on campus (guilty participants), read an article about the crime (informed innocent

participants), or read an unrelated article (naïve innocent participants). Afterward, all participants were presented with an eye tracking memory detection task. Results indicated that it was possible to distinguish the guilty participants from the naïve innocents. Interestingly, it was also possible to distinguish the guilty participants from the informed innocent ones, bypassing the leakage problem. While the results are promising, the detection efficiency observed in this study was lower than in previous memory detection studies.

In Chapter 4, as one of the major promises of memory detection based on eye movements lies in the possibility of conducting tests remotely and without a laboratory setup, I took a first step in assessing the validity of webcam-based eye tracking. In the study, three robust eye tracking paradigms -the cascade effect, the novelty preference, and the visual world paradigm- were replicated online using the participant's own webcam as an eye tracker. All three studies were replicated successfully, although the effect sizes were reduced by 20 to 27%. These findings indicate that online webcam-based eye tracking is feasible but also noisier than traditional laboratory-based eye tracking. However, as it has some important advantages as compared to traditional eye tracking, it could be preferable in certain cases. Based on these results, I suggest that eye tracking studies with relatively large effects that do not require extremely high eye tracking precision could be conducted online using participants' webcam as eye tracker.

Chapter 5 introduced a novel approach for detecting crime knowledge by examining eye movements toward modified areas of a crime scene. After committing either a mock crime or an unrelated task, guilty and innocent participants viewed the crime scene (taken with a 360° camera) in virtual reality (VR) to which several salient modifications were made. The study found that guilty participants looked more and earlier at these modifications as compared to innocent participants and that it was possible to significantly distinguish them from each other based on their eye movements. This new method has various advantages as

compared to classic memory detection. Primary because it could be less susceptible to both the leakage and the countermeasures problems. In addition, as it is based on a presentation of the entire crime scene, it does not require control items. However, the detection efficiency in the current design does seem lower than classic memory detection methods.

Practical and theoretical implications

Implications in the field of memory detection

The current dissertation has some important implications for the field of memory detection. Examiners should always try to match the modality of the test items to the modality in which the crime information was encoded, which is mostly pictorial in nature. This way, the detection efficiency of the test could be maximized. In cases where a laboratory setup or direct contact with the examinee would not be possible, memory detection based on eye movements could be considered. Using this method with a webcam will enable large-scale memory detection programs where groups of people could be tested in a quick way (Kleinberg et al., 2019). VR-based memory detection could be used in cases where investigators have only limited information about a crime (e.g., if they do not know the murder weapon or how the perpetrator fled the scene). In these cases, a classic memory detection test would not be possible. However, using a modified 3D image of the crime scene would allow the identification of suspects with crime knowledge.

Overall, these studies have the potential to improve memory detection in several ways. First, they suggest new ways to overcome the leakage problem (Chapters 2, 3, and 5). By using a pictorial test when leakage happened verbally, by using eye movement-based memory detection (as it was possible to distinguish informed innocents from guilty participants), or by using VR-based memory detection with a modified crime scene (where leakage is less likely), innocent suspects could possibly be protected from being falsely accused. Second, these

methods could provide a solution to the countermeasures problem (Chapters 4 and 5). It has been shown that eye movements can happen outside one's awareness, making them potentially more resistant to countermeasures (Lancry-Dayan et al., 2022; Ryan et al., 2000). Third, they could resolve the problem of the difficulty in selecting suitable control items (Chapter 5). The VR-based memory detection paradigm requires only a 360° picture of the crime scene, eliminating the need for control items. Last, these methods could be applicable to more settings than traditional memory detection (Chapters 2, 3, and 4). No lab, experienced examiner, or dedicated and expensive hardware is needed for these methods, which highly increases the test's applicability. In sum, the four chapters contribute to the growing body of research on memory detection and demonstrate the value of exploring different methods and measures for detecting crime knowledge.

Implications for the broader scientific community

Besides the field of memory detection, the current findings also have some implications for the broader scientific community. Chapter 2 contributes to the growing research on the picture superiority effect. There are two competing theories explaining pictures' mnemonic advantage over words. First, there are those attributing it to a perceptual processing advantage. Whereas pictures can be encoded through both the verbal and the nonverbal memory pathway, words can only be encoded through the verbal pathway (the dual-code theory; Paivio, 1971). Second, there are those attributing it to a conceptual processing advantage. Cognitive operations performed during encoding are reinstated during retrieval. Since a typical memory test requires the examinee to retrieve information visually, pictures are at advantage (the transfer-appropriate processing theory; Rugg et al., 2015). The results reported in Chapter 2 support both theories and suggest the two might not be mutually exclusive. As predicted by the dual code theory, items that were encoded pictorially were easier to detect than items that were encoded verbally. Furthermore, no conversion across

modalities was observed when items were encoded as words. As predicted by the transfer-appropriate processing theory, recognition effects were strongest when encoding and test modality were matched.

Chapters 3 and 5 contribute to the literature on the link between eye movements and memory. In these studies, guilty participants looked more at the crime-related items (Chapter 3) and the modifications (Chapter 5). This is in line with the orienting response theory according to which people look more at objects that are significant to them or differ from their expectations (Sokolov, 1990). This behavior is essential for adapting to a constantly changing environment, as it enables us to quickly make decisions and navigate through the world (Hannula et al., 2010). These responses are not only visible in eye-movements, but also in other physiological and behavioral responses (Lynn, 1966).

Chapter 4 has several important practical implications for the wider scientific community as it demonstrates the feasibility of eye tracking research to move beyond the confined lab and onto the web. This enables research to be conducted considerably faster, easier, and cheaper without the need for a lab, experimenter, or dedicated hardware. Moving online also opens up the possibility of recruiting a larger and more diverse participant pool, including even those who are hard to reach. The collection of data would no longer be limited by time or location as individuals could participate whenever they want, wherever they want.

Limitations and future directions

Lower detection efficiency

One of the major challenges of the methods presented in the current dissertation is that their detection efficiency seems generally lower than what is usually found with classical memory detection methods. For example, in Chapter 3, eye movement-based memory detection resulted in a detection efficiency of 0.71. In Chapter 5, VR-based memory detection

resulted in a detection efficiency of 0.75. In contrast, a recent review by Lancry-Dayan et al. (2022) showed that physiological measures (skin conductance, heart rate, and respiration) result on average in a detection efficiency of 0.81, neural activity result in a detection efficiency of 0.91, and eye movements (in a parallel design) in a detection efficiency of 0.86. As such, while the new methods presented in the current dissertation offer important advantages as compared to traditional memory detection, there is still much room for improvement in terms of their detection efficiency.

One of the reasons for the lower detection efficiencies found in the current dissertation could be that only newly learned items were used, while the averaged detection efficiency scores of 0.86 included studies in which highly familiar items (e.g., the face of a best friend) were used. Research has shown that newly learned items typically result in lower detection efficiency scores than highly familiar items (Lancry-Dayan et al., 2021; Millen et al., 2017, 2020). Furthermore, in the current studies, participants learned the crime-related items during a mock crime, whereas many other studies included in the average detection efficiency score used a memory game. While memory games guarantee optimal memory, a mock crime does not necessarily result in participants seeing and remembering all relevant information. This poorer memory for crime-related items might have decreased the efficiency scores even further. To compare the different methods more accurately, future studies should conduct a memory detection test while measuring eye movements, physiological measures, and neural measures simultaneously.

Webcam-based memory detection

In this dissertation, I did not use webcam-based eye tracking as a measure for memory detection. Although it would have seemed a logical next step, our studies suggested that it might be too early to combine the two methods. Chapter 3 showed that while memory detection based on eye movements was possible, effect sizes were relatively small. Chapter 4

revealed that large effects and few regions of interest are necessary for valid results. Therefore, I concluded that it might be premature to conduct a memory detection test with a participant's webcam at this point. Nonetheless, I predict that advancements in webcams, cameras, and machine learning will make this a possibility in the near future. For example, new methods that incorporate deep neural networks have been developed for measuring eye position in smartphones (Valliappan et al., 2020).

Moreover, eye tracking is not the only method that is constantly improving. Portable devices can now measure skin conductance, heart rate, breathing rate, and even neural signals (for example by portable EEG or functional near-infrared spectroscopy; Ferrari & Quaresima, 2012; Krigolson et al., 2017). This could make these measures more widely applicable as well. Future research could explore the detection efficiency of these portable devices and how the different measures could be combined to increase overall accuracy and reliability.

Memory detection to search for information

Chapter 5 found that someone with crime knowledge can be detected by taking them back to the crime scene in VR. This approach could possibly also be applied to the searching memory detection test (Breska et al., 2014; Osugi, 2011). For example, if investigators have already identified the perpetrator, they can use this test to detect the critical items (e.g., the location of a bomb). Clearly, this requires additional research.

Conclusion

In four chapters, the current dissertation explored several new methods to improve memory detection. I found that it is possible to detect crime knowledge by using eye movements, either by examining them with respect to crime items presented along with control items or pertaining to modified areas of a crime scene. In the future, these eye movement-based memory detection methods could possibly be used with a webcam, but for

now, webcam-based eye tracking should remain limited to studies with large effect sizes and few areas of interest. In all memory detection tests, the modality of test items should be matched to the modality in which crime information was encoded to improve detection efficiency. The methods explored have the potential to improve memory detection in several ways, working towards a solution to the leakage problem, the countermeasures problem, the difficulty of identifying proper items, and the limited test applicability. The findings have important implications for memory detection in criminal investigations and highlight the need for continued research to optimize its use. The findings also convey novel insights to the basic science literature related to the connection between memory and eye movements, as well as the picture superiority effect. In conclusion, while the idea of mind reading depicted in popular media remains futuristic, research proves that significant steps have already been made in the right direction. As technology continues to advance, we can only look forward to more innovative methods to detect crime knowledge, making it easier than ever to read the criminal mind.

References

Bradley, M. T., Barefoot, C. A., & Arsenault, A. M. (2011). Leakage of information to innocent suspects. *Memory detection: Theory and application of the Concealed Information Test*, 187-199. doi:10.1017/cbo9780511975196.011

Breska, A., Zaidenberg, D., Gronau, N., & Ben-Shakhar, G. (2014). Psychophysiological detection of concealed information shared by groups: An empirical study of the searching CIT. *Journal of Experimental Psychology: Applied, 20*(2), 136-146. doi:10.1037/xap0000015

Ferrari, M., & Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *NeuroImage, 63*(2), 921-935. doi:10.1016/j.neuroimage.2012.03.049

Hannula, D. E., Althoff, R. R., Warren, D. E., Riggs, L., Cohen, N. J., & Ryan, J. D. (2010). Worth a glance: using eye movements to investigate the cognitive neuroscience of memory. *Frontiers in Human Neuroscience, 4*, 166. doi:10.3389/fnhum.2010.00166

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Detecting Deceptive Intentions: Possibilities for Large-Scale Applications. *The Palgrave Handbook of Deceptive Communication*, 403-427. doi:10.1007/978-3-319-96334-1_21

Krigolson, O. E., Williams, C. C., Norton, A., Hassall, C. D., & Colino, F. L. (2017). Choosing MUSE: Validation of a low-cost, portable EEG system for ERP research. *Frontiers in Neuroscience, 11*, 109. doi:10.3389/fnins.2017.00109

Lancry-Dayan, O. C., Ben-Shakhar, G., & Pertzov, Y. (2022). The promise of eye-tracking in the detection of concealed memories. *Trends in Cognitive Sciences, 27*(1), 13-16. doi: 10.1016/j.tics.2022.08.019

Lancry-Dayan, O. C., Nahari, T., Ben-Shakhar, G., & Pertzov, Y. (2021). Keep an eye on your belongings: Gaze dynamics toward familiar and unfamiliar objects. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 47(11), 1888-1901. doi:10.1037/xlm0001086.

Lynn, R. (1966). *The orientation reaction. Attention, Arousal and the Orientation Reaction.* Pergamon, 1-13. doi:10.1016/b978-0-08-011524-5.50006-0

Millen, A. E., Hope, L., & Hillstrom, A. P. (2020). Eye spy a liar: assessing the utility of eye fixations and confidence judgments for detecting concealed recognition of faces, scenes and objects. *Cognitive Research: Principles and Implications, 5*(1), 1-18. doi:10.1186/s41235-020-00227-4

Millen, A. E., Hope, L., Hillstrom, A. P., & Vrij, A. (2017). Tracking the truth: the effect of face familiarity on eye fixations during deception. *Quarterly Journal of Experimental Psychology, 70*(5), 930-943. doi:10.1080/17470218. 2016.1172093.

Osugi, A. (2011). Daily application of the Concealed Information Test: Japan. *Memory Detection: Theory and application of the Concealed Information Test*, 253-275. doi:10.1017/cbo9780511975196.015

Paivio, A. (1971). *Imagery and Verbal Processes* (1st ed.)., New York: Holt, Rinehart and Winston.

Rugg, M. D., Johnson, J. D., & Uncapher, M. R. (2015). Encoding and Retrieval in Episodic Memory. *The Wiley Handbook on the Cognitive Neuroscience of Memory,* 84-107. doi:10.1002/9781118332634.ch5

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological Science, 11*(6), 454-461. doi:10.1111/1467-9280.00288

Sokolov, E. N. (1990). The orienting response, and future directions of its development. *The Pavlovian journal of biological science, 25*, 142-150. doi:10.1007/bf02974268

Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P.,

Shojaeizadeh, M., Guo, L., Kohlhoff, K., & Navalpakkam, V. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, *11*(1), 4553. doi:10.1038/s41467-020-18360-5

English summary

Reading the criminal mind

Exploring novel methods of memory detection

To examine whether a criminal suspect has implicating crime knowledge, recognition of crime-related details can be detected using memory detection methods. Despite decades of research demonstrating its effectiveness, only the Japanese National Police has been using it extensively in criminal investigations. Wider adoption of this method faces several challenges, including the leakage problem, the countermeasures problem, the difficulty in selecting suitable testitems, and the limited applicability of the test. In the current dissertation, various modalities, measurements, and techniques are explored in an effort to address these challenges and improve the effectiveness of the method.

The current dissertation comprises four lines of research. First, I examined which modality (verbal versus pictorial) is optimal for memory detection and whether encoding and testing modalities interact. The results highlight the importance of matching the modality of the test items to the modality in which crime-related information was encoded to improve detection efficiency. Second, I explored the promising new development of memory detection based on eye movements, including its potential of solving the leakage problem. Results indicated that it was possible to distinguish the guilty participants from the naïve innocent participants as well as the informed innocent ones based on their eye movements. Third, as a first step in assessing the potential of webcam-based eye tracking for memory detection, I examined whether a simple webcam could be used as an eye tracker. To this end, I replicated three robust eye tracking studies online using the participant's webcam as an eye tracker. The results suggest that webcam-based eye tracking is feasible, although it is best suited for studies with relatively large effects that do not require very high eye tracking precision. Fourth, I introduced a new approach for detecting crime knowledge by examining whether suspects' familiarity with a crime scene could be detected by bringing them back to a

modified version of it in virtual reality. We tested whether gaze behavior toward the modified parts of the crime scene could reveal its recognition. The study showed that guilty participants looked more and earlier at the modifications than innocent participants, making it possible to significantly distinguish the two groups from each other.

The findings have important practical implications for memory detection as well as theoretical implications for the memory literature. A wider application of memory detection can bring us one step closer to reading the criminal mind.

Dutch summary

Het lezen van de criminele geest

Onderzoek naar nieuwe methoden voor geheugendetectie

Om na te gaan of een verdachte van een misdrijf over daderkennis beschikt, kan herkenning van misdaad gerelateerde informatie worden getest met behulp van geheugendetectiemethoden. Hoewel onderzoek de doeltreffendheid van de methode reeds tientallen jaren heeft aangetoond, gebruikt tot op vandaag alleen de Japanse nationale politie deze methode op grote schaal in strafrechtelijke onderzoeken. Bredere toepassing van de methode kent verschillende uitdagingen, waaronder het lekprobleem, het probleem van tegenmaatregelen, de moeilijkheid om geschikte testitems te selecteren en de beperkte toepasbaarheid van de test. In het huidige proefschrift worden verschillende modaliteiten, metingen en technieken onderzocht in een poging deze uitdagingen aan te pakken en de effectiviteit van de test te verbeteren.

Het huidige proefschrift bestaat uit vier onderzoekslijnen. Ten eerste heb ik onderzocht welke modaliteit (verbaal versus pictoriaal) optimaal is voor geheugendetectie en of codering- en testmodaliteiten op elkaar inwerken. De resultaten benadrukken het belang van het afstemmen van de modaliteit van de testitems op de modaliteit waarin misdaad gerelateerde informatie werd gecodeerd om de efficiëntie van de detectie te verbeteren. Ten tweede onderzocht ik de veelbelovende nieuwe ontwikkeling van geheugendetectie op basis van oogbewegingen, inclusief de mogelijkheid om het lekprobleem op te lossen. De resultaten gaven aan dat het mogelijk was om de schuldige deelnemers te onderscheiden van de naïeve én de geïnformeerde onschuldige deelnemers op basis van hun oogbewegingen. Ten derde, als een eerste stap in het beoordelen van het potentieel van webcam-gebaseerde eye tracking voor geheugen detectie, onderzocht ik of een eenvoudige webcam gebruikt kan worden als eye tracker. Daartoe repliceerde ik drie robuuste eye tracking studies online met de webcam van de proefpersoon als eye tracker. De resultaten suggereren dat webcam-gebaseerde eye

tracking mogelijk is, hoewel het voorlopig het meest geschikt is voor studies met relatief grote effecten die geen zeer hoge precisie vereisen. Ten vierde introduceerde ik een nieuwe aanpak voor het detecteren van daderkennis door te onderzoeken of de herkenning van een plaats delict kon worden gedetecteerd door verdachten naar een aangepaste versie ervan te brengen in virtual reality. We testten of het kijkgedrag naar de aangepaste delen van de plaats delict de herkenning ervan kon onthullen. Uit het onderzoek bleek dat schuldige deelnemers meer en eerder naar de aanpassingen keken dan onschuldige deelnemers, waardoor de twee groepen significant van elkaar konden worden onderscheiden.

De bevindingen hebben belangrijke praktische implicaties voor geheugendetectie alsook theoretische implicaties voor de geheugenliteratuur. Een bredere toepassing van geheugendetectie kan ons een stap dichter brengen bij het lezen van de criminele geest.

Hebrew summary

קריאת המחשבה הפלילית

חקר שיטות חדשות לזיהוי זיכרון

כדי לבחון האם חשוד בפלילים קשור לפשע מסוים, ניתן לבחון זיהוי של פרטים הקשורים לפשע באמצעות שיטות לזיהוי מידע מוכמן. למרות עשרות שנים של מחקר שהראה את יעילות השיטה לזיהוי מידע מוכמן, עד היום נעשה בה שימוש נרחב בחקירות פליליות רק ביפן. אימוץ רחב יותר של שיטה זו מתמודד עם מספר אתגרים, לרבות בעיית דליפת המידע, בעיית אמצעי הנגד, הקושי בבחירת פריטים מתאימים והישימות המוגבלת של הבדיקה. בעבודת הגמר הנוכחית, נבחנים אופנים, מדידות וטכניקות שונות במאמץ להתמודד עם אתגרים אלו ולשפר את יעילות השיטה.

עבודת הגמר הנוכחית מורכבת מארבעה קווי מחקר. ראשית, בדקתי איזו מודאליות אופטימלית לזיהוי זיכרון והאם אופני קידוד המידע משפיעים על אופני הבדיקה האופטימלית. התוצאות מדגישות את החשיבות של התאמת האופנות של פריטי הבדיקה לאופן שבו קודד מידע הקשור לפשע כדי לשפר את יעילות הגילוי. שנית, בדקתי פיתוח חדש של זיהוי זיכרון המבוסס על תנועות עיניים. בנוסף, בדקנו האם שיטה זו עשויה לספק פתרון לבעיית הדליפה. התוצאות הצביעו על כך שניתן להבחין בין משתתפים אשמים לבין משתתפים תמימים וכן לתמימים בעלי מירע על סמך תנועות העיניים שלהם. שלישית, כצעד ראשון בהערכת הפוטנציאל של מעקב עיניים מבוסס מצלמת אינטרנט לזיהוי זיכרון, בדקתי את היכולת של מצלמת אינטרנט פשוטה לשמש למעקב עיניים. לשם כך, ביצעתי רפליקציה לשלושה מחקרים קלסיים עם אפקטים ברורים של תנועות עיניים באמצעות מצלמת הרשת של המשתתף. התוצאות מצביעות על כך שמעקב עיניים מבוסס מצלמת אינטרנט אפשרי, אם כי הוא מתאים ביותר למחקרים עם אפקטים גדולים יחסית שאינם דורשים דיוק גבוה במיוחד בזיהוי מיקום המבט. רביעית, הצגתי גישה חדשה לאיתור מידע מוכמן. בדקתי האם ניתן לזהות את ההיכרות של חשודים עם זירת פשע על ידי החזרתם לזירת הפשע שעברה שינויים במציאות מדומה. בדקתי האם התנהגות המבט כלפי חלקים בזירה שעברו שינוי יכולה לחשוף הכרות מוקדמת עם הזירה. המחקר הראה כי משתתפים אשמים הסתכלו יותר הרבה ויותר מוקדם על השינויים מאשר משתתפים תמימים, מה שאיפשר להבחין באופן משמעותי בין שתי הקבוצות.

לממצאים יש השלכה מעשית חשובה על זיהוי הזיכרון והשלכות תיאורטיות על מחקר זיכרון וקשב. המחקרים האלו מקרבים אותנו צעד אחד לכיוון זיהוי מדוייק של מידע פלילי מוכמן.

Author contributions

Chapter 2: Modality effects in memory detection

This chapter is published as: Van der Cruyssen, I., Regnath, F., Ben-Shakhar, G., Pertzov, Y., & Verschuere, B. (2021). Is a Picture Worth a Thousand Words? Congruency Between Encoding and Testing Improves Detection of Concealed Memories. *Journal of Applied Research in Memory and Cognition, 10*(4), 667-676. doi:10.1016/j.jarmac.2021.03.001

Ine Van der Cruyssen, Gershon Ben-Shakhar, Yoni Pertzov, and Bruno Verschuere developed the study concept. Ine Van der Cruyssen and Franziska Regnath set up the experiments, collected the data, and performed the statistical analyses. Gershon Ben-Shakhar, Yoni Pertzov, and Bruno Verschuere provided critical feedback. All authors contributed to and approved the final version of the manuscript for submission.

This research was financially supported by the Barbara and Morton Mendel Foundation.

Chapter 3: Memory detection based on eye movements

This chapter is submitted for publication as: Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., & Verschuere, B. (2022). *Detecting concealed familiarity using eye movements: The effect of leakage of mock crime details to innocents*. Manuscript submitted for publication.

All authors contributed to developing the study concept and design. Ine Van der Cruyssen programmed the experiment, performed the statistical analyses, and drafted the manuscript. All co-authors provided critical feedback on the manuscript and approved the final version of the manuscript for submission.

This research was financially supported by the Barbara and Morton Mendel Foundation.

Chapter 4: The validation of webcam-based eye tracking

This chapter is submitted for publication as: Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., Guy, N., Cabooter, Q., Gunschera, L. J., & Verschuere, B. (2022). *The validation of online webcam-based eye-tracking: The replication of the cascade effect, the novelty preference, and the visual world paradigm.* Manuscript submitted for publication.

Ine Van der Cruyssen, Gershon Ben-Shakhar, Yoni Pertzov, and Bruno Verschuere developed the study concept. Nitzan Guy programmed the experiments. Ine Van der Cruyssen, Quinn Cabooter, and Lukas Gunschera created the materials, collected the data, and performed the statistical analyses. Gershon Ben-Shakhar, Yoni Pertzov, and Bruno Verschuere provided critical feedback. All authors contributed to and approved the final version of the manuscript for submission.

This research was financially supported by the Barbara and Morton Mendel Foundation.

Chapter 5: Memory detection in virtual reality: The effect of scene modifications

This chapter is in preparation as: Van der Cruyssen, I., Ben-Shakhar, G., Verschuere, B., & Pertzov, Y. (2023). *Memory detection in virtual reality: The effect of scene modifications*. Manuscript in preparation.

All authors contributed to developing the study concept and design. Ine Van der Cruyssen set up the experiment and conducted the statistical analyses. Gershon Ben-Shakhar, Bruno Verschuere, and Yoni Pertzov provided critical feedback. All authors contributed to and approved the final version of the manuscript.

This research was financially supported by the Barbara and Morton Mendel Foundation.

Acknowledgements

No one would think that doing a PhD during a global pandemic would sound like a good idea, but thanks to the following people, it turned out to be a great one.

Thank you Bruno,

For being as approachable as a supervisor gets. It led to the most fascinating and entertaining conversations, which could go from power analyses to surf lessons in Indonesia, and always kept me thinking long after they were over. You always pushed me to keep learning, to keep growing, to stay hungry, and kicked my butt out of the comfort zone every time it was needed. Your excitement and drive always ended up making things bigger and better than I could have hoped for in the first place.

Thank you Yoni,

For creating such a relaxed and positive environment in your lab that made me feel like a member of the team from the day I arrived in Israel. Your down-to-earth attitude and cool ideas helped to elevate and enrich our studies. You did not only support me in my research, but also helped me navigate a new city and were always there to help with any non-work related issues from arranging a visa to lending your camping gear. Your kindness and generosity have made this experience unforgettable.

Thank you Gershon,

For always offering me the warmest encouragements and sincerest praises for my accomplishments. This meant a lot, coming from someone with such widespread knowledge and vast experience as you. But however vast your experience, you were always mindful of everyone's opinion. Your passion for good research practices was contagious during your course, and it felt as a privilege to have you as my supervisor.

Thank you members of the PhD committee,

For the time and effort you invest in reading and assessing my work. I am looking forward to the opportunity to present my findings to you and hear your thoughts during my defense.

Thank you to all the members of lielab and the visual cognition lab,

For the welcoming atmosphere, the interesting times at work and the fun times after work. A special thank you to you, Sera, for being my most consistent colleague throughout those four years – even if it was often only virtual. It was great to be able to discuss doubts, work related questions and -more importantly- non-work related banter. Having someone to share your experiences and walk the same path really is invaluable.

Thank you mom, dad, Bram, and Robbe,

For always being my safe haven. It was the most comforting thing to know that with COVID-19 and constantly changing plans you were always ready to welcome me back with open arms in between living everywhere and nowhere.

Thank you my BRRK-, scouts-, and Tel Aviv-friends,

For making these last four years not only educating but also just so much fun. This is a salute to our enduring, seen-some-shit, done-some-shit kind of unconditional love. I'm so happy that I can always count on you to make me cry with laughter whenever I need it.

And lastly, thank you Robin,

For enjoying it all with me. You really are my biggest smiles, my greatest support, and my happiest memories. Wherever we lived during these last four years, every time you walked

up to me with a bottle of wine and two glasses in your hands, ready to settle for another sunset, I knew I was home.