EYE MOVEMENT EFFECTS IN SIMULATED OBJECT RECOGNITION MEMORY IMPAIRMENT

by

Dmitriy Kazakov

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

In Psychology

at

The University of Wisconsin-Milwaukee

May 2016

ABSTRACT

EYE MOVEMENT EFFECTS IN SIMULATED OBJECT
RECOGNITION MEMORY IMPAIRMENT

by

Dmitriy Kazakov

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professor David C. Osmon, Ph.D., ABPP-CN

Malingering is the purposeful fabrication of symptoms for secondary gain. Memory problems are the most reported symptom, and object recognition tests are often used in clinical settings to evaluate these claims. Past research has shown that eye movements can indirectly index memory, in that greater viewing is directed at studied stimuli 500-750 ms after display onset. The present study evaluated eye movements as a potential method of detecting feigned memory impairment. Forty-eight participants, half simulators, studied standardized images and took a memory test. Several levels of analysis were used to detect broad trends and brief effects. Simulators performed significantly worse on the behavioral task, but also directed less viewing time towards studied stimuli overall and during the first 3s when providing correct responses. On Miss trials, they viewed studied stimuli even less in the last 3s. Although simulators demonstrated the early viewing effect, it occurred slightly later (750-1000 ms). The 250 ms data provided more useful information, as did Hit-Miss difference scores. A behavioral measure (corrected recognition score) emerged as the single best indicator of malingering. However, eyetracking methodology was able to provide five eye movement variables that demonstrated good psychometric properties and provided incremental diagnostic utility allowing for all cases to be correctly classified. Therefore, a multimethod approach proved to be most effective in detecting simulated memory impairment in this sample.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Analysis of Variance (ANOVA)

Area Under the Curve (AUC)

Intertrial Interval (ITI)

Negative Predictive Value (NPV)

Positive Predictive Value (PPV)

Receiver Operating Characteristic (ROC)

Social Security Administration (SSA)

Symptom Validity Test (SVT)

Performance Validity Test (PVT)

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. David Osmon. His continuing support of my graduate study and research is a testament to his patience and many years of experience as a mentor. His guidance has helped me stay on track throughout this entire process. In working closely with him, I have learned much and am becoming a better researcher and clinical neuropsychologist as a result.

Secondly, I would also like to thank Dr. Deborah Hannula, who has been a secondary advisor to me for this project. Dr. Hannula was gracious enough to allow me to use her paradigm, equipment, and lab space for this study. She also provided me with intensive training on using the eyetracker and much-needed support in understanding the methodology underlying her work. Her guidance in this endeavor has been invaluable.

Thirdly, I would like to thank Dr. David Copeland and Dr. Daniel Allen. Dr. David Copeland, my first mentor, is an extremely talented instructor whose love for the field led me to change my major a brief introductory psychology summer course and dive headlong into research. Dr. Daniel Allen was my first mentor in clinical neuropsychology during my undergraduate years. His passion for neuropsychology inspired me to pursue graduate study.

Lastly, I would like to thank my mother, grandmother, and late grandfather for their unwavering support in all areas of my life as I pursue graduate study. Ever since I was a child, they pushed me to never stop learning. Now that I am in graduate school, I see that there is so much to learn, but even more that has not been discovered. I want to make them proud by making significant contributions to the field for as long as I can.

Eye Movement Effects in Simulated Object Recognition Memory Impairment

Psychological assessment is a process that involves the evaluation of several different areas of functioning within a particular context. It relies to a great extent on test data, but should also include personal interviews and collateral information from third parties (e.g., family members, schools, etc.). Evaluations may be requested by the client themselves, an employer, a health provider, or a government agency. In many of these evaluation contexts, the question is whether the patient or client has impairment in one or more functional areas. Determining if such impairment exists involves examining brain-behavior relationships, which is the realm of neuropsychology. As many of these contexts offer some kind of compensation for impairment, the potential for malingering is likely to be high.

Malingering refers to the purposeful fabrication or exaggeration of symptoms for secondary gain. Secondary gain refers to a benefit to be gained from having a particular condition, disability, or diagnosis that has an external motivation (Heilbronner et al., 2009). Primary gain would describe the benefit of going through the evaluation process to relieve a physical or emotional symptom, as there is an internal motivation to alleviate discomfort. Sympathy gained from others by malingering would also fall under primary gain if the goal is to receive attention, but may also result in secondary gain if others choose to provide additional help. Stone and Boone (2007) describe humanity's long history of malingering for various reasons. Most cases of malingering involve monetary gain, access to services, and/or absolvement from responsibility. Monetary gain may include larger lawsuit settlements and receiving a disability pension. Access to services may include receiving medical treatment, subsidized government services, student accommodations, and access to medication (for self-

medicating or for selling). Absolvement from responsibility may include time off from school or work, discharge from military service, or avoiding criminal prosecution. All of these benefits represent incentives for people to engage in malingering in a variety of evaluation contexts.

Determining the validity of obtained test results is of paramount importance. The Neuropsychology Model LCD Taskforce (2011) has published a list of domains that should be evaluated during the course of a complete neuropsychological evaluation. Motivation and effort, the final category on the list, is most relevant to malingering and deception detection. The National Academy of Neuropsychology, one of the major neuropsychological practice organizations, released a position paper (Bush et al., 2005) specifically stating that:

> When the potential for secondary gain increases the incentive for symptom exaggeration or fabrication and/or when neuropsychologists become suspicious of insufficient effort or inaccurate or incomplete reporting, neuropsychologists can, and must, utilize symptom validity tests and procedures to assist in the determination of the validity of the information and test data obtained. (p. 425-426)

The American Academy of Clinical Neuropsychology, another major practice organization, released a consensus statement also supporting the need for detecting non-credible performance whether it be from insufficient effort or response bias in clinical and forensic contexts (Heilbronner et al., 2009).

It should also be noted that malingering may be confused with factitious disorder. Malingering is not categorized as a disorder in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5; American Psychiatric Association, 2013). Additionally, while malingering and factitious disorder both involve deceptive presentation, someone with factitious disorder would engage in this behavior in the absence of an external incentive. For example, their goal is

to adopt a "sick role," rather than to obtain compensation of some sort. The DSM-5 definition offers that a combination of factors may increase the likelihood that the individual may be malingering. These factors are: a medicolegal context, a marked discrepancy between subjective and objective claims, a lack of cooperation during evaluation and treatment plan adherence, and the presence of antisocial personality disorder.

Slick, Sherman, and Iverson (1999) developed a comprehensive set of diagnostic criteria for this kind of presentation, which was termed malingered neurocognitive dysfunction (MND). This approach sought to codify the different ways in which a patient's presentation may suggest the presence of malingering. In essence, these criteria are a checklist of potential discrepancies in the patient's data – the more discrepancies, the more likely that malingering is present. Criterion A requires the presence of a substantial external incentive at the time of the evaluation. Criterion B requires evidence from neuropsychological testing: definite negative response bias; a probable response bias on other kinds of well-validated tests of malingering; discrepancies between obtained and reference group data, observed behavioral presentation, reliable collateral reports, or documented background history. Criterion C requires evidence from self-report: discrepancies between reported and documented history, between self-reported symptoms and known reference group symptoms, between self-reported symptoms and behavioral observations, between self-reports and collateral information, or evidence of malingering through well-validated self-report scales of questionnaires. Criterion D requires that the relevant behaviors from criteria B and C are products of volitional decision making, rather than a known disorder that may account for such behavior.

Slick et al. (1999) propose three levels of MND: definite, probable, and possible. The definite and probable levels differ based on the number and quality of B and C criterion evidence

available. Definite MND requires criteria A and D, as well as a definite negative response bias. Negative response bias refers to statistically significant below-chance performance on at least one forced-choice cognitive measure. Probable MND requires criteria A and D, as well as either a minimum of two pieces of criterion B evidence or one from criterion B and at least one from Criterion C. Possible MND is used for the weaker cases for malingering given available information. This category requires criteria A and D and at least one piece of evidence from criterion C. Alternatively, one may meet criteria for definite or probable MND with the exception of criterion D, meaning that the deceptive behavior is fully accounted for by a psychiatric, neurological, or developmental factor. Slick et al. (1999) have placed a premium on negative response bias and results from tests specifically designed to detect malingering, rather than on individual performance that significantly deviates from known patterns of brain functioning. Here, "known patterns of brain functioning" refers to performance profiles obtained from reference groups having distinct conditions that exhibit a measurable impairment in one or more cognitive abilities. Additionally, evidence from discrepant patterns of performance in obtained test data is valued over discrepancies in self-reported symptoms.

Despite the apparent standard practice of including tests of malingering as part of an evaluation, this decision has its detractors. The Office of the Inspector General (2013) recently published a congressional response report regarding symptom and performance validity testing in Social Security Administration (SSA) disability determinations. Each U.S. state may have its own disability determination services and SSA allows them, or an administrative law judge overseeing the case, to purchase consultative examinations. These may include medical exams or procedures (e.g., X-rays), laboratory tests (e.g., blood, urine), or psychological testing. SSA does not allow for the purchasing of symptom and performance validity tests (SVTs and PVTs,

respectively) and has done so since the early 1990s. It was claimed that current S/PVTs were psychometrically flawed and would not contribute to a disability determination, and therefore, not worth the cost. Specifically, it was claimed that "there is no test that, when passed or failed, conclusively determines the presence of inaccurate self-reporting" (p. 2).

However, at the same time, SSA does state that symptom validity tests could provide useful information and that all relevant evidence (including medical evidence) in a case record should be considered. This also includes statements made by claimants about their symptoms or impairment, given that their statements are credible. Given this information, it may be concluded that the SSA acknowledges that S/PVTs may provide useful credibility information to SSA in order to make a claim determination, but that it is only willing to consider results from S/PVTs that it did not have to purchase. In the same document, the Office of the Inspector General's (2013) review of the extent of S/PVT use by other agencies indicates that they are indeed used to a great degree. In addition to position statements by major neuropsychological organizations (Bush et al., 2005; Heilbronner et al., 2009), private insurance companies allow the purchase of S/PVTs and seem to consider that they provide useful information for decision-making. It should be noted that the result from any S/PVT alone may not be conclusive proof of malingering, but that the result must be considered along with obtained test data, observations, and collateral information.

This policy creates an opportunity for arguments to be made about ethical and financial obligations. As of December 2011, SSA pays over $169 billion to approximately 16.9 million recipients each year, which includes Supplemental Security Income and Social Security Disability Insurance recipients with a range of medical and/or psychological conditions (Office of the Inspector General, 2013). Chafetz and Underhill (2013) estimated that $20.02 billion in

2011 was disbursed to claimants (excluding spouse and child beneficiaries, as well as SSI beneficiaries over age 65) who were malingering a mental disorder, given a base rate of 40% (Larrabee, Millis, & Meyers, 2009). This is significant because reported base rates of malingering from practicing clinical neuropsychologists for criminal and civil cases are two and three times greater than for cases with no clear financial component, respectively (Mittenberg, Patton, Canyock, & Condit, 2002).

The ethical guidelines put forth by the American Psychological Association (2010) provide five principles that seek to guide decision-making when we are functioning in professional capacities. Principle A states that we must take steps to benefit those we work with and do no harm to them. Principle B states that we have scientific and professional responsibilities to the community and to society as a whole, and that we form trusting relationships with those we work with. Principle C states that we promote truth in psychology and do not participate in misrepresentations of fact. Principle D states that we must work for justice, equality, and fairness so that all can gain from psychological services. Principle E states that we must respect everyone's individual rights and dignity to allow them to make their own decisions.

All five principles are in effect when the question of malingering is being considered, but SSA's guidelines seem to make it difficult to adhere to the ethics code. A psychologist's goal when completing an evaluation is to use their expertise and the best scientific assessments available to arrive at the most valid diagnosis. When a patient is being straightforward and adequately applying themselves on all assessments, a valid diagnosis is much easier to achieve. Based on the principles, when patients are actively trying to manipulate the system for personal gain, it is our responsibility to identify them and ensure that they do not receive undue benefits or

compensation. If a malingerer is not identified, SSA (or other funding source) is spending money on this individual that should instead be used to pay for treatment or compensation for a claimant with legitimate impairment.

Symptom and Performance Validity in Psychological Assessment

Malingering has been the subject of a large amount of scientific research in the past several decades. The goal of this kind of research is to determine the most valid and reliable ways of detecting malingering. The term *malingering* has traditionally implied some sort of symptom exaggeration or fabrication. Therefore, specific tests or methodologies that have been developed to detect malingering were named *symptom validity tests* (SVTs). Symptom invalidity, however, is not the only reason for noncredible performance. Larrabee (2011) has proposed that these tests be renamed *performance validity tests* (PVTs). Some SVTs are designed to be relatively easy to pass, even for those with legitimate and severe neurological conditions. Others use a forced-choice presentation model, wherein respondent performance is compared against chance. This is also why test results of all psychological tests, not just SVTs or PVTs, are compared against known reference groups, so as to determine if the patient in question is displaying a pattern of symptom reporting or test performance that departs from known patterns of brain functioning as evident from neuropsychological testing (Slick et al., 1999).

Larrabee (2012) suggests that even if severely impaired groups are able to pass such a test, then it requires but a small amount of neurocognitive capacity and the amount of effort exerted in the activity is rather small. Therefore, the argument is that anyone who fails a PVT is not exerting optimal, or even reasonable, amounts of effort. Larrabee (2012) has therefore spearheaded a movement to repurpose *symptom validity* as the accuracy of a self-reported symptomatic complaint. *Performance validity*, therefore, will be used to refer to the validity of

performance on an actual ability task. This distinction is important because tests measure behavioral output, which we infer as the result of cognitive processes. The cognitive processes being tested may very well be influenced by effort, among other things. Because we can only ever measure output using these tests, a test score that was produced with low effort may appear impaired (Millis, 2009). This is especially true on tests that were not designed specifically for the purpose of detecting deception, such as intelligence tests; suboptimal effort would result in the illusion of impaired performance and would possibly lead to a diagnosis of intellectual disability.

Good effort, on the other hand, does not mean obtaining a better score. Instead, it means that the obtained score is most indicative of the person's ability at the time. PVTs are necessary, then, to determine if obtained test scores are valid, or can be interpreted for the purposes of diagnosis or treatment. Bigler (2012), however, suggests that effort itself is a construct muddled by secondary influences, such as stress, anxiety, and depression. For example, mental flexibility, working memory, processing speed, new learning, and semantic fluency are among the most commonly affected abilities in depression (Basso, Miller, Estevis, & Combs, 2013), while attentional processes are significantly affected in patients suffering from anxiety (Clarke & Macleod, 2013). Because of the wide range of secondary influences and demographic variables to consider, it is highly important to select the best, but also most appropriate tests for detecting suspect effort and symptom invalidity (Ziegler & Boone, 2013).

A larger problem that Bigler (2012) identifies is that SVTs and PVTs are not infallible. Much like how actors practice their lines and actions to get into character, a disability claimant (or other individual with incentive) may do research about or be coached about what credible performance for a particular condition looks like or how a clinical population actually presents. Youngjohn (1995) published an account of coaching and information being provided to a client

by his attorney prior to a neuropsychological evaluation. Lawyers, although not specialists in medicine or psychology, can provide impaired test profiles similar to those provided by physicians if provided with relevant literature (Schwartz, Gramling, Kerr, & Morin, 1998). Clients with this advice could then potentially appear impaired with some success when under scrutiny. While raw test data cannot be presented in court due to copyright, lawyers are not barred from accessing scientific literature and from learning about specific tests from publicly available test interpretation books in order to assist with their cases. Therefore, even though research simulators provided with information could be identified when compared to legitimate brain injury profiles (Schwartz, Gramling, Kerr, & Morin, 1998), a lawyer knowledgeable in how clinical disorders truly present can help improve their client's chances of avoiding detection. Certainly, the influence of coaching and information is an important consideration in malingering detection and research for the clinical and forensic arenas (Lees-Haley, 1997).

In a study of simulators, it was found that some amount of preparation prior to testing is common and that the most common strategies were memory loss and slower reaction time (Tan, Slick, Strauss, & Hultsch, 2002). Memory loss is the favored malingering strategy for several groups, including inpatients and incarcerated individuals (Iverson, 1995). Simulators also find it difficult to maintain their chosen strategy due to issues with maintaining concentration or consistency, inhibition, or embarrassment (Tan et al., 2002). This suggests that malingerers may not adhere to the same strategy throughout the duration of the evaluation, or may adjust the amount of effort exerted based on the perceived importance of a test.

Coached simulators provide above-chance performances on validity tests which, while still below threshold for optimal effort, are nonetheless different from below-chance performances provided by uncoached simulators (Martin, Bolter, Todd, Gouvier, & Niccolls,

1993). They are also less likely to be identified than uninformed simulators (Rose, Hall, Szalda-Petree, & Bach, 1998). Even so, coached simulators still perform better than those who were not coached, even if they received information (Dunn, Shear, Howe, & Ris, 2003). This presents a practical problem because not all psychologists utilize the same batteries or even individual tests. While some tests are better at detection than others, the selection criteria for which specific tests are used in an evaluation are numerous. They include test availability, cost per administration, training, equipment, recommendation from fellow practitioners, use in scientific literature, preference for particular developers or publishers, and information from professional conferences. Therefore, a malingerer may have a better chance of evading detection depending on the cutoff scores of the specific tests administered to them. Administering multiple symptom and performance validity tests is recommended in order to have the best chances of identifying malingerers through convergent measures (Gorny & Merten, 2005). This approach is somewhat of a double-edged sword: using several can provide incremental diagnostic utility, but not all PVTs are equivalent in design and psychometrics, making it difficult to compare across results.

DenBoer and Hall (2007) found that only 32% of participants who were asked to simulate traumatic brain injury symptoms were successful in providing adequate effort. However, 76% of the successful participants were coached on actual symptom presentation and behaviors to avoid during testing. Simulators who received coaching on detection avoidance strategies were harder to detect than those provided with accurate symptom presentations or those receiving no coaching (Rüsseler, Brett, Klaue, Sailer, & Münte, 2008). In fact, a simple warning that efforts to malinger may be detectable led to fewer simulators being correctly identified (Suhr & Gunstad, 2000). Gorny and Merten (2005) found that a combination of providing symptom information, warning about malingering detection, and explaining SVT principles led to greater pass rates.

These highly informed simulators were also better able than other groups to identify more SVTs and PVTs in the battery. This suggests that well-informed litigants can be taught to identify malingering tests, provide optimal effort on them, but then malinger on standard non-malingering neuropsychological tests.

Despite the recent advances in symptom and performance validity test development, no single test is perfect and the influence of preparation can affect detection accuracy on a wide variety of tasks (Coleman, Rapport, Millis, Ricker, & Farchione, 1998; DiCarlo, Gfeller, & Oliveri, 2000; Lamb, Berry, Wetter, & Baer, 1994; Rapport, Farchione, Coleman, & Axelrod, 1998). Simulators, regardless of the amount of provided coaching or information, also have a tendency to respond slower (Rose, Hall, & Szalda-Petree, 1995; Tan, Slick, Strauss, & Hultsch, 2002). Dunn et al. (2003) replicated this finding and found no difference between different coaching and information simulator groups. It was noted that simulators also had greater variability in response times, while participants asked to do their best evidenced very little variability. Known base rates are concerning, especially given the ability of healthy volunteers to pass an effort test administered in a language other than their own (Richman et al., 2006). Therefore, Dunn et al. (2003) suggest that response time may also have value as a clinical indicator of malingering, for which there is experimental evidence (Vendemia, Buzan, & Simon-Dack, 2005; Willison & Tombaugh, 2006).

Additionally, functional imaging studies support the idea of differential recruitment of brain areas when feigning compared to truth-telling (Abe et al., 2006; Browndyke et al., 2008; Ganis, Kosslyn, Stose, Thompson, & Yurgelun-Todd, 2006; Langleben et al., 2005; Mohamed et al., 2006; Nunez, Casey, Egner, Hare, & Hirsch, 2005; Spence et al., 2001). For example, control recognition performance in Browndyke et al.'s (2008) sample was only associated with

dorsomedial parietal and inferior occipitotemporal areas. Simulators had increased medial temporal lobe and inferior parietal activity during both misses and false alarms. Simulators also had secondary activation: dorsomedial and dorsolateral prefrontal areas were specifically associated with misses, while the ventrolateral prefrontal area was more active during false alarms. Successful malingering would benefit from persistent monitoring of past responding and probabilistic outcomes, which is associated with medial temporal and inferior parietal activity (Lee et al., 2002; Spence et al., 2001; Spence et al., 2004). This differential recruitment, especially in prefrontal areas, suggests that deceptive responses require a greater deal of resources for cognitive control, and are therefore more effortful, relative to true responses.

Symptom and performance validity testing has been established as a critical part of the neuropsychological examination. Current malingering detection methods, such as the Word Memory Test (Green, 2005), have demonstrated high sensitivity and specificity but are not infallible. Despite evidence of their utility, secondary influences may impact patient performance on these tests, and in different ways depending on current emotional state, as well as comorbid psychological and medical disorders. There is a need for the identification and development of newer methodologies for malingering detection that are less, or even completely not, reliant on verbal responses such as MRI, EEG, eyetracking, and others. Because memory has been identified as the chief complaint reported by both credible and noncredible patients, as well as the area most tapped in existing PVTs, a new memory test would have the greatest impact on malingering detection practice.

Eye Movements as a Visual Index of Memory

Much of the foundational oculomotor research focused on properties of visual stimuli and how they affect viewing patterns. Visual stimuli are rarely perceived alone without the context of

other objects in the visual field. Therefore, many of the early studies on eye movements investigated the viewing of multi-object scenes or objects with multiple features. The visual system categorizes certain features as important and others as redundant, with important features receiving greater viewing time or gaze duration (Buswell, 1935; Mackworth & Morandi, 1967). The eyes are led to fixate on these important features, while the so-called redundant features are perceived through peripheral vision. Edges, and especially those which are rougher and less predictable evidenced greater viewing than contours (Mackworth & Morandi, 1967). The luminance gradient, responsible for edges, is one of the ways in which the brain determines what stimuli to prioritize and perceive more closely than others. Buswell (1935) also noted that providing viewers with specialized instructions or a question to answer about a scene could change viewing patterns with different features receiving importance based on the provided guidelines. Antes (1974), however, showed that while informative details receive greater attention overall, this pattern occurs in early viewing and less relevant details of a picture were fixated on for longer in later viewing. Additionally, stimuli that trigger a spontaneous recollection of the encoding experience received more fixations than those that were novel or familiar to participants without triggering recollection (Kafkas & Montaldi, 2012).

Although perceptual features of visual stimuli are naturally processed first, semantic properties are also taken into account (Henderson, Weeks, & Hollingworth, 1999). Loftus and Mackworth (1978) also showed that there is an early viewing effect, in addition to more frequent and longer fixation, for stimuli that are out of place in a scene, providing further evidence of the importance of semantic aspects. Stark and Ellis (1981) showed that subsequent viewing of familiar stimuli produced nonrandom viewing behavior. Initial viewing involves scanning for features, as mentioned previously, but viewing the same stimulus again constrains viewing

behavior such that “scanpaths” begin to emerge and the salient features of a stimulus receive the most attention.

Neal Cohen and his colleagues were instrumental in building a strong evidence base for the use of eye movements in memory research. Althoff and Cohen (1999) demonstrated the emergence of different viewing patterns for famous (i.e., familiar) and nonfamous (i.e., novel) faces, which is in accordance with past exposure. While participants fixated on main facial features more than other face areas for both types of faces, viewing patterns were more constrained for famous faces in that the eyes received significantly greater viewing than the mouth. Essentially, upon exposure to a previously viewed stimulus, participants examine less of the stimulus and make fewer fixations. This was also proposed to be an obligatory, rather than a cognitively controlled, response to having previously processed the presented images (Althoff, 1998; Althoff et al., 1998; Althoff & Cohen, 1999).

This eye movement-based viewing effect has been used indirectly to measure memory deficits, especially in amnesia. The indirect approach here refers to measuring a behavior (i.e., viewing) indicative of memory for a stimulus that occurs independently of an explicit statement of memory. Amnesic patients and controls both engaged in this repetition effect to the same degree; however, only controls exhibited a relational manipulation effect (Ryan, Althoff, Whitlow, & Cohen, 2000). The relational manipulation effect is an increase in viewing directed to regions of a visual scene that have been manipulated in some way from its initial presentation. This finding is significant because it partialed out relational learning deficits from general deficits in explicit or declarative memory. Although some implicit or domain-specific procedural memory can be compromised in amnesia (Cohen & Eichenbaum, 1993), performance on explicit memory tests is expected to be significantly worse. An indirect memory test, such as the increase

in viewing constraint resulting upon subsequent viewing of stimuli, is one measure that can be used to determine if learning has taken place. Eye movements are also able to detect if an individual recognizes studied faces early in viewing (1000-2000ms after display onset) when studied faces were presented with two morphed versions of the studied face. This disproportionate viewing effect was evident before and regardless of explicit responses (Hannula, Baym, Warren, & Cohen, 2012).

This is important because explicit statements about memory necessarily reference a past learning experience. Ryan et al. (2000) argue that poor explicit memory test performance, therefore, could be indicative of deficits in declarative memory. The authors concluded that amnesics were unable to engage in creating relational memory binding for objects in the original scenes, such that they did not attend to the scene manipulations as the controls did. Hannula et al. (2010b) also found that eye movement behavior patterns indicate relational memory impairment in schizophrenic patients, but no impairment on item memory. Additionally, more severe symptoms of schizophrenia were associated with worse task performance. Williams et al. (2010) investigated relational memory time course differences in the early viewing effect in schizophrenic patients, finding that viewing directed at studied faces is significantly reduced and occurs later in the trial following display onset (1250-1500ms vs. 500-750ms) compared to controls. From this, we know that eye movements have potential for providing detailed information about several kinds of memory deficits and may be a useful implicit measure.

Subsequent research has identified an early viewing effect of previously studied stimuli. Disproportionately greater viewing was directed at critical objects 500-750ms after test display onset in a relational memory task (Hannula, Ryan, Tranel, & Cohen, 2007). In this study, participants must use relational memory to learn associations between scenes and faces. The

early viewing effect was observed regardless of whether an explicit response was required. However, not providing a scene preview led to the emergence of the effect 1000-2000ms following onset of the full stimulus display. It should be noted, however, that eye movement patterns were sensitive to studied stimuli associates approximately one second before an explicit response was given. Given these results, it is believed that the scene preview enables faster retrieval of the associated face due to pattern completion, resulting in the observed eye movement effect occurring earlier in the trial. Amnesics tested using the same paradigm performed significantly worse than controls on the behavioral task, and their data did not indicate a relational memory effect even when recognition responses were correct. On incorrect trials, though, they directed disproportionate viewing to the face associated with their incorrect response rather than the correct face (response intention effect). When associated stimuli were presented together for amnesics, disproportional viewing did not emerge at any point during the trial.

Task demands also exert a significant influence on eye movements. When participants were instructed to not look at familiar faces, more time was spent viewing novel faces throughout the course of the trial. Similarly, when told to freely view three-face displays in preparation for a memory test, greater viewing was also directed to novel faces. Presumably, this is because familiar faces already have representations in memory, prompting participants to learn the novel faces in order to perform better on the test. When instructed to explicitly identify familiar faces, greater viewing was directed to familiar rather than novel faces (Ryan, Hannula, & Cohen, 2007). However, before any task demands can be met, a stimulus must first be identified as being familiar. In the condition requiring participants to not look at familiar faces, when presented with a famous face that was not presented in the experimental context,

disproportionate viewing occurred as early as 500-1000ms following display onset, following which viewing was diverted to novel faces. It should also be noted that people are almost always confident in their memory judgments, such that disproportionate viewing of familiar items in a relational task occurred sooner and was more likely when participants were most confident in their responses on correct trials only (Chua, Hannula, & Ranganath, 2012).

This viewing effect for novel views of familiar faces is explained by the absence of an identical memory representation for that specific image of a famous face within the experimental context (Ryan et al., 2007). When a studied famous face was present in the display, preferential viewing away from it emerged as early as 1000-1500ms, and as early as 2000-2500ms when presented with a studied non-famous face. It is argued that the absence of an identical match in recent memory requires more effort to correctly identify a novel famous face as famous, which led to the emergence of a novel face viewing preference 3500-4000ms following display onset. When participants were instructed to identify familiar faces, disproportionate viewing emerged 500-1000ms into the trial for both famous and non-famous studied faces, but at 1000-1500ms after display onset for novel famous faces. Ryan et al. (2007) state that this eye movement behavior is obligatory and is a function of past exposure, such that memory is the primary factor driving scanning and task demands being secondary.

Schwedes and Wentura (2012) replicated this early viewing effect using a visual concealed knowledge task and evaluating the first three fixations. By reducing the presentation time for a visual stimulus, it is also possible to detect deception, as well as prevent potential coaching or countermeasures employed to subvert the test (Bowman, Filetti, Alsufyani, Janssen, & Su, 2014). This, along with the early viewing effects, suggests that deception can be detected at both early and late stages in a given trial. If information is presented rapidly, there is little to

no opportunity to consider or formulate an alternate response. But if information is presented for several seconds, early viewing indicative of previous exposure followed by avoidance of target stimuli may be a potential indirect indicator of malingering.

The purpose of this study was to evaluate a novel approach to measuring recognition deficits by comparing inconsistencies between explicit and implicit behavior on a memory task. The importance of eye movement effects cannot be overstated, as this methodology has shown that, even in amnesia, the eyes can indicate which stimuli have been previously seen (Althoff, Maciukenas, & Cohen, 1993; Althoff, 1998; Ryan et al., 2000). This is a type of old/new effect (Gardner, Mo, & Borrego, 1974), meaning that there is a differential reaction depending on the novelty of a presented stimulus. This, combined with the knowledge that familiar stimuli receive disproportionately more viewing shortly after display onset, allows for the development of a rough framework for expected outcomes.

Eye movement methodology has shown to be a reliable way to study memory processes, but also has the potential to be a powerful tool in applied settings (Hannula et al., 2010a; Hannula, Baym, Warren, & Cohen, 2012). Both a true amnesic individual and one who is malingering a memory deficit are expected to provide explicit denials of memory for a studied stimulus, but their implicit behaviors should allow for them to be distinguished from each other. If using a relational memory task, the amnesic individual would not exhibit disproportionate viewing for manipulated scenes. Although the majority of the foundational research presented here involves relational memory, it is recognition memory that is most often evaluated in neuropsychological assessments. Therefore, investigating the discrepancy between explicit and implicit performance on a recognition task would provide additional support for the use of eyetracking as a means of indexing memory effects.

Because the majority of malingering detection occurs in the context of recognition memory tests, a measure or technique that evaluates memory without the need for explicit responding would be an attractive option for clinicians looking to better distinguish patients with legitimate memory complaints from those that are malingering. Despite the majority of this research being on relational memory, some effects are expected to carry over. For example, although amnesics would likely view novel objects for longer than familiar ones overall (repetition effect), they would not exhibit disproportionately greater viewing of familiar objects. Malingerers would be expected to have an early disproportionate viewing effect, which could then be compared to their explicit responses, foiling their attempts to feign impairment. Using recognition rather than relational memory allows the experimental task to be less time-consuming and somewhat easier, as the relational paradigms often use complex scenes or scenes paired with faces. Overall, this project is an effort to employ an experimental methodology in an applied context in order to determine its utility in a clinical setting.

Hypotheses

Hypothesis 1: Controls will perform significantly better than simulators on the behavioral accuracy measure during both the test and posttest phases.

Hypothesis 2: Simulators will perform significantly above chance on the behavioral accuracy measure during the test phase.

For the test phase, when simulators will be actively faking memory impairment, control participants are expected to achieve higher accuracy scores. Additionally, because simulators will receive only minimal coaching and will be warned against obviously poor performance, it is also expected that their average accuracy will be significantly above a chance level of 50%. In the posttest phase, however, simulators will no longer be faking impairment and could

potentially exhibit improved performance on the task. However, some interference from the experimental instructions is expected to prevent them from achieving performance equal to that of controls. This interference may occur during encoding (i.e., poor learning during the study phase) and/or retrieval (i.e., providing incorrect responses during the test phase). Control participants were likely to perform equally well in both phases due to the absence of an incentive to perform poorly.

Hypothesis 3: Controls will demonstrate significantly greater discriminability (d') than simulators during the test and posttest phases.

In addition to accuracy, it is useful to know if one group is more likely to engage in errors than the other. The discriminability index (d') is a statistic used in signal detection research to determine sensitivity. Simulators would necessarily need to make errors on either target-present trials, target-absent trials, or both to further their cause of exhibiting memory impairment. Therefore, the signal and noise distributions for simulators should have greater overlap than those for controls.

Hypothesis 4: Simulators and controls will spend significantly more time viewing studied stimuli over the course of the trial in target-present trials than non-studied critical items in target-absent trials in test and posttest phases.

Hypothesis 5: Disproportionately greater viewing of studied stimuli on target-present trials for simulators and controls will emerge 0-1000 ms and 500-750 ms following display onset using the 1000 ms and 250 ms levels of analyses, respectively.

The 500-750 ms time bin has been previously identified as the earliest that a within-groups difference emerges between viewing time direct to critical objects in target-present and target-absent trials (Hannula et al., 2007; Ryan et al., 2007). This difference emerged early and

was seen throughout the course of the trial. Therefore, it is expected that the same pattern of results should emerge in the 0-1000 ms and 500-750 ms time bins, using the 1000 ms and 250 ms levels of analysis, respectively. Investigating both levels of analysis will help determine if a more fine-grained approach provides more useful data, as opposed to simply more specific data.

Methods

Participants

A sample of 53 undergraduate students from the University of Wisconsin-Milwaukee was recruited through an online department subject pool (SONA). Participants without normal or corrected-to-normal vision (i.e., contact lenses, surgery) were excluded from the study. Although glasses do fall under the "corrected-to-normal" vision category, wearing them decreases the eyetracker's ability to reliably track eye position. Five participants were excluded from the study due to having an inadequate number of valid trials further described in the analysis of eye movement data. All participants were compensated with course credit.

The sample consisted of 38 females (79.2%) and 10 males (20.8%). The average age of participants was 22.125 years (SD = 5.63), while the median age was 21 years. The sample's race/ethnicity distribution was as follows: 66.7% Caucasian (n = 32), 12.5% African American (n = 6), 10.4% Asian (n = 5), 6.25% Latino/a (n = 3), 2.08% Native American or Alaska Native (n = 1), and 2.08% declined to answer (n = 1). Forty-two participants were right-handed, five were left-handed, and one was ambidextrous.

Materials

Equipment. Visual stimuli were presented using Neurobehavioral Systems Presentation experiment software running on a Windows-based computer and color monitor. This same computer also recorded behavioral responses and response time. Eye position and gaze duration

was obtained using an ASL D6 desktop eyetracking system sampling at 60 Hz. The eyetracker was controlled using proprietary software developed by ASL on a second Windows-based computer. Behavioral responses were obtained through a keyboard-like game controller intended for use with a single hand with number labels on three buttons.

Stimuli. For this experiment, 144 images were obtained from a standardized set of stimuli designed for research purposes (Snodgrass & Vanderwart, 1980). The stimuli are images of commonly-used nouns and include things such as animals, food, man-made tools, buildings, and widely-recognized symbols (e.g., heart). The original unedited stimuli were black line drawings on a white background and were meant to be projected onto a screen. Bitmap files of these images were obtained for this study. The size of each of the 144 images was increased from 150 by 150 pixels to 280 by 280 pixels to increase their visibility.

Each image was categorized as belonging to one of three distinct categories: a) organic (i.e., animals, plants, food, body parts), b) large inanimate objects (e.g., buildings, vehicles), or c) small inanimate objects most often manipulated using the hands (e.g., tools, clothing, jewelry, etc.). These categories were selected so that each would have roughly equal membership. Images from each category were randomly assigned to one of six 24-item lists with equal contribution from each of the three stimulus categories. One list served as the study list (e.g., A), in that participants were asked to view and remember those images in the study phase. These stimuli served as critical objects for target-present trials. Two lists (e.g., B and C) served as distractors for the study list in target-present trials. Target-present trials are those in which a previously-studied stimulus exists, for the purpose of testing an individual's memory. The remaining three lists (e.g., D, E, and F) served as stimuli in target-absent trials. Stimuli from one of these

remaining lists were used as critical objects in target-absent trials. Lists of images and their categories are provided in Appendix A.

Using a separate set of stimuli for target-absent trials ensures that there is no memory overlap between target-present and target-absent trials. For example: if a distractor from a target-present trial is then reused in a subsequent target-absent trial, this may lead to a "present" response rather than the correct "absent" response. Even though participants are instructed to only respond "present" on trials that have studied stimuli, the familiarity of reused stimuli may lead to unintended errors. Additionally, reusing a stimulus in a target-absent trial would be detrimental to the study design. A novel stimulus presented alongside two other novel stimuli is expected to receive approximately 33% of the viewing time for the trial. A previously viewed stimulus, even if originally presented alongside a more salient stimulus (i.e., a stimulus from the study phase), would likely have a viewing time advantage over a completely novel stimulus when presented in a target-absent trial. This advantage defeats the purpose of having target-absent trials, which are meant to serve as baselines with which target-present trial data can be compared. Such a confound could confuse participants that are attempting to follow directions, but would also likely obscure statistical differences both between and within participants.

Stimuli within each list also had an equal likelihood of appearing on one of the three locations in the test displays. Lists were counterbalanced across participants such that stimuli from each list had an equal likelihood of being studied. For example, participant 1 would study stimuli from list A, have list B and C stimuli appear as distractors in target-present trials, and have list D, E, and F stimuli appear in target-absent trials. Participant 2 would study list B stimuli, have list C and A stimuli as distractors in target-present trials, and have list E, D, and F stimuli in target-absent trials. For participant 2, list E serves as the critical object list for target-

absent trials, whereas list D would serve that function for participant 1. Participant 4, however, would study list D, have lists E and F as distractors, and have lists A, B, and C as target-absent stimuli. Participants 5 and 6 would cycle through the lists in the same way. A complete diagram of stimulus counterbalancing is available in Appendix B. Trial order was randomized for each participant. The locations of critical objects in displays were also randomized to reduce the probability of guessing the correct answer.

As luminance changes contribute to ocular reflexes (Porter et al., 2010; Porter & Troscianko, 2003), the goal was to minimize the luminance differential while still maintaining good visibility and allowing encoding to take place. Bitmap files are comprised of pixels, and as such, have luminance gradients at object edges. Changing luminance values in a bitmap file would produce noticeable and odd-looking luminance gradients that would significantly distinguish it from its original version. Instead, a freeware vector editor, Inkscape 0.48.4, was used to convert each image into a vectorized line drawing, change its luminance and that of the background, and then convert it back into a bitmap file. The luminance of the background was set at 75 cd/m^2, while that of the object was set at 40 cd/m^2. This procedure allowed the foreground and background luminances to be altered independently and minimized the extent of the luminance gradient. This procedure was also partially adapted from Porter et al. (2010), who used values of 75.2 cd/m^2 and 63.6 cd/m^2 for background and foreground, respectively. As the Porter study featured a feature search task in which the stimuli were visually identical but were oriented in different directions, the low contrast in that study was necessary to encourage search behavior and prevent pop-out effects, as well as to minimize pupillary reflexes (Porter & Troscianko, 2003). In the present study, such low contrast could potentially interfere with proper

encoding of stimuli in the study phase, as well as recognition and decision making in the test and posttest phases. Therefore, the contrast was increased slightly to facilitate improved viewing.

Procedure

Descriptions of the experiment, the equipment involved in the procedure, and any risks associated with participation were provided to the participants by a research assistant before obtaining informed consent. After obtaining informed consent, a research assistant randomly assigned participants to the control or simulator condition and read the appropriate instructions. Participants assigned to the control condition were instructed to perform their best on a memory test. Participants assigned to the simulator condition were provided with a vignette and asked to imagine that the situation in the vignette had happened to them. The vignette describes a car crash scenario after which the participant had a concussion, was taken to the hospital, and did not have any impairment. However, the vignette states that the participant has decided to fake memory impairment in order to receive a larger settlement from legal proceedings against the driver who caused the car accident. The participant was instructed that they would take a test to determine if they have impairment, and that they will receive no settlement if they were caught faking. The full text of the vignette is provided in Appendix C.

Simulators were provided with minimal coaching. They were instructed to respond to every display, and to not always respond incorrectly; engaging in either of these two behaviors would make the evaluator suspicious of their performance. Following this, the simulators were asked to explain the objective in their own words to ensure comprehension. The participant was then allowed to ask questions about the objective before the evaluator entered the room, but not after. Research assistants were instructed to only provide further explanation of already provided instructions, but not to provide any additional instructions, constraints, or demand characteristics.

Essentially, simulators were free to manipulate their performance in any way they saw fit to complete the objective. While they were free to not respond to some items or provide only incorrect responses, it was ensured that they understood that engaging in these strategies would actually make it more likely to be detected. Simulator preparation instructions were significantly longer than that of controls, and could reveal a participant's group assignment to the experimenter. To avoid this, research assistants were instructed to ask control participants general questions about their coursework for several minutes.

The evaluator then entered the room and measured the distance between the participant's eye and a computer screen directly in front of them. The ideal distance for the eye behavior recording with this equipment is 24 inches. Participants were instructed to sit in a comfortable position in which they believed they could sit still without making large or unnecessary movements. Following this, the eyetracking software was used to track the participant's head. This feature eliminated the need for a chinrest and provided for a less intrusive and more ecologically valid experience. Calibration was accomplished by presenting a 3 by 3 spatial array with letters on the screen. The participant was asked to look directly at the middle of each letter, which would allow the software to obtain reference points to be used in mapping that individual's eye movements in two-dimensional space.

The evaluator described the study phase and instructed the participant to study the images in preparation for a subsequent memory test. A PowerPoint slide was displayed that showed what the display will look like, with the evaluator explaining the process one more time. Participants were then instructed to remain as still as possible and to not look away from the screen. The study phase involves showing a full sequence of 24 pictures, followed by the same 24 pictures in a new randomized order, for a total of 48 trials. Stimuli were presented

individually for 3000 ms each, for a total of 6000 ms of exposure to each studied item. To ensure that participants were viewing the images, a fixation cross would appear on the screen for a minimum of 1000 ms. Once the evaluator saw that the participant's gaze is centered on the fixation cross, the evaluator manually triggered the trial to begin.

Once the study phase was complete, participants were shown a second PowerPoint slide that demonstrated the test displays with the evaluator providing instructions on how to respond. In the test phase, participants were shown 48 three-item displays. An illustration of sample displays and task progression is provided in Figure 1. Each display has one item in the upper left region, one item in the upper right region, and one item in the lower center region. Half of the displays contain only novel stimuli that have not been shown in the study phase (target-absent displays). The remaining displays contain one studied image (critical item) and two novel images (distractors) and will be referred to as target-present displays. The fixation cross procedure between trials, as described for the study phase, was also used for the test and post-test phases. The three-item displays were presented for 6000 ms each, followed by an intertrial interval of 10000 ms where only a blank screen was displayed. The length of the intertrial interval was selected to allow for a rest period from studying the stimuli, as well as to allow ocular reflexes, such as pupillary diameter, to return to baseline before presenting the next set of stimuli (Porter et al., 2010). Participants responded with their dominant hand using a keyboard-like game controller. If a three-item display contained a familiar picture (i.e., an image presented during the study phase) participants were to press the '1' key, otherwise, they were to press the '2' key. Trials were experimenter-initiated based on verification of central fixation.

Once the test phase was complete, a display detailing the posttest phase was shown and the evaluator provided additional instructions. Before anything about the task content was

revealed, the evaluator instructed the participant to do their best on the task, regardless of whether they were previously instructed to simulate memory impairment. Because the evaluator was blind to group assignment, this instruction was given to all participants. This manipulation was used to determine if simulators actually learned the stimuli during the study phase and volitionally adjusted their performance in the test phase. In other words, if they truly learned the stimuli, they should be able to perform well in the posttest phase.

Here, participants viewed the three-item displays identical to those they saw in the test phase, but in a re-randomized order. The displays were also shown for 6000 ms, with a 10000 ms intertrial interval and a fixation cross procedure of at least 1000 ms. In contrast to the test phase, the posttest phase involved slightly more involvement from the participants. For each display, participants would first use the keyboard-like game controller to provide the locations of studied items. If a studied item appeared in the upper left region, they were to press the '1' key. Studied items located in the upper right and lower center regions were assigned the '2' and '3' keys, respectively. If a studied item was not present, participants were instructed to press any of the three keys as soon as they have identified the absence of a studied item in the display. Once the display was removed from the screen and the intertrial interval had begun, the participant would provide a yes/no verbal response. This response refers to whether they believe a studied item was in the display they just saw. This allowed for the backsorting of button responses in this phase to determine reaction times for target–present and –absent trials. Verbal responses were recorded by the research assistant on a separate form.

After the posttest phase is finished, participants completed an exit questionnaire about objectives, effort, motivation, confidence, and strategies. This questionnaire is provided in Appendix E. Once the participant completed the questionnaire, they were debriefed by the

evaluator and/or research assistant and could ask questions about the study. Barring any major technical issues, completing all parts of the study protocol took approximately 60 minutes.

Analyses

Behavioral Data

Behavioral responses from the Presentation software were first compared against a list of known correct responses. A standard signal detection paradigm was used: a target-present trial that received a "present" response was marked as a hit, while an "absent" response was marked as a miss. For target-absent trials, a "present" response was counted as a false alarm, while an "absent" response was marked as a correct rejection. Corrected recognition score was the sole behavioral accuracy measure in the study. This is calculated by summing a participant's hit and correct rejection rates, and then dividing by two. This adjustment better accounts for the occurrence of false alarms in a response set.

In the event that a participant provided two different responses on a single trial, behavioral data from that trial was excluded from analysis. It was speculated that using those trials would not be a reliable measure of their true response; an input error may be caused by carelessness, and a response change may be the result of a participant trying to change a correct answer to an incorrect one to appear more impaired. It may also potentially reflect decreased confidence in their own responses or an inability to properly follow the task directions. Because the simulator's goal is to not arouse suspicion, a high incidence of changing answers would likely work against them in a genuine incentive-seeking situation.

Behavioral accuracy was evaluated using a two-way mixed model ANOVA using corrected recognition as the dependent variable. Both between-group (i.e., controls vs. simulators) and within-group (i.e., test vs. posttest) performance were compared. Simulator

performance was also evaluated to determine the success of the experimental manipulation. Above-chance performance for simulators was evaluated using binomial one-sample *t*-tests for the test and posttest phases separately. Performance at chance level on both phases would yield an accuracy score of 50%. This calculation will determine if simulators, as a group, effectively manipulated their performance in accordance with experimental instructions to not make their deception obvious to the experimenter.

D' is calculated as the difference in z-scores between a participant's hit rate and false alarm rate. In signal detection theory, it is the distance between the peaks of the signal and noise distributions. Higher values of d' indicate a greater ability to correctly identify studied stimuli. A two-way mixed model ANOVA was performed to compare d' scores between simulators and controls for the test and posttest phases.

Eye Movement Data

Recent relational memory eyetracking research using a similar paradigm has shown a pattern of greater viewing to be directed at familiar stimuli than novel stimuli emerging shortly after display onset (500-750ms) and continuing throughout the duration of the trial (Hannula et al., 2007; Ryan et al., 2007). Although this study investigated recognition item memory, this early viewing effect was also expected to occur. Because this study also dealt with deception, the above hypothesis will need to be unpacked to some extent. Participants should spend more time viewing familiar stimuli, regardless of whether they provide a correct (hit) or incorrect (miss) behavioral response on target-present trials. On target-absent trials, all stimuli are novel and none should receive significantly more viewing than the others in a three-item display.

For test and posttest phase trials, three regions of interest were identified on the stimulus displays: left, right, and bottom. Raw output was processed by an algorithm that groups samples

collected at 60 Hz into fixations. A fixation is defined as when eye position is maintained within 0.5 degrees of viewing across at least six consecutive samples. Using this algorithm, a fixation is considered complete if two consecutive samples differed by at least one degree of viewing angle. A second algorithm was used to separate the data into 1000ms time bins (i.e., 0-1000ms, 1000-2000ms, etc.). A separate procedure was used to apply a similar algorithm to separate the data into 250ms time bins (i.e., 0-250ms, 250-500ms, etc.) for a deeper level of analysis. Evaluating data at the 250 ms level allows for more specific temporal detection of viewing effects, at the cost of increasing the number of data to be analyzed.

Viewing times for critical objects were isolated and calculated both across trials and within each time bin. Following this, proportions of viewing time were calculated to facilitate comparison across trials, as not all trials had the same amount of viewing time. Although trials are 6000 ms in length, given the size of the screen and the dimensions of each of the three regions of interest, it is unlikely that all of the 6000 ms were spent viewing the three regions by each participant. Trials with less than 65% (3900 ms) of the trial time spent viewing the three regions of interest was excluded from eye data analysis.

For this study, eye movement data analysis required a minimum of four trials per signal detection category for an adequate sample of viewing time. As there is no learned stimulus to detect in target-absent trials, false alarm and correct rejection trial data was collapsed into a single “target-absent” category. Therefore, to be retained for analysis, simulators needed a minimum of four hits, four misses, and four target-absent trials based on the viewing time criterion established above. Because controls were not specifically instructed to provide incorrect answers on some trials, and were actually encouraged to do their best, it was not expected that a minimum of four miss trials would be obtained from each control participant. Therefore, controls

only needed four hits and four target-absent trials with adequate viewing time to be retained for analysis.

Using this approach, five participants (3 simulators, 2 controls) were excluded from the sample, resulting in a final sample size of 48. From this final sample, 543 of the 2304 total trials (23.57%) were rejected using the viewing time criterion. Determining if one of the groups is significantly more likely to spend less time viewing the stimuli overall would be of interest. A significant difference in rejected trials between groups could indicate a simulation strategy or decreased effort expended over the course of the task. However, it could also indicate difficulties in obtaining reliable eyetracking data, as there are multiple reasons why a trial could be rejected. Therefore, a direct statistical comparison of rejected trials would not be appropriate.

All eye movement data of a particular type (e.g., hit, target-absent) and time bin (e.g., 0-250 ms, 1000-2000 ms) was first averaged within-participants, then between-participants by group. A mixed model ANOVA was used to identify significant differences in proportion of viewing time between simulators and controls, and within groups by trial type (i.e., target-present and target-absent) at the trial level. A one-way repeated measures ANOVA evaluated within-group differences using three trial types (i.e., hit, miss, and target-absent) at the trial level for simulators only. The goal of the second analysis was to identify differences between hit and miss trials, as the miss trials are critical for the identification of simulators. Partial eta-squared (η^2) was used as a measure of effect size. Commonly accepted interpretations for this statistic are .01 (small), .06 (medium), and .14 (large). For data obtained from 1000 ms and 250 ms time bins, focused mixed model repeated measures ANOVAs, as well as two-tailed independent and dependent *t*-tests were used to compare the proportion of viewing time between groups and within groups by trial type (i.e., hit, miss, target-absent), as appropriate.

Secondary Analyses

To evaluate the utility of each of these measures as a tool for detecting feigned memory impairment, receiver operating characteristic (ROC) analysis was employed. ROC curves plot the true positive rate and false positive rates for various cutoff scores on a particular variable. Traditionally, the true positive rate is termed *sensitivity*, while the false positive rate is displayed as *one minus specificity*. For the purposes of this study, the ROC plot provides a visual representation of how many participants are correctly classified into their experimentally assigned groups on a given measure using a variety of cutoff scores. In fact, as sensitivity increases, specificity tends to decrease, leading to the difficult decision of deciding on an appropriate cutoff score.

Test developers seek to maximize both of these properties, but it is highly unlikely that a psychometrically perfect measure exists or will be developed. Therefore, instruments designed for clinical use should have sufficiently high sensitivity and specificity to justify their use in settings outside of research. Practically, a measure or technique should have at least 0.80 on both of these indices. However, in the case of malingering, clinicians are more interested in correctly identifying those who put forth good effort, suggesting that a specificity of at least .90 is more desirable.

In order for the cutoff scores to be useful, the test or variable itself must be useful. The ROC analysis provides a measure of this by calculating the area under the ROC curve (AUC). An AUC of 0.5 represents a test whose true positive and false positive rates are equal, suggesting that it is functionally worthless. Generally, the greater the AUC, the better the test is at being a diagnostic instrument due to its higher sensitivity and specificity values. Tests yielding AUCs under 0.5 should be avoided at all costs.

ROC curves were plotted for a variety of behavioral and eye movement measures to identify useful clinical indicators of malingering. Because performance on the behavioral accuracy measure produces only one score per participant in the test phase, a single ROC analysis was conducted using behavioral data. For eye movement data, two sets of ROCs were plotted. The first set includes 31 separate ROC curves: one for each of six 1000ms time bins, 24 250ms time bins, and one for collapsed trial-level data from the test phase. The test variable for these 31 ROCs is the proportion of viewing time directed to the studied item in target-present test phase trials. The second set of eye movement ROCs was identical to the first, save for the test variable. Here, a difference score was calculated using the proportion of viewing time directed to studied items between correct (hit) and incorrect (miss) target-present trials. As controls are unlikely to have many incorrect trials individually, an average of all control miss trials for this variable was used for this calculation instead.

Each represents an index that could potentially be used for clinical practice, and combinations of multiple such variables could create an even more powerful index. For this study, only one test variable was selected for each type of data because they represent the most useful individual indices of memory in research settings. The identification of a multimethod index is time-consuming and somewhat beyond the scope of this study. Due to the importance of such information, however, an attempt to create such an index will be made using available variables that demonstrate good diagnostic utility.

Results

Behavioral Data

The 2 (controls vs. simulators) x 2 (test vs. posttest) two-way mixed model ANOVA compared corrected recognition scores. There was a main effect of group, $F(1,46) = 123.778$, $p <$

.001, partial η^2 = .729, as well as a main effect of time, $F(1,46) = 32.317$, $p < .001$, partial η^2 = .413. There was also a significant interaction effect, $F(1,46) = 48.432$, $p < .001$, partial η^2 = .513.

Post-hoc two-tailed independent *t*-test comparisons revealed that controls (95.69%) performed significantly better than simulators (61.39%) in the test phase, $t(46) = 12.482$, $p < .001$. Controls (93.58%) also significantly outperformed simulators (82.38%) in the posttest phase, $t(46) = 4.454$, $p < .001$. However, post-hoc two-tailed dependent *t*-test comparisons showed that simulators were able to significantly improve their performance in the posttest phase when compared to the test phase, $t(23) = 6.524$, $p < .001$. Interestingly, controls evidenced a small, but significant decrease in performance from the test phase to the posttest phase, $t(23) = -2.582$, $p < .05$. Because controls performed significantly better than controls on the behavioral task in both phases, these results fully support Hypothesis #1.

Additionally, the binomial one-sample two-tailed t-tests found that simulators performed significantly above chance level (>50%) on the test phase, $t(23) = 4.433$, $p <.001$, and on the posttest phase, $t(23) = 14.809$, $p < .001$. Therefore, simulators successfully manipulated their performance by making a notable amount of errors, but not so many as to appear to be at or below chance, which supports Hypothesis #2.

A separate 2 (controls vs. simulators) x 2 (test vs. posttest) two-way mixed model ANOVA was used to compare discriminability index (d') scores. There was a main effect of group, $F(1,46) = 104.539$, $p < .001$, partial η^2 = .694, as well as a main effect of time, $F(1,46) = 20.654$, $p < .001$, partial η^2 = .310. There was also a significant interaction effect, $F(1,46) = 46.183$, $p < .001$, partial η^2 = .501.

Post-hoc two-tailed independent *t*-test comparisons revealed that controls (d' = 3.71) had better discriminability than simulators (d' = 0.66) in the test phase, $t(46) = 13.910$, $p < .001$.

Controls (d' = 3.41) also had greater discriminability than simulators (d' = 2.20) in the posttest phase, $t(46) = 4.362$, $p < .001$. Post-hoc two-tailed dependent *t*-test comparisons showed that simulators significantly improved their discriminability in the posttest phase, $t(23) = 6.548$, $p < .001$. Controls evidenced a small, but significant decrease in discriminability from the test to the posttest phase, $t(23) = -2.251$, $p < .05$. These results support Hypothesis #3.

Eye Movement Data

Trial level analysis.

Test phase. Proportions of viewing time for the test phase are presented in Table 1. The 2 (controls vs. simulators) X 2 (Hit trials vs. Absent trials) mixed model ANOVA found that there was a significant main effect of Group, $F(1,46) = 4.325$, $p = .043$, partial $\eta^2 = .086$. There was also a significant main effect of Trial type, $F(1,46) = 103.818$, $p < .001$, partial $\eta^2 = .693$. The interaction was only marginally significant, $F(1,46) = 3.811$, $p = .057$, partial $\eta^2 = .077$. A post-hoc two-tailed independent *t*-test found a significant difference between controls and simulators on Hit trials (.560 vs. .483), $t(46) = 2.143$, $p = .037$. Post-hoc two-tailed dependent *t*-test comparisons found significant differences between Hit and Target-Absent trials within controls (.560 vs. .339), $t(23) = 8.487$, $p < .001$, as well as simulators (.483 vs. .333), $t(23) = 5.894$, $p < .001$. These results provide test phase support for Hypothesis #4.

As the simulator group had a greater number of Miss trials, a follow-up analysis of all trial types was completed in order to detect differences in viewing behavior between Hit and Miss trials. A one-way repeated measures ANOVA found significant within-subjects mean differences, $F(2,46) = 17.065$, $p < .001$, partial $\eta^2 = .426$. Proportion of viewing time directed at the critical object was significantly different between Hit trials (.483) and Miss trials (.398), $t(23)$

= 3.188, p = .004. The mean difference between Miss trials and Absent trials (.333) was also significant, $t(23) = 2.586$, p = .017.

Given that each participant undergoes 48 trials in the test phase, each group collectively undergoes 1152 trials. Overall, controls had 318 rejected trials (27.6%), while simulators only had 225 (19.5%), for a total of 543. While it would seem more likely that simulators would have more rejected trials due to purposeful invalidation of trials through off-screen viewing or response issues, it was not expected that controls would actually have more rejected trials overall.

Posttest phase. The 2 (controls vs. simulators) X 2 (Hit trials vs. Absent trials) mixed model ANOVA found that there was no main effect of Group, $F(1,45) = 1.088$, p = .303, partial η^2 = .024. There was a significant main effect of Trial type, $F(1,45) = 234.237$, $p < .001$, partial η^2 = .839. The Group X Trial type interaction was not significant, $F(1,45) = .336$, p = .565, partial η^2 = .007. Post-hoc two-tailed dependent t-test comparisons found greater viewing directed at critical objects during Hit trials for controls (.597 vs. .339), $t(22) = 12.556$, $p < .001$, as well as simulators (.566 vs. .326), $t(23) = 9.579$, $p < .001$. These results support Hypothesis #4. Additionally, a two-tailed independent t-test found no significant difference between controls and simulators on Hit trials, $t(45) = 0.950$, p = .347. One control participant had insufficient valid Hit trials due to poor tracking, resulting in that data point being entered as missing data for these analyses.

While collapsed trial level data provides important information about overall differences between group trial types, it does not reveal *when* significant differences occur. Because attention is a continuous phenomenon, we looked to the 1000 ms and 250 ms levels of analysis to

determine whether the groups demonstrate implicit recognition at the same time and whether they exhibit differing patterns of viewing behavior throughout the trial.

Time course analysis at the 1000 ms level. The full time course is presented graphically in Figure 2. Data regarding the proportions of viewing time for the test phase are presented in Table 1. A 2 (controls vs. simulators) X 2 (Hit trials vs. Absent trials) X 6 (six 1000 ms timebins) mixed models ANOVA was performed. There was a significant main effect of Group, $F(1,46) = 4.962$, $p = .031$, partial $\eta^2 = .097$. More significant was the main effect of Trial type, $F(1,46) = 106.052$, $p < .001$, partial $\eta^2 = .697$. Mauchly's Test of Sphericity indicated that the independent variable of Time had violated the assumption of sphericity, $\chi^2(14) = 29.762$, $p = .008$. Utilizing Greenhouse-Geisser correction, there was a main effect of Time, $F(3.919,180.269) = 2.870$, $p = .025$, partial $\eta^2 = .059$. The Trial type X Group interaction was marginally significant, $F(1,46) = 3.876$, $p = .055$, partial $\eta^2 = .078$. The Time X Group interaction was also marginally significant, $F(5,230) = 1.964$, $p = .085$, partial $\eta^2 = .041$. Mauchly's Test of Sphericity indicated that the Trial type X Time interaction had violated the assumption of sphericity, $\chi^2(14) = 26.017$, $p = .026$. Utilizing a Greenhouse-Geisser correction, this interaction was also significant, $F(4.053,186.426) = 4.843$, $p < .001$, partial $\eta^2 = .095$. The three-way interaction was also significant, $F(5,230) = 4.514$, $p < .001$, partial $\eta^2 = .089$.

A follow-up repeated measures ANOVA was performed using Hit, Miss, and Absent data from 1000 ms time bins from the simulator group alone. These results are presented graphically in Figure 3. There was a significant main effect of Trial type, $F(2,46) = 17.819$, $p < .001$, partial $\eta^2 = .437$. Mauchly's Test of Sphericity indicated that the sphericity assumption was violated for Time, $\chi^2(14) = 25.339$, $p = .032$, and the Trial type X Time interaction, $\chi^2(54) = 96.166$, $p < .001$. Using a Greenhouse-Geisser correction, there was a marginally significant main effect of Time,

$F(3.560,81.876) = 2.451$, $p = .059$, partial $\eta^2 = .096$. The interaction was also significant following a Greenhouse-Geisser correction, $F(5.888,135.427) = 2.671$, $p = .018$, partial $\eta^2 = .104$.

A series of two-tailed dependent *t*-tests were performed comparing proportion of viewing time to critical objects in Hit and Absent trials for the test phase. Both controls and simulators directed significantly greater viewing towards critical objects in Hit trials throughout all six timebins, all $p < .01$. Bonferroni correction was used to adjust the rate of family-wise error for 12 comparisons, changing the significance criterion to $p < .004167$. Following this correction, 11 of the 12 comparisons remained significant, with the exception of controls in the 5000-6000 ms timebin ($p = .0051$). The significant difference in viewing time in the 0-1000 ms timebin for both groups supports Hypothesis #5.

A series of two-tailed independent *t*-tests were performed comparing controls and simulators on the proportion of viewing time to critical objects in Hit trials only. Controls directed greater viewing to critical items than simulators on Hit trials for the first half of the trial (0-3000 ms), all $p < .05$, but not during the timebins in the second half (3000-6000 ms), all $p > .05$. Bonferroni correction was used to adjust family-wise error for six comparisons, changing the significance criterion to $p < .0083$. Following this correction, only the differences for the 1000-2000 ms ($p = .0002$) and the 2000-3000 ms ($p = .002$) timebins remained significant.

A series of two-tailed dependent *t*-tests were performed comparing proportion of viewing time to critical objects in Hit and Miss trials made by simulators. Simulators did not statistically differ in their viewing of critical objects during the first half of the trial, all $p > .05$. However, they directed significantly less viewing time to critical objects during Miss trials in the second half of the trial, all $p < .05$. Bonferroni correction was used to adjust family-wise error for six

comparisons, changing the significance criterion to $p < .0083$. Following this correction, only the differences for the 4000-5000 ms and 5000-6000 ms timebins remained significant, both $p < .01$.

Time course analysis at the 250 ms level. The full time course of the test phase is presented graphically in Figure 4. Data regarding the proportions of viewing time for the test phase are presented in Tables 2 and 3. A 2 (controls vs. simulators) X 2 (Hit trials vs. Absent trials) X 8 (eight 250 ms timebins) mixed models ANOVA was performed. From the 1000 ms analyses, it was clear that significant differences occurred within the first two seconds of the trial. Therefore, only the first eight timebins were evaluated here. There was a significant main effect of Group, $F(1,37) = 7.278$, $p < .05$, partial $\eta^2 = .164$, as well as a significant main effect of Trial type, $F(1,37) = 94.119$, $p < .001$, partial $\eta^2 = .718$. Mauchly's Test of Sphericity indicated that the sphericity assumption was violated for Time, $\chi^2(27) = 149.730$, $p < .001$, and for the Trial type X Time interaction, $\chi^2(27) = 159.634$, $p < .001$. Using Greenhouse-Geisser corrections, there was a significant main effect of Time, $F(2.692,99.586) = 5.996$, $p < .002$, partial $\eta^2 = .139$, and a significant Trial type X Time interaction, $F(2.894,107.063) = 2.930$, $p < .05$, partial $\eta^2 = .073$. The Group X Trial type and Group X Time interactions were also both significant, $F(1,37) = 10.989$, $p = .002$, partial $\eta^2 = .229$ and $F(7,259) = 3.618$, $p < .001$, partial $\eta^2 = .089$, respectively. The three-way interaction was also significant, $F(7,259) = 3.686$, $p < .001$, partial $\eta^2 = .091$.

Due to the large number of data points, even within this limited scope, performing all possible pairwise comparisons would greatly increase family-wise error rates. Therefore, only comparisons relevant to the hypotheses and the supported literature were completed (i.e., identifying the *earliest* timebin with statistically significant differences). Controls directed greater viewing at critical objects during Hit trials 500-750 ms following display onset, $t(23) =$

4.289, $p < .001$. Simulator data for the 500-750 ms timebin showed that they did not direct significantly more viewing time at critical objects, $t(23) = 1.649$, $p = .11$. They did, however, evidence significantly greater viewing in the 750-1000 ms timebin, $t(23) = 3.217$, $p < .01$. When comparing the groups at the 500-750 ms timebin, it was found that controls had significantly greater viewing time than simulators, $t(46) = 2.617$, $p < .05$.

The full time course of the posttest phase is presented graphically in Figure 5. Data regarding the proportions of viewing time for the posttest phase are presented in Tables 4 and 5. A 2 (controls vs. simulators) X 2 (Hit trials vs. Absent trials) X 8 (eight 250 ms timebins) mixed models ANOVA was performed. There was no significant difference between groups, $F(1,32) = 2.195$, $p = .148$, partial $\eta^2 = .064$. There was a significant main effect of Trial type, $F(1,32) = 144.876$, $p < .001$, partial $\eta^2 = .819$. Mauchly's Test of Sphericity indicated that the sphericity assumption was violated for Time, $\chi^2(27) = 119.705$, $p < .001$, and for the Trial type X Time interaction, $\chi^2(27) = 113.031$, $p < .001$. Using Greenhouse-Geisser corrections, there was a significant main effect of Time, $F(3.213,102.829) = 8.800$, $p < .001$, partial $\eta^2 = .216$, and a significant Trial type X Time interaction, $F(3.530,112.952) = 7.945$, $p < .001$, partial $\eta^2 = .199$. The Group X Trial type interaction was statistically significant, $F(1,320 = 7.995$, $p < .01$, partial $\eta^2 = .200$, but the Group X Time interaction was not, $F(7,224) = .914$, $p = .496$, partial $\eta^2 = .028$. The three-way interaction was also not statistically significant, $F(7,224) = 1.362$, $p = .223$, partial $\eta^2 = .041$.

Similarly to the approach taken to test phase data, only relevant pairwise comparisons were performed. For the 500-750 ms timebin, controls and simulators both directed greater viewing at critical objects in Hit trials than in Absent trials, $t(23) = 3.414$, $p < .01$ and $t(23) = 3.811$, $p < .001$, respectively. To determine if there was an early between-group difference for

Hit trials, pairwise comparisons were conducted using four consecutive timebins that include when early-trial effects occurred in the test phase: 500-750, 750-1000, 1000-1250, and 1250-1500 ms. The greatest mean difference occurred at the 750-1000 ms timebin with controls having greater viewing time, but this difference was only marginally significant, $t(46) = 1.909$, $p = .0625$. The remaining three comparisons were not significant, $p > .05$, and all four comparisons were not significant following Bonferroni correction for four comparisons.

Secondary Analyses

Hit-Miss difference calculation. While simulators are expected to make errors on target-present trials that translate into Misses, controls performed at their best and were not expected to have many Miss trials. Only 20 Miss trials were collectively made by controls. Proportion of viewing time directed at critical objects in these trials was averaged together by level of analysis (i.e., trial, 1000 ms, 250 ms). To calculate the Hit-Miss difference, the appropriate Miss data point was subtracted from a control participant's Hit data point for that level of analysis. Miss trial and Hit-Miss difference scores are presented in Table 6 for trial and 1000 ms data and in Tables 7 and 8 for 250 ms data.

Behavioral ROC data. Corrected recognition scores for both groups were entered into a ROC analysis to determine the diagnostic utility of this measure. The ROC curve is presented in Figure 6. The analysis revealed that corrected recognition score is a near-perfect diagnostic variable, AUC = .998, $p < .001$. Using a cutoff score of 80.2% yields a sensitivity of .958 and a specificity of 1.0. A cutoff of 84.1% will instead produce a sensitivity of 1.0 and a specificity of .9583.

Trial level data. Separate ROC analyses were performed for the proportion of viewing time directed at critical objects during Hit trials and the calculated difference in proportion of

viewing time directed at critical objects between Hit and Miss trials. For Hit data, lower values were specified to be more indicative of simulator status, whereas greater values were to be more indicative when using the Hit-Miss difference. The associated ROC curves are presented in Figures 7 and 8, respectively. Tables 9 and 10 provide AUC data and selected cutoff scores for the two curves. The analysis revealed that viewing time during Hit trials is a marginally significant diagnostic variable, AUC = .660, $p = .058$. Having perfect specificity would mean lowering sensitivity to .125, which is unadvisable. Decreasing specificity to .958 would boost sensitivity to .333. The Hit-Miss difference score appears to be a more significant diagnostic variable, AUC = .767, $p < .01$. A sensitivity of .417 would yield a specificity of .917. Using a specificity of .958 provides a sensitivity of .208, suggesting that trial-level Hit data alone may be more useful at this specificity.

1000 ms level data. Six ROC analyses each were performed for Hit trial and Hit-Miss difference data at the 1000 ms level of analysis. Smaller values were specified as more indicative of simulator status for Hit data. The ROC curves for Hit data are presented in Figures 9 and 10, with associated ROC analysis statistics presented in Table 9. Of the six Hit data analyses, only three were significant. The 1000-2000 ms timebin was the most significant, AUC = .79, $p < .001$. The 2000-3000 ms timebin was the second most significant, AUC = .748, $p < .01$. The 0-1000 ms timebin was the third most significant, AUC = .682, $p < .05$. Cutoff values presented in Table 9 were selected based on a minimum specificity of 0.80. In addition to being the most statistically significant, the 1000-2000 ms timebin has the most useful cutoff values. By lowering the specificity to .833, a sensitivity of .667 can be achieved, which is higher than any other timebin in this analysis. The 2000-3000 ms timebin has three useable cutoffs, but only two have

a sensitivity of 0.5 and above. While the 0-1000 ms timebin has cutoffs with high specificity, sensitivity values are below 0.4, making them not as useful as cutoffs from other timebins.

The ROC curves for Hit-Miss difference data are presented in Figures 12 and 13, with associated ROC analysis statistics presented in Table 10. For these analyses, greater values were specified to be more indicative of simulator status. Of the six analyses, four were statistically significant. The 5000-6000 ms timebin was the most significant, AUC = .901, $p < .001$. The 3000-4000 ms timebin was second most significant, AUC = .816, $p < .001$. The 4000-5000 ms timebin was the third most significant, AUC = .811, $p < .001$. The fourth significant timebin was 1000-2000 ms, $p < .05$, but its diagnostic value is highly questionable, AUC = .332. Cutoff values presented in Table 10 were selected based on a minimum specificity of 0.80. In addition to being the most statistically significant, the 5000-6000 ms timebin also has a cutoff value with sensitivity and specificity of .833, as well as another with a sensitivity of .625 and a specificity of .958. The 4000-5000 ms timebin also has two usable cutoff values with at least 0.8 specificity and sensitivity in the 0.7 range. Comparatively, the 3000-4000 ms timebin only has one useful cutoff; its specificity of .958 is high, but its sensitivity of .542 is somewhat lacking compared to cutoffs generated for the other timebins.

250 ms level data. Twenty-four ROC analyses each were performed for Hit trial and Hit-Miss difference data from the test phase at the 250 ms level of analysis. Smaller values were specified as more indicative of simulator status for Hit data. The ROC curves for Hit data are presented in Figure 11, and associated ROC analysis statistics are provided in Table 11. Of the 24 Hit data ROCs, 14 were not significant ($p > .05$), one was marginally significant ($p = .07$), and nine were significant ($p < .05$ or less). However, 14 of the 24 analyses (including four significant analyses) had at least one tie between positive and negative state variable groups and

may have yielded biased statistics, leading to their exclusion from further investigation. Therefore, only five timebins were evaluated further for diagnostic cutoff values and are the ones presented in Figure 11 and Table 11. The most statistically significant timebin was 1250-1500 ms, AUC = .797, $p < .001$. The second most significant timebin was 1500-1750 ms, AUC = .757, $p < .01$. The third was 2500-2750 ms, AUC = .738, $p < .01$, the fourth was 750-1000 ms, AUC = .736, $p < .01$, and the fifth was 500-750 ms, AUC = .696, $p < .05$. The cutoff values generated by the ROC analyses for the 500-750 ms and 2500-2750 ms timebins both had sensitivity values below 0.4 at specificity values above 0.8. The 750-1000 ms timebin improved on that with a cutoff with sensitivity of .458 and a specificity of .875. The 1500-1750 ms timebin offered a sensitivity of .583 with a specificity of .833. The 1250-1500 ms, however, offered two of the best cutoffs: sensitivities of .708 and .667 and specificities of .875 and .917, respectively.

The ROC curves for Hit-Miss difference data are presented in Figures 14-16, with associated ROC analysis statistics presented in Table 12. For these analyses, greater values were specified to be more indicative of simulator status. Of the 24 Hit-Miss difference data ROCs, six were not significant ($p > .05$), three were marginally significant ($.05 \leq p \leq .10$), and 15 were significant ($p < .05$ or less). Three significant results and the three marginally significant analyses had AUCs below 0.5, suggesting no diagnostic utility and were excluded from further investigation. Only one timebin (0-250 ms) had at least one tie between positive and negative state variable groups and may have yielded biased statistics, but its analysis was not significant ($p > .05$) despite having an AUC above 0.5. Therefore, only those twelve timebins were evaluated further for diagnostic cutoff values and are the ones presented in Figures 14-16 and Table 12. The remaining twelve timebins had AUCs ranging from .727 (4500-4750 ms) to .939 (5750-6000 ms). Interestingly, all twelve timebins collectively make up the second half of the

trial. Eleven of the twelve significant timebins have at least one useful cutoff score with a sensitivity above 0.5 and a specificity above 0.8. Three timebins (i.e., 4750-5000 ms, 5250-5500 ms, and 5750-6000 ms) each have a cutoff value that boasts a sensitivity and specificity of 0.833.

Diagnostic classification. Using AUC as a criterion, the ROC analyses revealed a number of potentially useful diagnostic variables. Of the numerous available variables, seven in particular will be evaluated for their utility in classifying the participants in this sample. They include Corrected Recognition (CR), proportion of viewing time directed to critical objects during Hit trials at 1000-2000 ms and 1250-1500 ms, and difference in proportion of viewing time directed at critical objects between Hit and Miss trials at 5000-6000 ms, 5750-6000 ms, 5250-5500 ms, and 4750-5000 ms. Each variable was evaluated using its selected cutoff values, as mentioned above for a total of 22 analyses. All relevant variable and diagnostic data is presented in Table 13. Corrected Recognition outperformed every other variable chosen for analysis with its sensitivity of .958 and specificity of 1.0. A variable with such high signal detection characteristics is incredibly rare, such that it only misclassified one of 48 cases (a simulator). Although it was not expected for the behavioral measure to be so effective, no test is inherently perfect. Therefore, it was of practical interest to find a second variable that could supplement Corrected Recognition to make a hypothetically perfect measure in this sample, at least.

False positive rate, or incorrectly labeling a control as a simulator (i.e., Type I error), is negatively tied to specificity and positive predictive value (PPV). On the other hand, false negative rate, or incorrectly labeling a simulator as a control (i.e., Type II error), is negatively tied to sensitivity and negative predictive value (NPV). While sensitivity and specificity are good test evaluation criteria, they are essentially probabilities of detecting a condition .The predictive

values, however, measure the likelihood that the test result matches the person's actual condition state. For this reason, they are considered to be a more real-world examination of the test's utility. For example, Corrected Recognition has a PPV of 1.0 and a NPV of .96, meaning that all cases classified as simulators were actually simulators (23 correct simulators and 0 incorrect controls), but that 96% of cases classified as controls were actually controls (24 correct controls, but one incorrect simulator).

Only 13 of 21 analyses were able to correctly identify the single case misclassified by the Corrected Recognition variable, all coming from five of the six remaining variables. When prioritizing PPV, however, the top three analyses were able to complete the sample. While all three had perfect PPV, their NPV ranged from .774 to .686. The analyses with the three highest NPVs in the sample were not able to complete the sample by correctly identifying the final simulator using their respective cutoffs. The next five analyses, however, were able to complete the sample. These analyses had NPVs ranging from .84 to .808 with PPVs ranging from .87 to .864, respectively. Technically, because the 13 analyses had correctly classified the final simulator case, any one of those variables (using those specific cutoffs) would add incremental utility to the task by boosting its sensitivity and specificity to 1.0.

Discussion

The aim of the present study was to determine if eye movement methodology had utility as a way to identify individuals who were feigning memory impairment. Modern psychological assessment approaches to the question of performance validity largely utilize memory tests and rely on explicit responding. Therefore, it was believed that an indirect measure of recognition could potentially detect malingerers as good as, or even better than, an explicit measure. In the present study, both explicit and implicit data were collected in an effort to compare the efficacy

of the two. However, as no single measure is psychometrically perfect both under control conditions and when applied to the rigors of real-world use, no great expectations were placed on any single measure.

Although the simulator group performed at above-chance levels, not all individual simulators performed above 50%. Volitional maintenance of impaired performance is cognitively draining, such that actively keeping track of correct and incorrect responses is difficult when done alongside the study's actual task. Therefore, not all simulators may have been able to actively adjust their performance on a trial-by-trial basis, whether effectively or not. The posttest questionnaire, presented in Appendix E, revealed that simulators engaged in a variety of behaviors that could potentially impact encoding, retrieval, or both. All options were endorsed at least once and simulators often endorsed multiple strategies. Therefore, it should be noted that not everyone tasked with manipulating their performance by adopting an impaired persona approached the task the same way.

All experimental hypotheses involving eye movement measures were supported by the available data. Significant differences were found in viewing time between critical objects in Hit and Absent trials at the trial level for both groups in the test phase. Simulators also showed differences in viewing time to critical objects between Hit, Miss, and Absent trials, suggesting a potential pattern of viewing behavior resulting from the experimental instructions. In the posttest phase, controls and simulators were not statistically distinguishable using trial-level eye movement data while maintaining differences between Hit and Absent trials that were previously established. This data provides evidence that the trial-level data from eye movements is somewhat sensitive to instructions to feign memory impairment.

The 1000 ms level of analysis provided additional information about when group differences might occur. Both groups directed greater viewing towards critical objects in Hit trials than in Absent trials for nearly the full duration of the trial, consistent with expectations. The earliest timebin to show this difference was 0-1000 ms, providing additional evidence that early-trial eye movements are sensitive to memory. Interestingly, simulators directed significantly less viewing to critical objects in Hit trials 1000-3000 ms following display onset. Additionally, simulators also had differences between critical object viewing for Hit and Miss trials 4000-6000 ms following display onset.

Integrating this information suggests that simulators a) indirectly identify studied objects early in the trial, b) engage in "distributed viewing" (i.e., spending more time viewing other objects or looking off-screen) during the early stages of Hit trials, and c) engage in "distributed viewing" in the late stages of Miss trials. One interpretation of this phenomenon is that simulators are keeping the experimental instructions in mind (i.e., answer incorrectly sometimes) and that distributed viewing is a volitional strategy employed to make the impression that they have not found the studied object. However, because their indirect recognition is intact and occurs very early in the trial, their viewing patterns in the middle and late stages of the trial are actually quite telling of their motivations. The late-trial avoidance of studied objects in Miss trials may very well be a volitional manipulation meant to imply that they have visually identified a non-studied object as the studied one. This is remarkably similar to results reported for when participants are instructed to look away from famous faces, but end up directing greater early viewing to those faces first before avoiding them (Ryan, Hannula, & Cohen, 2007). Again, the early viewing effect allows for the identification of indirect recognition.

Essentially, these departures from expected behavior (when compared to controls) are compared to the individual's explicit responses to determine how useful they are for group classification. However, instead of using a subtest score or an equation to detect malingering, this approach uses the body's own systems to demonstrate which group it belongs to. Of course, the earlier an effect occurs in a trial, the more likely that the effect results from an obligatory process. But just because a more obligatory and likely less cognitively demanding process can be an index of indirect memory, it does not mean that a late-trial difference is useless. Some of the planned eye movement analyses focused on identifying the earliest timebin in which differences were detected. This information is useful experimentally to provide evidence for past research, but can also be used to justify the changing the task design.

For example, the earliest significant differences were all detected within the first 1000 ms of the trial. From a utilitarian perspective and knowing only this information, it would be reasonable to reduce the trial length by at least 50% to enable faster administration. However, there is a wealth of useful information in the second half of the trial that provides insight about the more volitional processes at work during the task. Additionally, going beyond the trial-level data to the 1000 ms level provided useful information about group differences that are only visible through the lens of time.

Going even further, the 250 ms analyses also provided additional evidence to support the earlier finding of the 500-750 ms timebin as the earliest point during a trial in which studied critical objects receive greater viewing, as long as participants are performing at their best. For simulators, the earliest timebin in which this difference occurred was 750-1000 ms. Although a difference of 250 ms may not have immediate practical significance, it has the potential to be a useful diagnostic measure. It should also be noted that the 0-250 ms timebin was not particularly

useful for analysis, as the data was highly variable both between and within-participants, or was missing due to eyetracker error.

From a purely visual standpoint, the 250 ms time courses for controls and simulators in the test phase follow remarkably similar patterns. While the control Hit curve rises faster and maintains greater magnitude, the simulator curve also has a small decrease in Hit trial viewing during the 1750-2000 ms timebin. Also, the two Hit curves almost overlap during the second half of the trial. While simulator Miss data at the 250 ms level was not specifically examined, the 1000 ms level data suggests a continued downward trend in viewing of studied critical objects. Together, this suggests that in the second half of target-present trials, simulators "behave" like controls when they provide a correct response, but noticeably avoid looking at the studied critical object when they provide an incorrect response.

When observing the 250 ms level posttest time course, there is significantly more overlap between the control and simulator Hit curves. At this stage of the study, there was no need for simulators to continue adhering to the experimental instructions. Not only did their explicit recognition accuracy significantly increase, but their viewing behavior also became more control-like. This includes a temporal shift of the early viewing effect in simulators back to the 500-750 ms timebin from where it was in the test phase. Although a marginally significant difference was detected between groups on Hit trials early in the trial (750-1000 ms), that difference was of the largest magnitude found in that time course. The function of the posttest phase was to determine if there were some lasting effects of the experimental instructions on viewing that persist, despite clear instructions to perform optimally. These results suggest that while there may be some influence, it is largely absent and may only be important statistically rather than practically.

It was rather unexpected to see such near-perfect psychometrics from the sole behavioral variable, Corrected Recognition. At the same time, explicit recognition tests are fairly common and one of the most well-researched in psychological assessment. Most, if not all, aspects of the study design were geared towards drawing a clear distinction between performance arising from optimal and suboptimal effort. The set of visual stimuli (Snodgrass & Vanderwart, 1980) was chosen for its ease of identification (of individual stimuli) and were specifically adjusted so as to decrease distractibility (Porter & Troscianko, 2003; Porter et al., 2010). The car crash scenario presented to simulators appeared to result in a measurable change in behavior, as simulators were outperformed by controls, yet did not provide a completely noncredible performance at the group level. The manipulation persisted, as simulators were not able to raise their performance to match that of controls. This suggests that participants may have engaged in flawed learning leading to poor encoding of stimuli. Alternatively, the cognitive effort of following the simulator instructions during the test phase may have altered how the studied stimuli were re-processed and responded to during the posttest phase.

No trial-level variables made for useful diagnostic indicators, which is unsurprising given the amount of information collected over 6000 ms compressed into a single data point. Regarding the remaining useful variables, based on ROC analyses, Hit-Miss difference scores appeared to do better than Hit-only data. This is not necessarily surprising, as late-trial Hit and Miss data at the 1000 ms level suggested that this may be the case. Also, more of the useful variables were from the 250 ms level rather than the 1000 ms level. The 1000 ms level subsumes just enough data points to yield significant differences, but is not as precise about when changes occur. Because significant differences occur early and rapidly, this suggests that more in-depth data analysis was warranted and is consistent with earlier research (Hannula et al., 2007).

Limitations

It should be noted that this is an analog study; participants in the simulator group were instructed to behave differently so that it could be determined whether the measures were of any use diagnostically. While the task yielded several good and useful indicators of malingering, it is certainly possible that the experimental environment influenced participants to fulfill their roles either optimally or suboptimally. The apparent success of the Corrected Recognition variable is surprising, especially for an early task design. It is entirely possible that a larger and more representative sample may yield very good sensitivity and specificity, but not as high as those obtained in the present study.

In a true malingering study, the task would be used with individuals who have a definite motivation to underperform and during the course of a standard evaluation. In such a study, the clinician would have to make a judgment about the individual's group state (i.e., neurologically normal, impaired, or malingering). It would also be against the individual's self-interest to reveal providing suboptimal effort in a situation where they stand to gain something. There would also be no group assignment document that could be referenced to check the accuracy of the clinician's judgment after the fact. Therefore, analog studies should expect to obtain more impressive ROCs than studies using real cases.

In order to make such an important decision, reference data would need to be collected from a variety of groups for the purpose of creating score profiles. In using reference groups in which members have clearly identified conditions, new cases would be compared to each group for best fit. The present study was a preliminary evaluation of the methodology's potential in malingering detection. It is also significantly easier to recruit participants from a college population than neurologically impaired community members. At the same time, in order for a

test or indicator to even be considered viable, demonstrating proof of concept via an analog simulator study is a first step.

The test phase involved reporting an explicit dichotomous decision about recognizing an object in the display. Had the test phase required specific identification of the studied object, similar to the posttest phase, an analysis of viewing time directed to the selected object would have been made. The combination of the early viewing effect for the studied object, as well as the match between greater late viewing and behavioral selection of an incorrect object via the response intention effect (Hannula et al., 2007) could have provided another powerful diagnostic variable.

Looking to Ryan et al. (2007), several modifications to the task could have potentially improved it by allowing for more between-group comparisons. At different stages of the task, individuals would be asked to identify familiar objects, ignore familiar objects, or focus on novel objects with their eyes while measuring viewing time. Following this, recognition and recall tasks would be administered. Such a series of tasks could potentially be highly sensitive to feigned memory impairment, as malingerers would engage in normal viewing patterns during the viewing phases, but demonstrate a response intention effect during the test phases (Hannula et al., 2007). Such a mismatch between implicit and explicit behavior could also have great utility in clinical settings.

Eyetracking equipment and methodology is still currently used almost exclusively in the experimental realm. Any methodology that hopes to match or exceed an industry "gold standard" requires extensive research and development before widespread use and availability. The methodology and applications used in this study are all highly experimental and will require further analysis and design alterations to perfect. At the present time, the time and financial costs

of using a diagnostic eyetracking task may outweigh its benefits. If eyetracking can be shown to be as useful a methodology clinically as it is experimentally, more financially accessible models may be developed in the future. For optimal ease of use, the equipment should be as unintrusive as possible. The eyetracker used in this study did not require a chinrest and allowed the participant to sit in a chair at a comfortable distance away from the screen. This, along with head tracking, allowed participants some freedom to move their heads without compromising the eyetracker's ability to record data was also important during data collection. An eyetracker with good temporal resolution (60 Hz), like the one used here, may be more ecologically appropriate. However, an eyetracker operating at 120 Hz may collect more and higher quality information, at the cost of having to use a chinrest and keeping the head immobile.

Conclusion

All diagnostic tests must be developed with an eye for proper classification of cases and rigorously tested to ensure its viability. The present study was an attempt to develop a task to be used with an eyetracker that would determine if eyetracking methodology could rival some modern malingering detection techniques. Due to its implicit nature and experimental success in identifying viewing effects, it was believed that a clinical application may be possible. This analog simulator study produced a behavioral measure that has very high psychometric viability on its own. Of the multitude of eyetracking variables examined, six were promising, and only five of those were of clinical utility at some cutoff values. Of interest is that the majority of those useful variables were from the late stages of the trial, rather than the early stages where important experimental findings reside. Time course patterns, along with other indices, should be considered when evaluating clinical populations.

As the use of this equipment and methodology with a clinical agenda in mind is still new and experimental, the viable variables and associated cutoff values should not be used as independent indicators of malingering at this time. However, supplementing Corrected Recognition with one of the five eyetracking variables was able to correctly classify all 48 participants in the sample. Therefore, there is some evidence that eyetracking may provide incremental diagnostic utility to identify cases missed by traditional malingering detection measures. Future studies should seek to replicate this methodology with larger samples, but also with known reference groups in order to build an evidence base for this technique. Additional investigation into the task design, such as altering the difficulty or trial length would also be useful. Identifying the best and fewest number of variables needed to make an effective judgment about an individual case would be highly valuable, as well. Because analysis of eyetracking data can be quite time-consuming, automation of the technical aspects of the process would go a long way in demonstrating that eyetracking is easy to use and should improve the method's adoption among researchers and practitioners.

As mentioned earlier, the clinical and experimental approaches to the same problem emphasized different parts of the task in terms of the task's time course. This is an interesting analogy for how practitioners and scientists tend to operate in their separate spaces, seldom interacting. By being aware of both approaches, researchers should be able to appreciate the differential utility of findings based on context and scope. By integrating them, however, researchers stand to gain much more understanding of the populations and methodologies they use.

References

Abe, N., Suzuki, M., Tsukiura, T., Mori, E., Yamaguchi, K., Itoh, M.,…Fujii, T. (2006). Dissociable roles of prefrontal and anterior cingulate cortices in deception. *Cerebral Cortex, 16*(2), 192-199.

Althoff, R. R. (1998). *Eye movement-based memory assessment: The use of eye movement monitoring as an indirect measure of memory.* Unpublished doctoral dissertation, Urbana, University of Illinois at Urbana-Champaign.

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.

Althoff, R. R., Cohen, N. J., McConkie, G., Wasserman, S., Maciukenas, M., Azen, R., …Romine, L. (1998). Eye movement-based memory assessment. In W. Becker, H. Deubel, & T. Mergner (Eds.), *Current oculomotor research: Physiological and psychological aspects* (pp.293-302). New York: Plenum.

Althoff, R. R., Maciukenas, M., & Cohen, N. J. (1993). Indirect assessment of memory using eye movement monitoring. *Society for Neuroscience Abstracts, 19*, 439.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). Ethical principles of psychologists and code of conduct. Washington, DC: Author.

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology, 103*(1), 62-70.

Basso, M. R., Miller, A., Estevis, E., & Combs, D. (2013). Neuropsychological deficits in major depressive disorder: Correlates and conundrums. In P. A. Arnett (Ed.), *Secondary influences on neuropsychological test performance* (pp. 39-66). New York, NY: Oxford University Press.

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*, 1-11.

Bowman, H., Filetti, M., Alsufyani, A., Janssen, D., & Su, L. (2014). Countering countermeasures: Detecting identity lies by detecting conscious breakthrough. *PLoS ONE, 9*(3): e90595, 1-17.

Browndyke, J. N., Paskavitz, J., Sweet, L. H., Cohen, R. A., Ticker, K. A., Welsh-Bohmer, K. A.,…Schmechel, D. E. (2008). Neuroanatomical correlates of malingered memory impairment: Event-related fMRI of deception on a recognition memory task. *Brain Injury, 22*(6), 481-489.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., … Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity. *Archives of Clinical Neuropsychology, 20*, 419-426.

Buswell, G. T. (1935). *How people look at pictures.* Chicago, IL: University of Chicago Press.

Chafetz, M., & Underhill, J. (2013). Estimated costs of malingered disability. *Archives of Clinical Neuropsychology, 28*, 633-639.

Chua, E. F., Hannula, D. E., & Ranganath, C. (2012). Distinguishing highly confident accurate and inaccurate memory: Insights about relevant and irrelevant influences on memory confidence. *Memory, 20*(1), 48-62.

Clarke, P. & Macleod, C. (2013). The impact of anxiety on cognitive task performance. In P. A. Arnett (Ed.), *Secondary influences on neuropsychological test performance* (pp. 93-116). New York, NY: Oxford University Press.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system.* Cambridge, MA: MIT Press.

Coleman, R. D., Rapport, L. J., Millis, S. R., Ricker, J. H., & Farchione, T. J. (1998). Effects of coaching on detection of malingering on the California Verbal Learning Test. *Journal of Clinical and Experimental Neuropsychology*, *20*, 201–210.

DenBoer, J. W., & Hall, S. (2007). Neuropsychological test performance of successful brain injury simulators. *The Clinical Neuropsychologist, 21*, 943-955.

DiCarlo, M. A., Gfeller, J. D., & Oliveri, M. V. (2000). Effects of coaching on detecting feigned cognitive impairment with the Category Test. *Archives of Clinical Neuropsychology*, *15*, 399–413.

Dunn, T. M., Shear, P. K., Howe, S., & Ris, M. D. (2003). Detecting neuropsychological malingering: Effects of coaching and information. *Archives of Clinical Neuropsychology, 18*, 121-134.

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: An fMRI investigation. *Cerebral Cortex, 13*, 830-836.

Gardner, R. M., Mo, S. S., & Borrego, R. (1974). Inhibition of pupillary orienting reflex by novelty in conjunction with recognition memory. *Bulletin of the Psychonomic Society, 3*(3B), 237-238.

Gorny, I., & Merten, T. (2005). Symptom information-warning-coaching: How do they affect successful feigning in neuropsychological assessment? *Journal of Forensic Neuropsychology, 4*(4), 71-97.

Green, P. (2005). *Word memory test for Windows: User's manual and program.* Edmonton, Alberta: Green's Publishing.

Hannula, D. E., Baym, C. L., Warren, D. E., & Cohen, N. J. (2012). The eyes know: Eye movements as a veridical index of memory. *Psychological Science, 23*(3), 278-287.

Hannula, D. E., Althoff, R. R., Warren, D. E., Riggs, L., Cohen, N. J., & Ryan, J. D. (2010a). Worth a glance: Using eye movements to investigate the cognitive neuroscience of memory. *Frontiers in Human Neuroscience, 4*(166), 1-16.

Hannula, D. E., Ranganath, C., Ramsay, I. S., Solomon, M., Yoon J., Niendam, T. A., … Ragland, J. D. (2010b). Use of eye movement monitoring to examine item and relational memory in schizophrenia. *Biological Psychiatry, 68*, 610-616.

Hannula, D. E., Ryan, J. D., Tranel, D., & Cohen, N. J. (2007). Rapid onset relational memory effects are evident in eye movement behavior, but not in hippocampal amnesia. *Journal of Cognitive Neuroscience, 19*(10), 1690-1705.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference Participants. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23* (7), 1093-1129.

Henderson, J. M., Weeks, P. A., Jr. & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 210-228.

Iverson, G. (1995). Qualitative aspects of malingered memory deficits. *Brain Injury, 9*,35-40.

Kafkas, A., & Montaldi, D. (2012). Familiarity and recollection produce distinct eye movement, pupil and medial temporal lobe responses when memory strength is matched. *Neuropsychologia, 50*, 3080-3093.

Lamb, D. G., Berry, D. T. R., Wetter, M. W., & Baer, R. A. (1994). Effects of two types of information on malingering of closed head injury on the MMPI-2: An analog investigation. *Psychological Assessment*, *6*, 8–13.

Langleben, D. D., Loughead, J. W., Bilker, W. B., Ruparel. K., Childress, A. R., Busch, S. I.,…Gur, R. C. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. *Human Brain Mapping, 26*(4), 262-272.

Larrabee, G. J. (2011). Assessment of malingering. In G. J. Larrabee (Ed.). *Forensic neuropsychology: A scientific approach* (2nd ed., pp. 116-159). New York: Oxford University Press.

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*, 1-7.

Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2009). 40 plus or minus 10, a new magical number: Reply to Russell. *The Clinical Neuropsychologist, 23*, 841–849.

Lee, T. M., Liu, H. L., Tan, L. H., Chan, C. C., Mahankali, S., Feng, C. M.,…Gao, J. H. (2002). Lie detection by functional magnetic resonance imaging. *Human Brain Mapping, 15*(3), 157-164.

Lees-Haley, P. R. (1997). Attorneys influence expert evidence in forensic psychological and neuropsychological cases. *Assessment*, *4*, 321–324.

Loftus, G., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 565-572.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics, 2*(11), 547-552.

Martin, R. C., Bolter, J. F., Todd, M. E., Gouvier, W. D., & Niccolls, R. (1993). Effects of sophistication and motivation on the detection of malingered memory performance using a computerized, forced-choice task. *Journal of Clinical and Experimental Neuropsychology, 15*, 867–880.

Millis, S.R. (2009). Methodological challenges in assessment of cognition following mild head injury: Response to Malojcic et al. (2008). *Journal of Neurotrauma, 26*, 2409–2410.

Mittenberg, W., Patton., C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*(8), 1094-1102.

Mohamed, F. B., Faro, S. H., Gordon, N. J., Platek, S. M., Ahmad, H., & Williams, J. M. (2006). Brain mapping of deception and truth telling about an ecologically valid situation: Functional MR imaging and polygraph investigation-initial experience. *Radiology, 238*(2), 679-688.

Neuropsychology Model LCD Taskforce. (2011, July). *Neuropsychology Model LCD.* Retrieved from https://www.theaacn.org/userdocuments/neuropsychology_model_lcd-1.pdf

Nuñez, J. M., Casey, B. J., Egner, T., Hare, T., & Hirsch, J. (2005). Intentional false responding shares neural substrates with response conflict and cognitive control. *Neuroimage, 25*(1), 267-277.

Office of the Inspector General. (2013). The Social Security Administration's policy on symptom validity tests in determining disability claims. Retrieved from http://oig.ssa.gov/sites/default/files/audit/full/pdf/A-08-13-23094.pdf

Porter, G., Tales, A., Troscianko, T., Wilcock, G., Haworth, J., & Leonards, U. (2010). New insights into feature and conjunction search: I. Evidence from pupil size, eye movements and ageing. *Cortex, 46*, 621-636.

Porter, G., & Troscianko, T. (2003). Pupillary response to grating stimuli. *Perception, 32*, 156.

Rapport, L. J., Farchione, T. J., Coleman, R. D., & Axelrod, B. N. (1998). Effects of coaching on malingered motor function profiles. *Journal of Clinical and Experimental Neuropsychology*, *20*, 89–97.

Richman, J., Green, P., Gervais, R., Flaro, L., Merten, T., Brockhaus, R., & Ranks, D. (2006). Objective tests of symptom exaggeration in independent medical examinations. *Journal of Occupational and Environmental Medicine, 48*, 303-311.

Rose, F. E., Hall, S.,& Szalda-Petree, A. D. (1995). Portland Digit Recognition Test-computerized: Measuring the response latency improves the detection of malingering. *Clinical Neuropsychologist*, *9*, 124–134.

Rose, W. E., Hall, S., Szalda-Petree, A. D., & Bach, P. J. (1998). A comparison of four tests of malingering and the effects of coaching. *Archives of Clinical Neuropsychology, 13*(4), 349-363.

Rüsseler, J., Brett, A., Klaue, U., Sailer, M., & Münte, T. F. (2008). The effect of coaching on the simulated malingering of memory impairment. *BMC Neurology, 8*(37).

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.

Ryan, J. D., Hannula, D. E., & Cohen, N. J., (2007). The obligatory effects of memory on eye movements. *Memory, 15*(5), 508-525.

Schwartz, S. M., Gramling, S. E., Kerr, K. L., & Morin, C. (1998). Evaluation of intellect and deficit specific information on the ability to fake memory deficits. *International Journal of Law and Psychiatry*, *21*, 261–272.

Schwedes, C., & Wentura, D. (2012). The revealing glance: Eye gaze behavior to concealed information. *Memory & Cognition, 40*, 642-651.

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13*, 545-561.

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.

Spence, S. A., Farrow, T. F., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport, 12*, 2849-2853.

Spence, S. A., Hunter, M. D., Farrow, T. F., Green, R. D., Leung, D. H., Hughes, C. J.,…Ganesan, V. (2004). A cognitive neurobiological account of deception: Evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 359*(1451), 1755-1762.

Stark, L., & Ellis, S. R. (1981). Scanpaths revisited. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception* (pp. 193-226). Hillsdale, NJ: Erlbaum.

Stone, D. C., & Boone, K. B. (2007). Feigning of physical, psychiatric, and cognitive symptoms: Examples from history, the arts, and animal behavior. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment* (pp. 3-12). New York, NY: Guilford Press.

Suhr, J. A., & Gunstad, J. (2000). The effects of coaching on the sensitivity and specificity of malingering measures. *Archives of Clinical Neuropsychology*, *15*, 415–424.

Tan, J. E., Slick, D. J., Strauss, E., & Hultsch, D. F. (2002). How'd they do it? Malingering strategies on symptom validity tests. *The Clinical Neuropsychologist, 16*(4), 495-505.

Vendemia, J. M., Buzan, R. F., & Simon-Dack, S. L. (2005). Reaction time of motor responses in two-stimulus paradigms involving deception and congruity with varying levels of difficulty. *Behavioral Neurology, 16*, 25-36.

Williams, L. E., Must, A., Avery, S., Woolard, A., Woodward, N. D., Cohen, N. J.,…Heckers, S. (2010). Eye-movement behavior reveals relational memory impairment in schizophrenia. *Biological Psychiatry, 68*, 617-624.

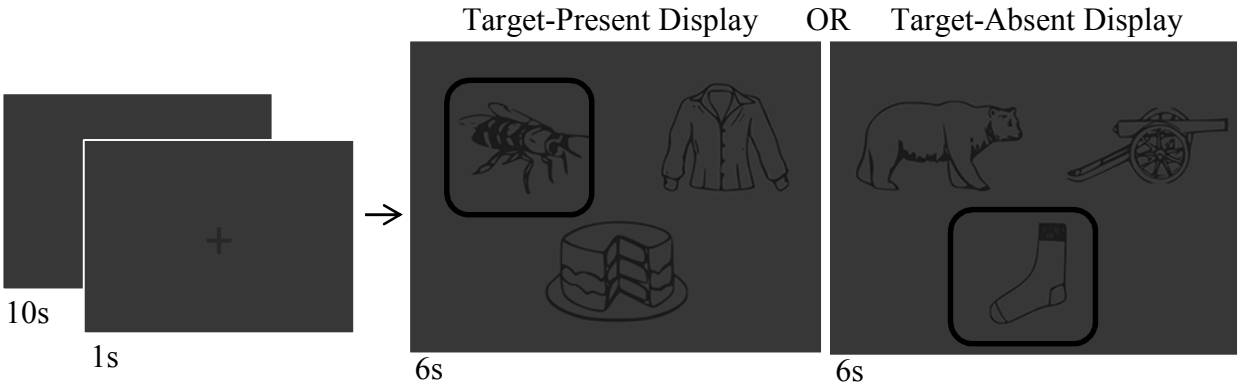
Willison, J., & Tombaugh, T. N. (2006). Detection simulation of attention deficits using reaction time tests. *Archives of Clinical Neuropsychology, 21*, 41-52.

Youngjohn, J. R. (1995). Confirmed attorney coaching prior to neuropsychological evaluation. *Assessment, 2*, 279–283.

Ziegler, E. A., & Boone, K. B. (2013). Symptom invalidity on neuropsychological testing. In P. A. Arnett (Ed.), *Secondary influences on neuropsychological test performance* (pp. 7-38). New York, NY: Oxford University Press.

STUDY PHASE
Two blocks – 24 objects seen once per block
No responses from participant

TEST PHASE
One block – 48 three-object displays (24 Target-Present / 24 Target-Absent)



Behavioral Response: Is a studied object present in the display?

Highlighted objects indicate critical objects used for calculation of proportion of viewing time. For example, the bee was one of the stimuli in this participant's study phase, so it is used as the critical item in a target-present trial. The blouse and cake were not in the study phase and serve as distractors. All three stimuli in the target-absent display did not appear in the study phase. After counterbalancing, the sock was selected to be the critical item for this display. Note that none of the three objects in this display should have an advantage over the other two based on previous viewing, as all three are novel. Highlighting is for illustration only, and was not present in actual trials.

POSTTEST PHASE
One block – 48 three-object displays (24 Target-Present / 24 Target-Absent)
Identical displays from test phase presented in re-randomized order
Behavioral Response: Where is the studied object located in the display? (Left, Right, Bottom)
Verbal Response: Is a studied object present in the display? (Yes or No)
If no studied object is present, instructions were to press any button, then respond, "No."

Figure 1. Visual Illustration and Brief Descriptions of Task Phases

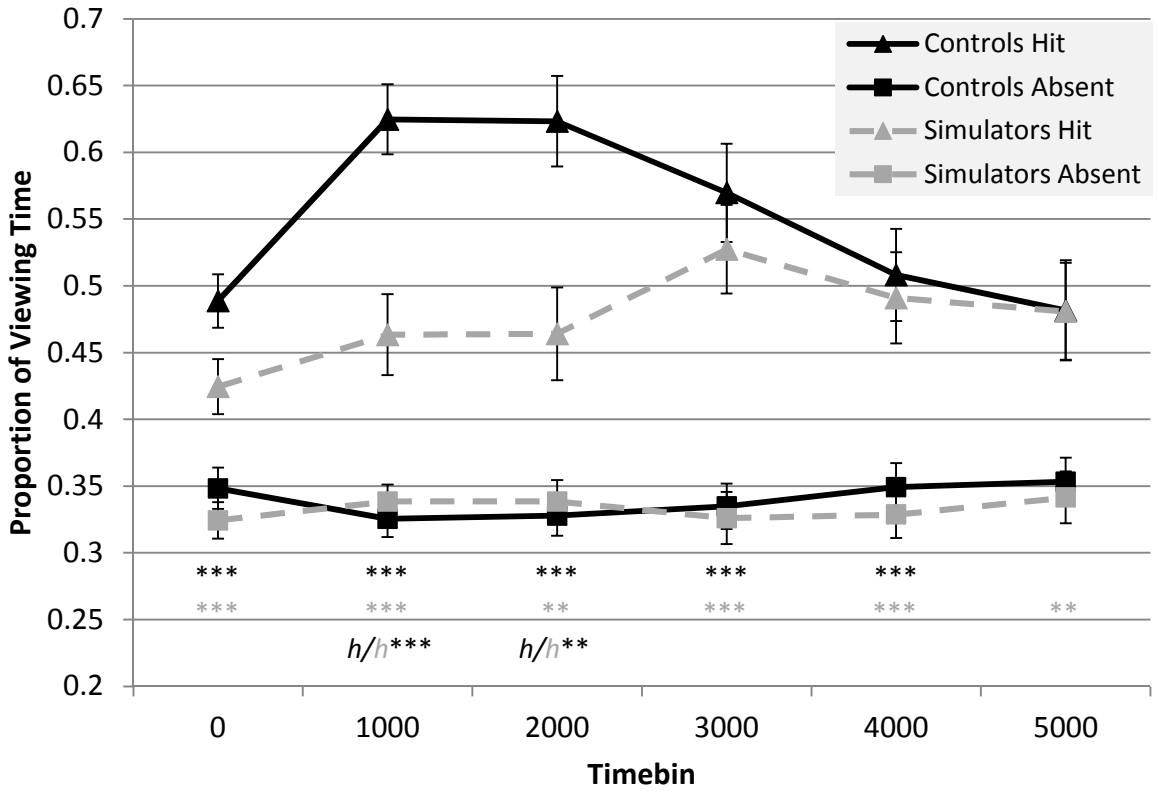


Figure 2. Proportion of Viewing Time Directed at Critical Objects in Hit and Absent Trials by Controls and Simulators in the Test Phase Parsed into 1000 ms Timebins. Error bars represent Standard Error of the Mean (SEM). Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-1000 ms following display onset). ** $p < .01$, *** $p < .001$, *h* denotes Hit trials.

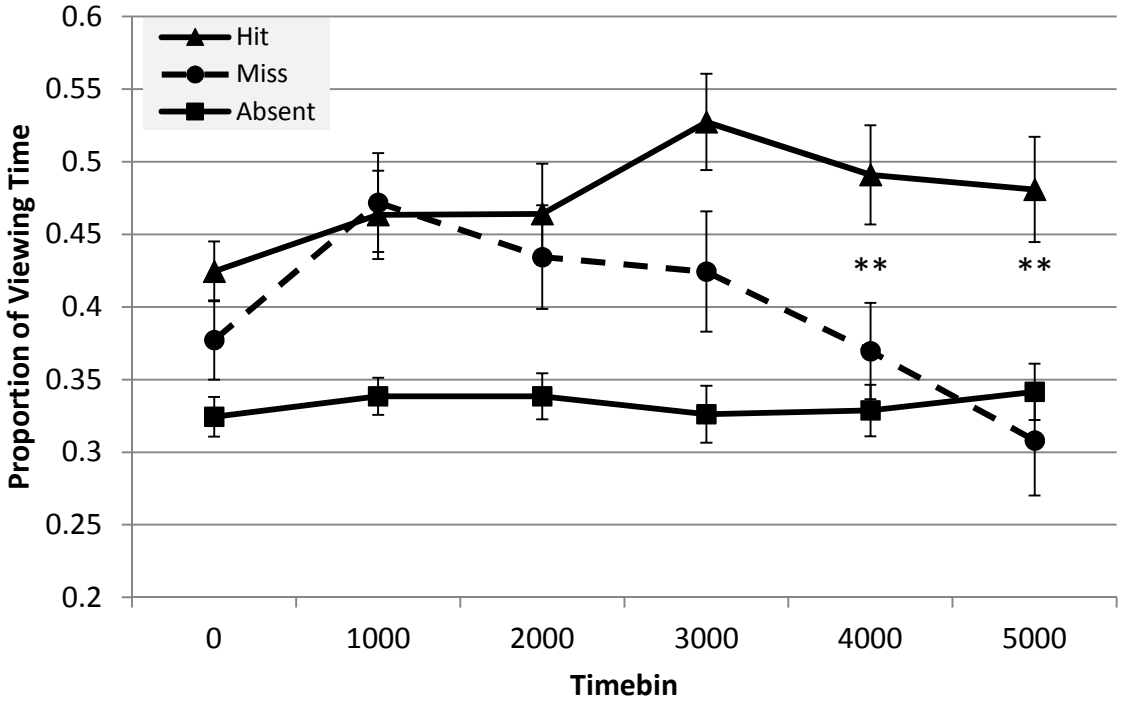


Figure 3. Proportion of Viewing Time Directed at Critical Objects in Hit, Miss, and Absent Trials by Simulators in the Test Phase Parsed into 1000 ms Timebins. Error bars represent Standard Error of the Mean (SEM). Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-1000 ms following display onset). ** $p < .01$.

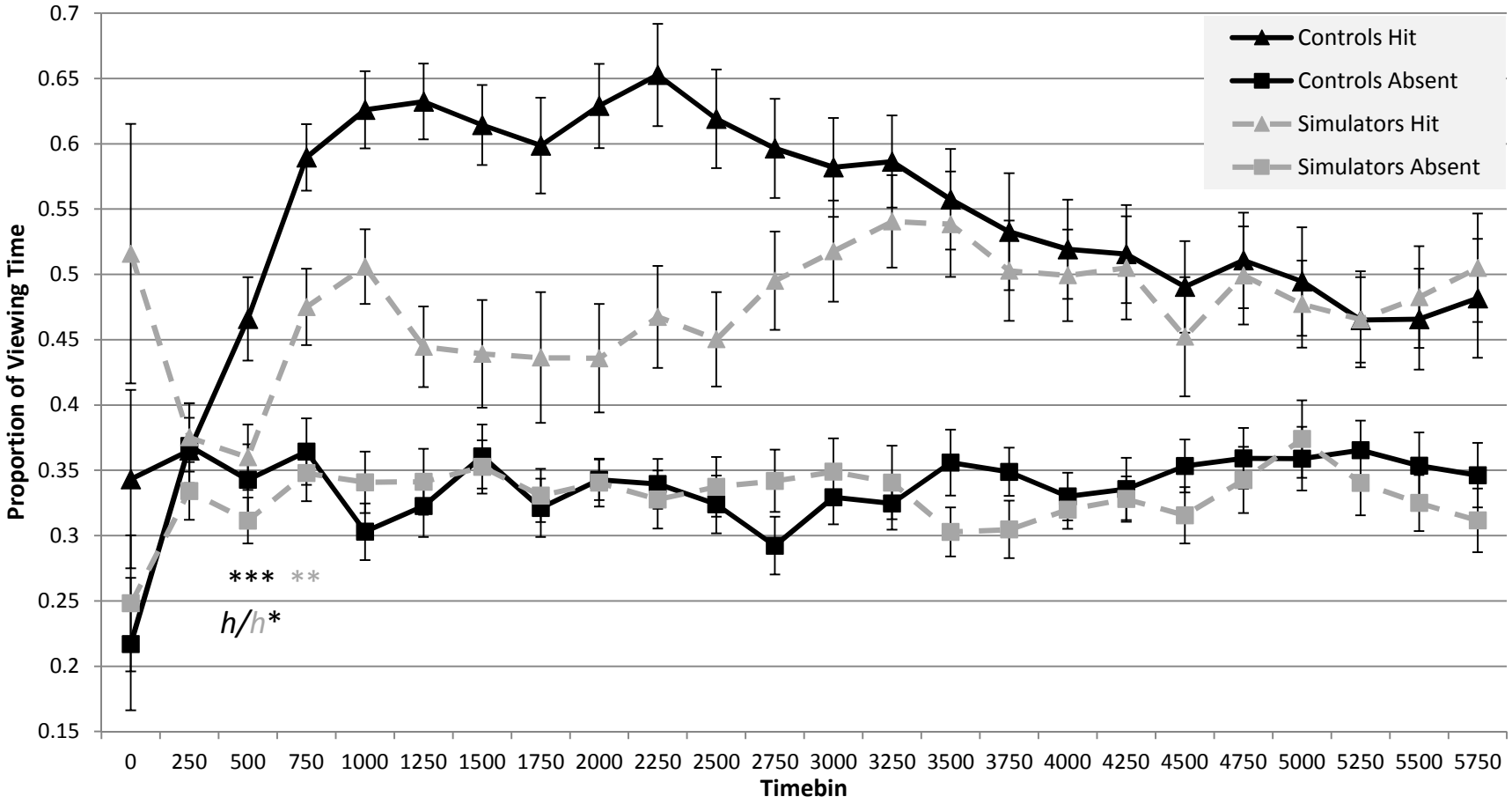


Figure 4. Proportion of Viewing Time Directed at Critical Objects in Hit and Absent Trials by Controls and Simulators in the Test Phase Parsed into 250 ms Timebins. Error bars represent Standard Error of the Mean (SEM). Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset). * $p < .05$, ** $p < .01$, *** $p < .001$, *h* denotes Hit trials.

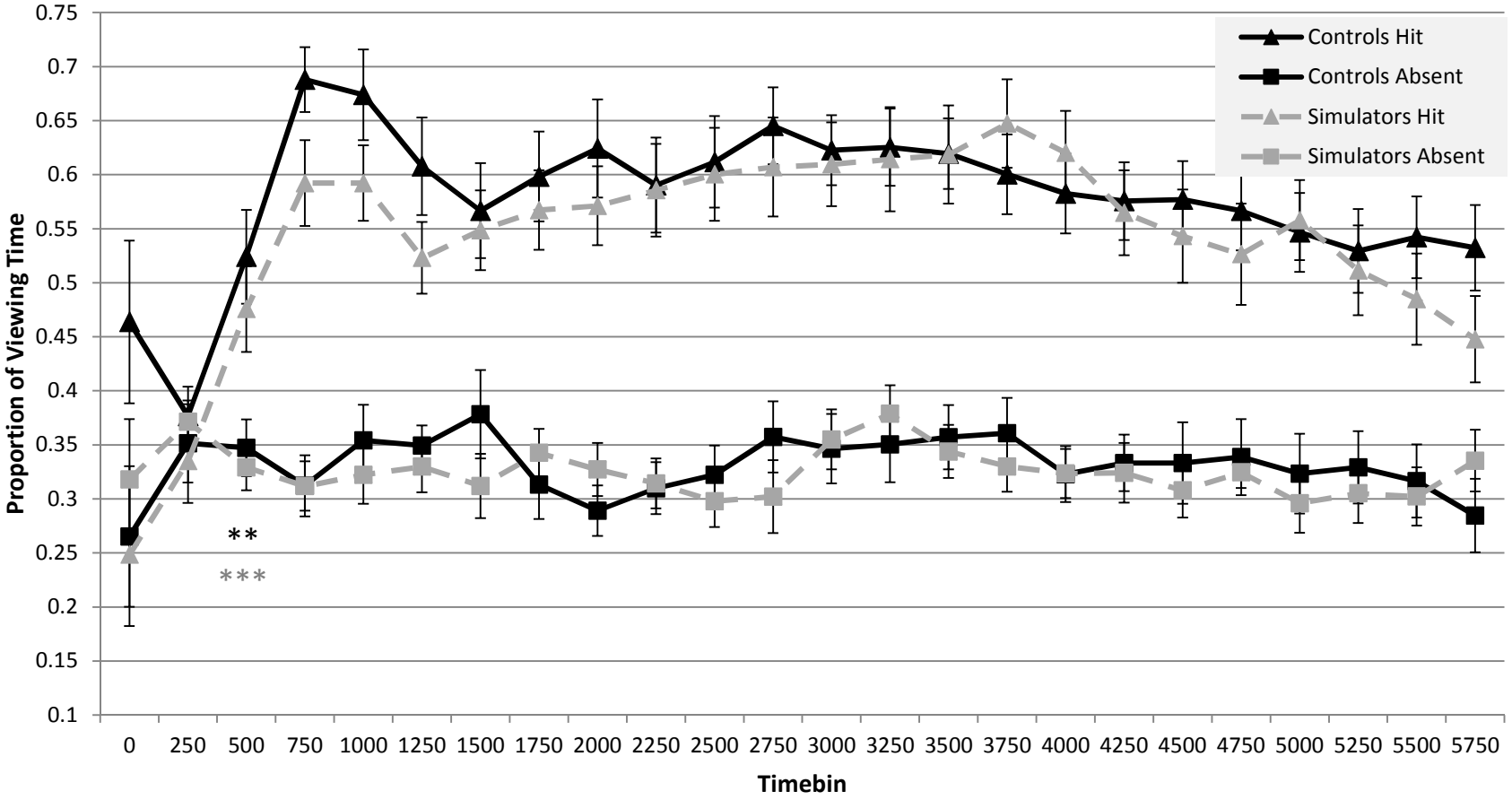


Figure 5. Proportion of Viewing Time Directed at Critical Objects in Hit and Absent Trials by Controls and Simulators in the Posttest Phase Parsed into 250 ms Timebins. Error bars represent Standard Error of the Mean (SEM). Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset). ** $p < .01$, ** $p < .001$.

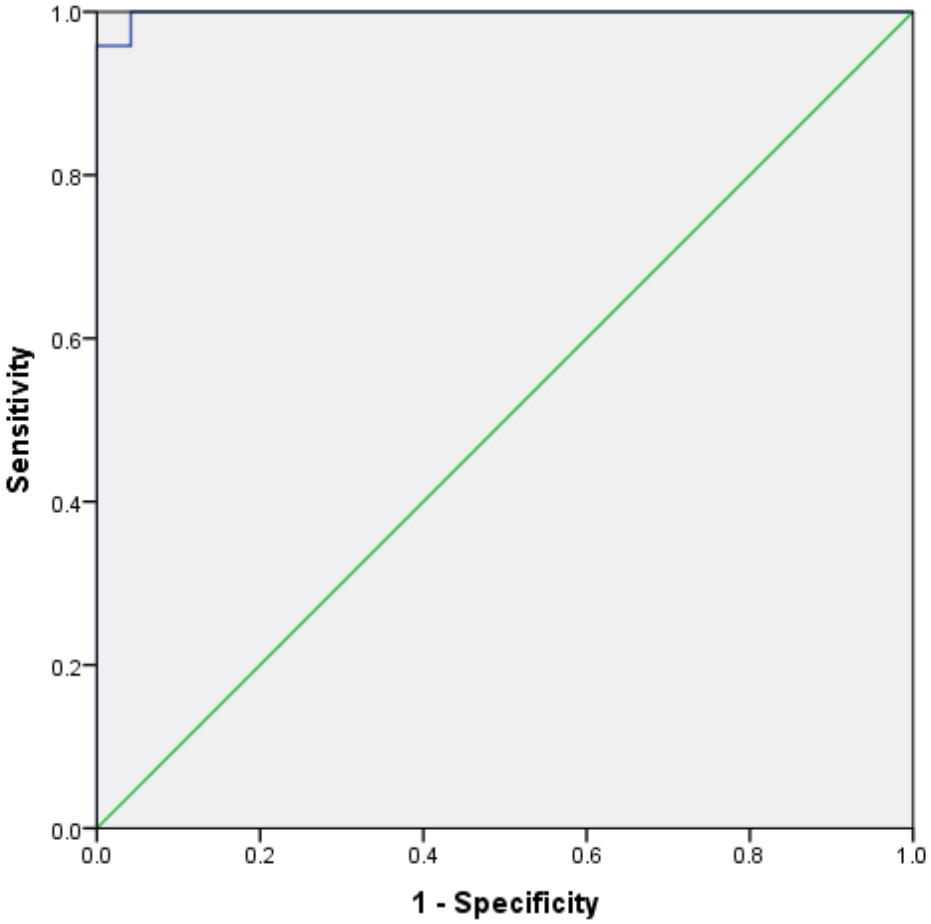


Figure 6. ROC Curve for Corrected Recognition Data.

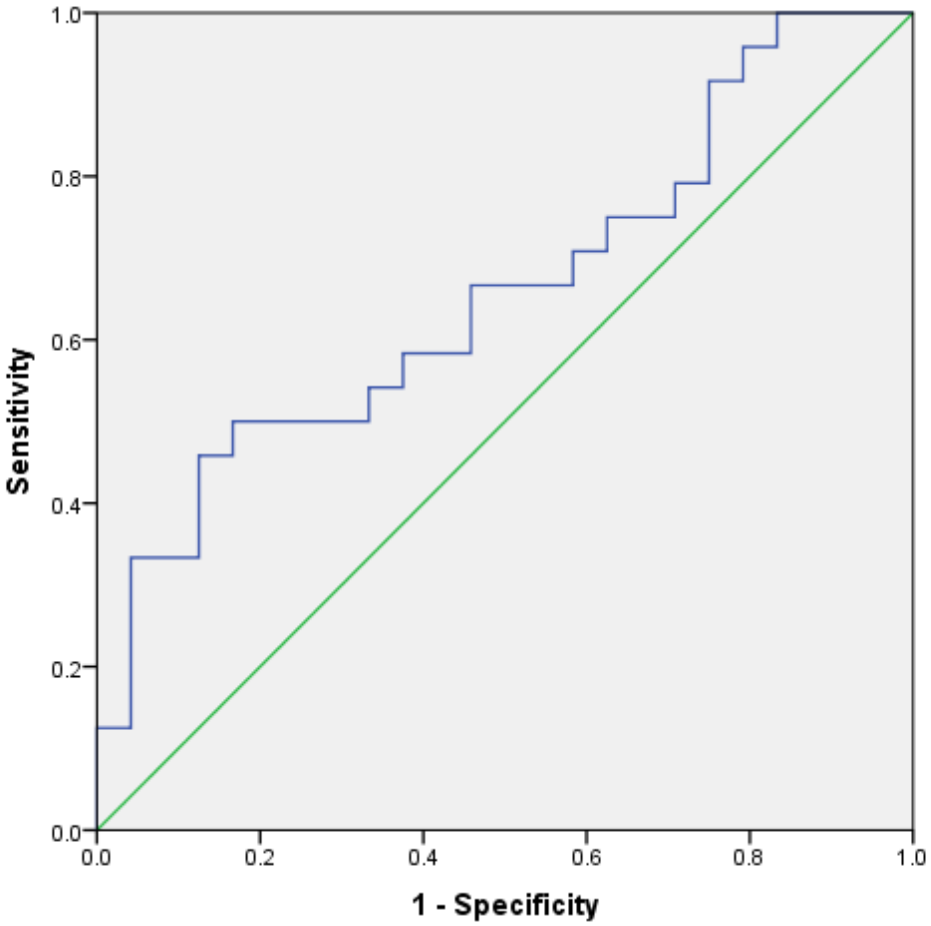


Figure 7. ROC Curve for Proportion of Viewing Time Directed at Critical Objects During Hit Trials at the Trial Level.

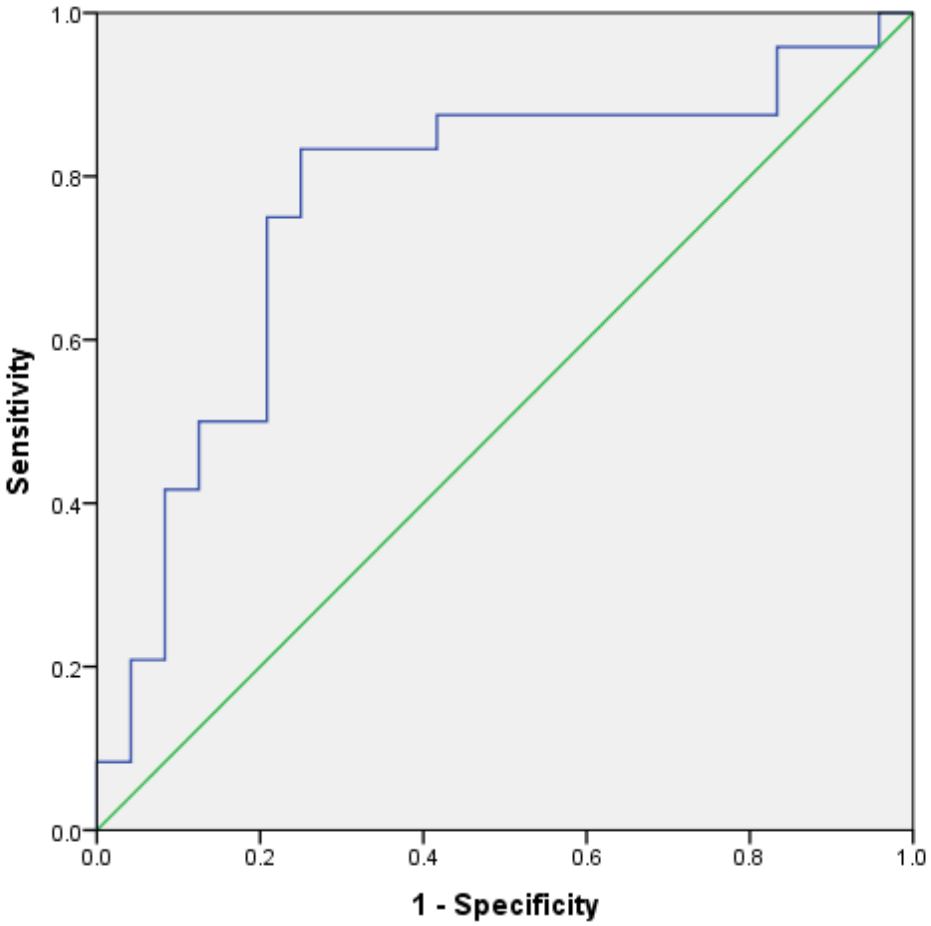


Figure 8. ROC Curve for Hit-Miss Difference in Proportion of Viewing Time Directed at Critical Objects During Hit Trials at the Trial Level.

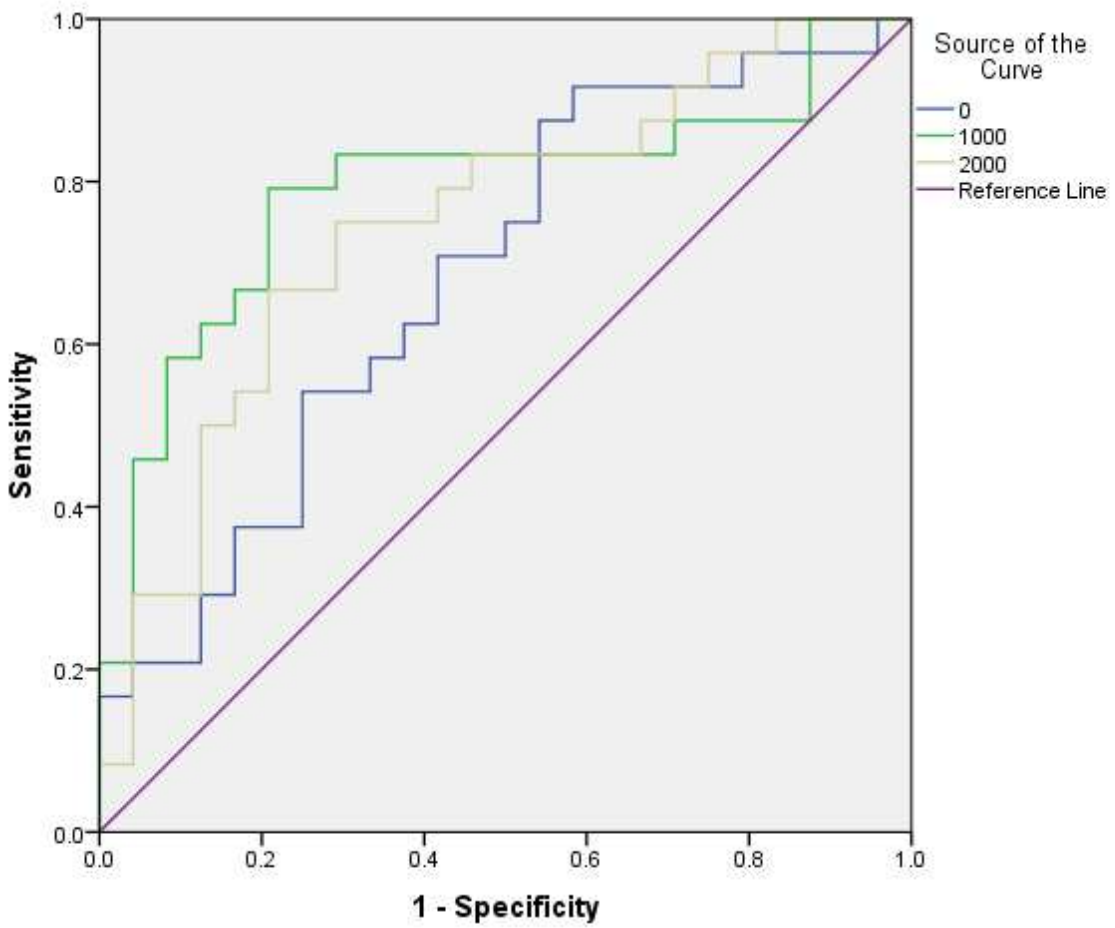


Figure 9. ROC Curve for Proportion of Viewing Time Directed at Critical Objects During Hit Trials at the 1000 ms Level (0, 1000, and 2000 ms timebins).

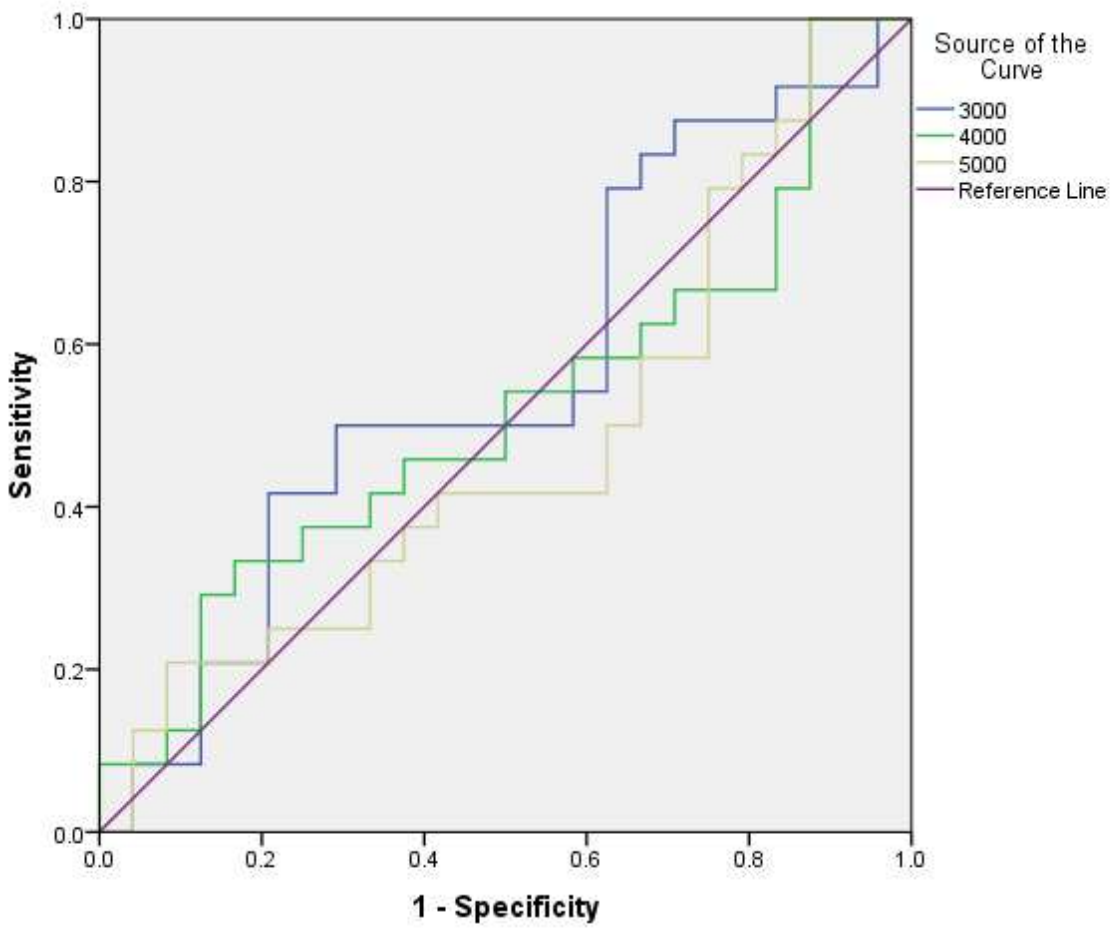


Figure 10. ROC Curve for Proportion of Viewing Time Directed at Critical Objects During Hit Trials at the 1000 ms Level (3000, 4000, and 5000 ms timebins).

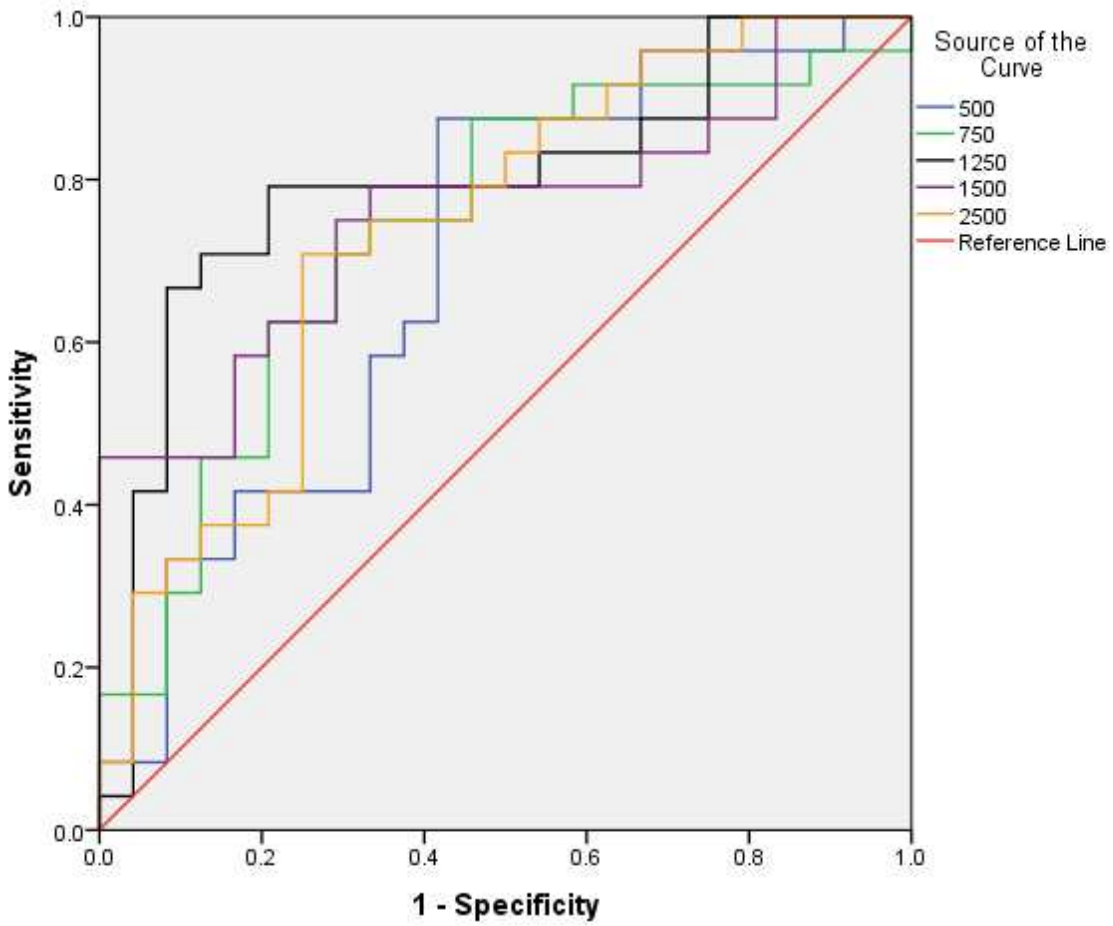


Figure 11. ROC Curve for Proportion of Viewing Time Directed at Critical Objects During Hit Trials at the 250 ms Level (500, 750, 1250, 1500, and 2500 ms timebins).

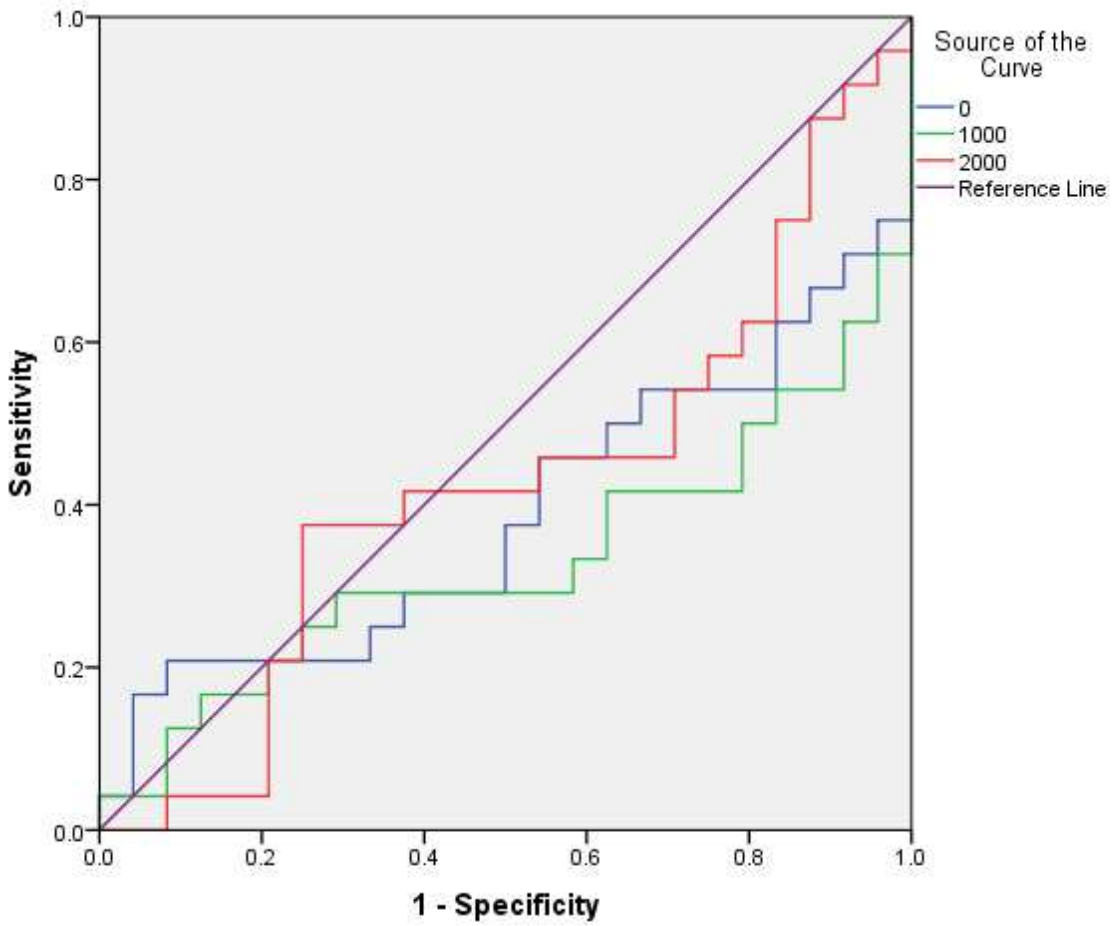


Figure 12. ROC Curve for Hit-Miss Difference in Proportion of Viewing Time Directed at Critical Objects at the 1000 ms Level (0, 1000, and 2000 ms timebins).

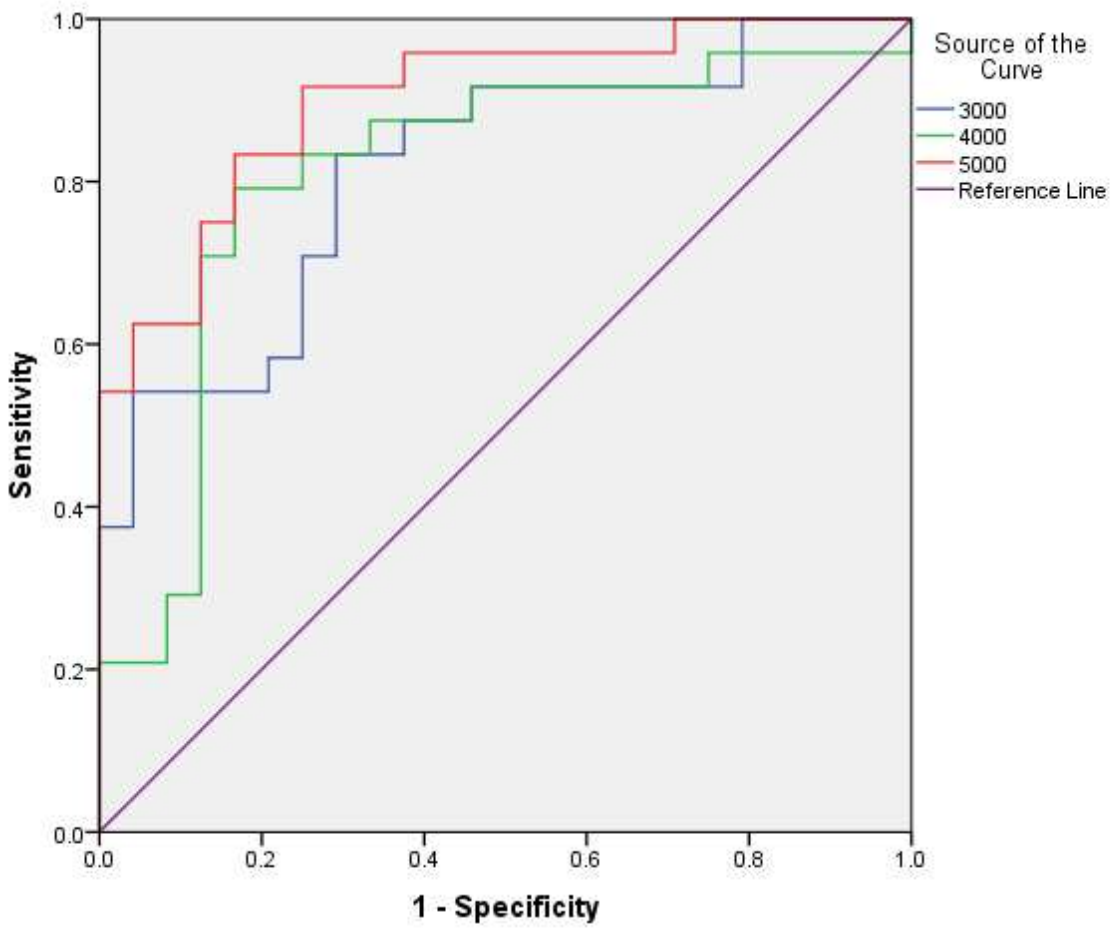


Figure 13. ROC Curve for Hit-Miss Difference in Proportion of Viewing Time Directed at Critical Objects at the 1000 ms Level (3000, 4000, and 5000 ms timebins).

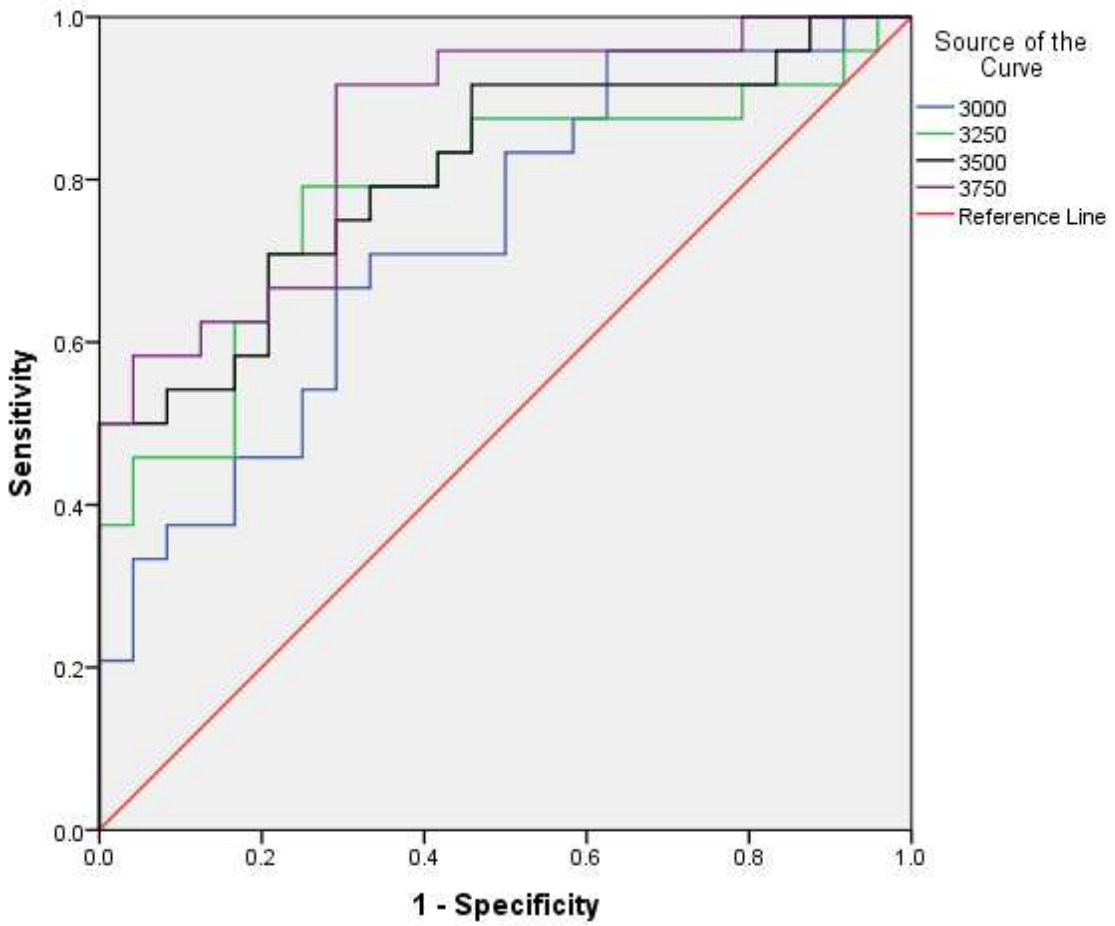


Figure 14. ROC Curve for Hit-Miss Difference in Proportion of Viewing Time Directed at Critical Objects at the 250 ms Level (3000, 3250, 3500, and 3750 ms timebins).

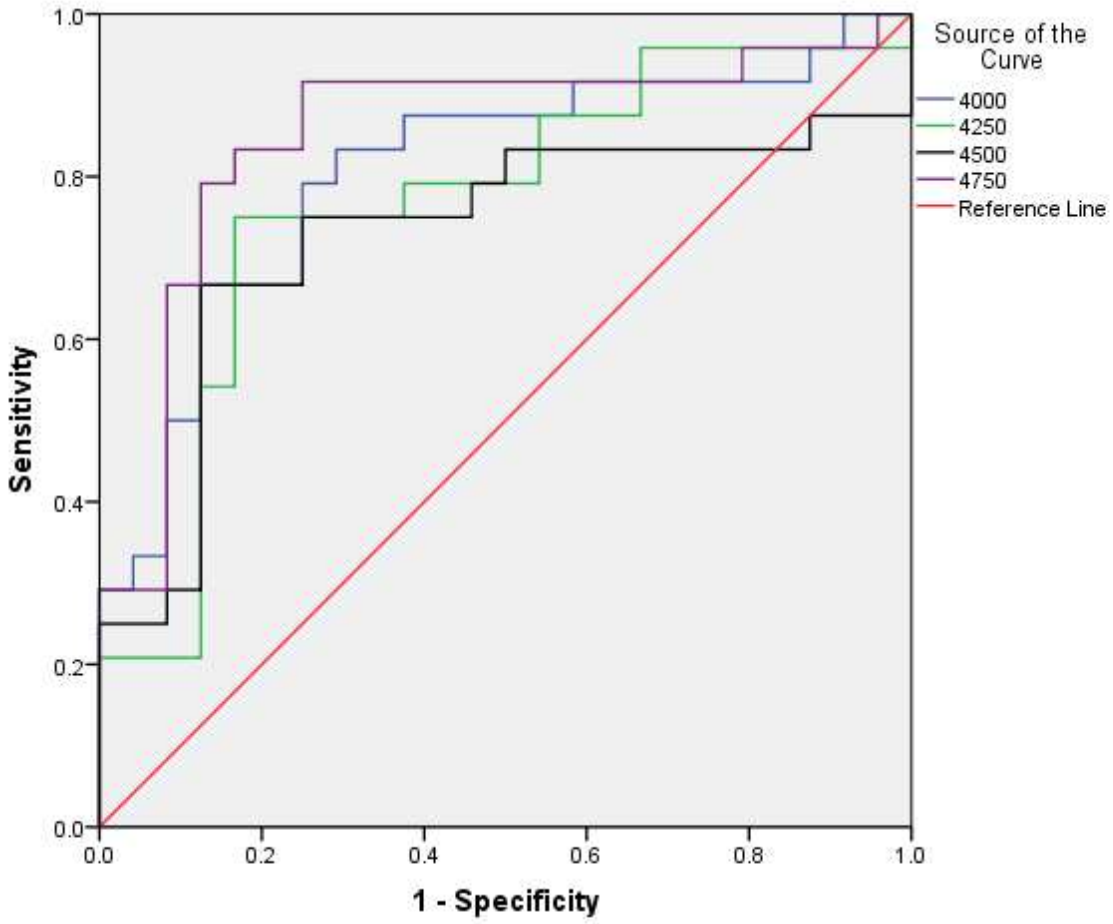


Figure 15. ROC Curve for Hit-Miss Difference in Proportion of Viewing Time Directed at Critical Objects at the 250 ms Level (4000, 4250, 4500, and 4750 ms timebins).

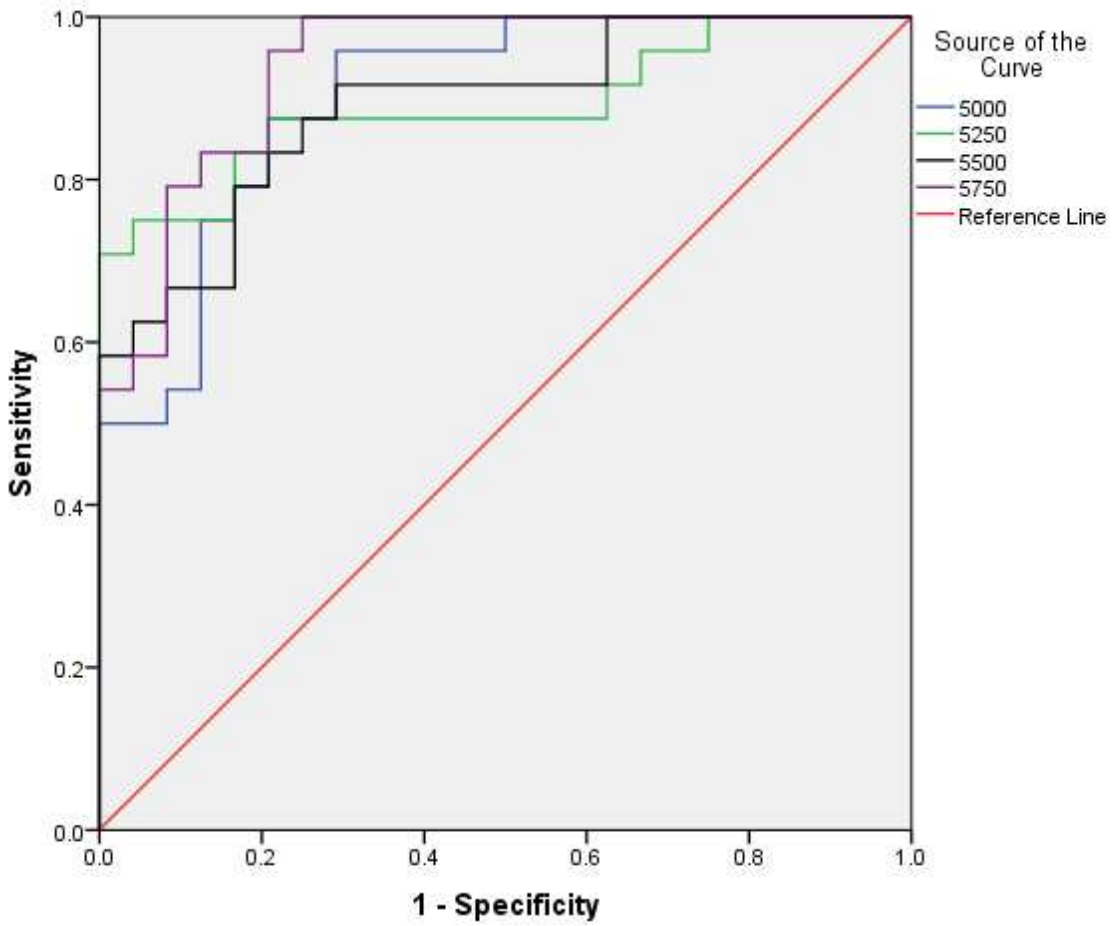


Figure 16. ROC Curve for Hit-Miss Difference in Proportion of Viewing Time Directed at Critical Objects at the 250 ms Level (5000, 5250, 5500, and 5750 ms timebins).

Table 1

Mean Proportion of Viewing Time Directed at Critical Objects Across Trial Types by Controls and Simulators in the Test Phase

		Timebin						
		0	1000	2000	3000	4000	5000	Trial
Controls	Hit	0.489 (0.02)	0.625 (0.026)	0.623 (0.034)	0.570 (0.037)	0.508 (0.035)	0.482 (0.038)	0.560 (0.025)
	Absent	0.348 (0.016)	0.325 (0.014)	0.328 (0.015)	0.335 (0.017)	0.349 (0.018)	0.353 (0.018)	0.339 (0.009)
Simulators	Hit	0.425 (0.021)	0.463 (0.03)	0.464 (0.035)	0.527 (0.033)	0.491 (0.034)	0.481 (0.036)	0.483 (0.025)
	Miss	0.377 (0.027)	0.472 (0.034)	0.434 (0.036)	0.424 (0.041)	0.370 (0.033)	0.308 (0.038)	0.398 (0.023)
	Absent	0.324 (0.014)	0.338 (0.013)	0.339 (0.016)	0.326 (0.02)	0.329 (0.018)	0.342 (0.019)	0.333 (0.009)

Note. Values in parentheses are Standard Error of the Mean. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-1000 ms following display onset).

Table 2

Mean Proportion of Viewing Time Directed at Critical Objects Across Trial Types by Controls and Simulators in the Test Phase at the 250 ms Level (0-3000 ms)

		Timebin											
		0	250	500	750	1000	1250	1500	1750	2000	2250	2500	2750
Controls	Hit	0.343 (0.068)	0.365 (0.025)	0.466 (0.032)	0.59 (0.025)	0.626 (0.029)	0.632 (0.029)	0.614 (0.031)	0.599 (0.037)	0.629 (0.032)	0.653 (0.039)	0.619 (0.038)	0.597 (0.038)
	Absent	0.217 (0.051)	0.369 (0.032)	0.343 (0.027)	0.364 (0.025)	0.303 (0.022)	0.323 (0.024)	0.361 (0.025)	0.321 (0.022)	0.343 (0.016)	0.34 (0.019)	0.324 (0.022)	0.292 (0.022)
Simulators	Hit	0.516 (0.099)	0.375 (0.026)	0.36 (0.025)	0.475 (0.029)	0.506 (0.029)	0.445 (0.031)	0.439 (0.041)	0.436 (0.05)	0.436 (0.042)	0.467 (0.039)	0.45 (0.036)	0.495 (0.038)
	Miss	0.315 (0.092)	0.351 (0.04)	0.377 (0.03)	0.415 (0.044)	0.467 (0.051)	0.449 (0.043)	0.474 (0.041)	0.451 (0.036)	0.445 (0.039)	0.43 (0.05)	0.405 (0.035)	0.414 (0.033)
	Absent	0.248 (0.052)	0.334 (0.022)	0.312 (0.018)	0.348 (0.022)	0.341 (0.024)	0.341 (0.025)	0.353 (0.02)	0.331 (0.02)	0.341 (0.018)	0.328 (0.022)	0.337 (0.023)	0.342 (0.024)

Note. Values in parentheses are Standard Error of the Mean. Values in parentheses are Standard Error of the Mean. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset).

Table 3

Mean Proportion of Viewing Time Directed at Critical Objects Across Trial Types by Controls and Simulators in the Test Phase at the 250 ms Level (3000-6000 ms)

		Timebin											
		3000	3250	3500	3750	4000	4250	4500	4750	5000	5250	5500	5750
Controls	Hit	0.582 (0.038)	0.586 (0.035)	0.558 (0.038)	0.533 (0.045)	0.519 (0.038)	0.516 (0.038)	0.49 (0.035)	0.511 (0.037)	0.495 (0.042)	0.465 (0.033)	0.466 (0.039)	0.482 (0.046)
	Absent	0.329 (0.021)	0.325 (0.02)	0.356 (0.025)	0.349 (0.018)	0.33 (0.018)	0.336 (0.024)	0.353 (0.02)	0.359 (0.023)	0.359 (0.024)	0.365 (0.023)	0.353 (0.025)	0.346 (0.025)
Simulators	Hit	0.518 (0.039)	0.541 (0.035)	0.539 (0.04)	0.503 (0.038)	0.499 (0.035)	0.505 (0.04)	0.452 (0.046)	0.499 (0.038)	0.477 (0.033)	0.466 (0.037)	0.483 (0.039)	0.505 (0.042)
	Miss	0.404 (0.036)	0.444 (0.05)	0.44 (0.045)	0.412 (0.046)	0.414 (0.047)	0.397 (0.039)	0.347 (0.032)	0.309 (0.033)	0.296 (0.038)	0.308 (0.045)	0.293 (0.043)	0.333 (0.039)
	Absent	0.349 (0.026)	0.341 (0.028)	0.303 (0.019)	0.305 (0.022)	0.32 (0.015)	0.328 (0.017)	0.316 (0.022)	0.343 (0.025)	0.374 (0.03)	0.34 (0.025)	0.325 (0.021)	0.312 (0.024)

Note. Values in parentheses are Standard Error of the Mean. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset).

Table 4

Mean Proportion of Viewing Time Directed at Critical Objects Across Trial Types by Controls and Simulators in the Posttest Phase at the 250 ms Level (0-3000 ms)

		Timebin											
		0	250	500	750	1000	1250	1500	1750	2000	2250	2500	2750
Controls	Hit	0.464 (0.075)	0.377 (0.027)	0.524 (0.043)	0.688 (0.03)	0.674 (0.042)	0.608 (0.045)	0.567 (0.044)	0.598 (0.042)	0.624 (0.045)	0.59 (0.044)	0.612 (0.042)	0.645 (0.036)
	Absent	0.265 (0.065)	0.351 (0.036)	0.347 (0.026)	0.312 (0.028)	0.354 (0.033)	0.349 (0.019)	0.378 (0.041)	0.313 (0.032)	0.289 (0.023)	0.31 (0.024)	0.322 (0.027)	0.357 (0.033)
Simulators	Hit	0.248 (0.066)	0.335 (0.039)	0.476 (0.04)	0.592 (0.04)	0.592 (0.035)	0.523 (0.033)	0.549 (0.037)	0.567 (0.037)	0.571 (0.037)	0.586 (0.043)	0.6 (0.043)	0.607 (0.046)
	Absent	0.318 (0.056)	0.371 (0.02)	0.329 (0.021)	0.312 (0.023)	0.322 (0.027)	0.33 (0.024)	0.312 (0.03)	0.343 (0.022)	0.327 (0.024)	0.314 (0.023)	0.298 (0.024)	0.302 (0.034)

Note. Values in parentheses are Standard Error of the Mean. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset).

Table 5

Mean Proportion of Viewing Time Directed at Critical Objects Across Trial Types by Controls and Simulators in the Posttest Phase at the 250 ms Level (3000-6000 ms)

		Timebin											
		3000	3250	3500	3750	4000	4250	4500	4750	5000	5250	5500	5750
Controls	Hit	0.623 (0.032)	0.625 (0.036)	0.62 (0.033)	0.6 (0.037)	0.583 (0.037)	0.576 (0.036)	0.577 (0.036)	0.567 (0.037)	0.547 (0.037)	0.529 (0.039)	0.542 (0.038)	0.532 (0.04)
	Absent	0.347 (0.032)	0.351 (0.035)	0.357 (0.03)	0.361 (0.033)	0.323 (0.026)	0.333 (0.026)	0.333 (0.038)	0.339 (0.035)	0.323 (0.037)	0.329 (0.033)	0.317 (0.034)	0.285 (0.034)
Simulators	Hit	0.61 (0.039)	0.614 (0.048)	0.619 (0.045)	0.647 (0.041)	0.62 (0.039)	0.565 (0.039)	0.543 (0.043)	0.526 (0.047)	0.558 (0.037)	0.512 (0.042)	0.485 (0.042)	0.448 (0.04)
	Absent	0.355 (0.028)	0.379 (0.026)	0.344 (0.025)	0.33 (0.023)	0.324 (0.023)	0.324 (0.028)	0.308 (0.025)	0.325 (0.015)	0.296 (0.027)	0.305 (0.028)	0.302 (0.027)	0.335 (0.029)

Note. Values in parentheses are Standard Error of the Mean. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset).

Table 6

Proportion of Viewing Time Directed at Critical Objects During Miss Trials by Controls and Hit-Miss Difference Data for Both Groups at the Trial and 1000 ms Levels

		Timebin						
		0	1000	2000	3000	4000	5000	Trial
Controls	Miss Trials	0.389 (0.072)	0.516 (0.064)	0.547 (0.083)	0.735 (0.082)	0.611 (0.101)	0.713 (0.101)	0.601 (0.048)
Controls	Difference	0.1 (0.02)	0.109 (0.026)	0.077 (0.034)	-0.165 (0.037)	-0.103 (0.035)	-0.231 (0.038)	-0.042 (0.025)
Simulators	Difference	0.047 (0.031)	-0.008 (0.047)	0.03 (0.036)	0.103 (0.046)	0.121 (0.039)	0.173 (0.054)	0.085 (0.027)

Note. Values in parentheses are Standard Error of the Mean. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-1000 ms following display onset).

Table 7

Proportion of Viewing Time Directed at Critical Objects During Miss Trials by Controls and Hit-Miss Difference Data for Both Groups at the 250 ms Level (0-3000 ms)

		Timebin											
		0	250	500	750	1000	1250	1500	1750	2000	2250	2500	2750
Controls	Miss Trials	0.333 (0.167)	0.458 (0.122)	0.251 (0.085)	0.442 (0.103)	0.552 (0.103)	0.52 (0.1)	0.467 (0.105)	0.469 (0.109)	0.501 (0.108)	0.536 (0.108)	0.593 (0.109)	0.527 (0.11)
Controls	Difference	0.01 (0.068)	-0.092 (0.025)	0.215 (0.032)	0.147 (0.025)	0.074 (0.029)	0.112 (0.029)	0.148 (0.031)	0.13 (0.037)	0.128 (0.032)	0.117 (0.039)	0.026 (0.038)	0.07 (0.038)
Simulators	Difference	0.196 (0.162)	0.024 (0.056)	-0.017 (0.041)	0.06 (0.039)	0.039 (0.045)	-0.004 (0.051)	-0.035 (0.066)	-0.014 (0.06)	-0.009 (0.054)	0.037 (0.051)	0.045 (0.037)	0.081 (0.036)

Note. Values in parentheses are Standard Error of the Mean. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset).

Table 8

Proportion of Viewing Time Directed at Critical Objects During Miss Trials by Controls and Hit-Miss Difference Data for Both Groups at the 250 ms Level (3000-6000 ms)

		Timebin											
		3000	3250	3500	3750	4000	4250	4500	4750	5000	5250	5500	5750
Controls	Miss Trials	0.65 (0.104)	0.737 (0.104)	0.722 (0.109)	0.8 (0.092)	0.702 (0.098)	0.6 (0.112)	0.57 (0.111)	0.61 (0.11)	0.7 (0.105)	0.7 (0.105)	0.7 (0.105)	0.765 (0.106)
Controls	Difference	-0.068 (0.038)	-0.15 (0.035)	-0.165 (0.038)	-0.267 (0.045)	-0.183 (0.038)	-0.084 (0.038)	-0.079 (0.035)	-0.099 (0.037)	-0.205 (0.042)	-0.235 (0.033)	-0.234 (0.039)	-0.283 (0.046)
Simulators	Difference	0.114 (0.05)	0.096 (0.055)	0.099 (0.047)	0.091 (0.049)	0.085 (0.048)	0.108 (0.048)	0.105 (0.054)	0.19 (0.046)	0.181 (0.049)	0.157 (0.059)	0.19 (0.06)	0.172 (0.047)

Note. Values in parentheses are Standard Error of the Mean. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset).

Table 9

ROC Analysis of Proportion of Viewing Time Directed at Critical Objects During Hit Trials at the Trial and 1000 ms Levels

Timebin	AUC	SE	*p*	Cutoff	Se	Sp
0	0.682	0.077	<.05	.3237770	.167	1.000
				.3316638	.208	0.958
				.3723176	.292	0.875
				.3974044	.375	0.833
1000	0.79	0.07	<0.001	.3410504	.208	1.000
				.4059858	.458	0.958
				.4738923	.583	0.917
				.4819337	.625	0.875
				.4904446	.667	0.833
2000	0.748	0.072	<.01	.2049895	.083	1.000
				.3469355	.292	0.958
				.4369877	.500	0.875
				.4577337	.542	0.833
3000	0.561	0.085	0.47	--	--	--
4000	0.519	0.085	0.821	--	--	--
5000	0.477	0.086	0.789	--	--	--
Trial	0.66	0.079	0.058	.3064914	.125	1.000
				.3995003	.333	0.958
				.4505981	.458	0.875
				.4790418	.500	0.833

Note. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-1000 ms following display onset). AUC = Area Under the Curve, SE = Standard Error of the Mean, *p* = significance value, Cutoff = Test variable cutoff value resulting in given Se and Sp values, Se = Sensitivity, Sp = Specificity.

Table 10

ROC Analysis of Hit-Miss Difference in Proportion of Viewing Time Directed at Critical Objects at the Trial and 1000 ms Levels

Timebin	AUC	SE	*p*	Cutoff	Se	Sp
0	0.387	0.084	0.18	--	--	--
1000	0.332	0.081	<.05	.244489	.125	0.917
				.426436	.042	1.000
2000	0.425	0.085	0.375	--	--	--
3000	0.816	0.061	<.001	.092874	.542	0.958
				.194747	.375	1.000
4000	0.811	0.067	<.001	-.002003	.792	0.833
				.064732	.708	0.875
				.223531	.292	0.917
				.275184	.208	1.000
5000	0.901	0.044	<.001	-.05558	.833	0.833
				.081228	.625	0.958
				.147905	.542	1.000
Trial	0.767	0.073	<.01	.110708	.500	0.875
				.135777	.417	0.917
				.176551	.208	0.958
				.269441	.083	1.000

Note. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-1000 ms following display onset). AUC = Area Under the Curve, SE = Standard Error of the Mean, *p* = significance value, Cutoff = Test variable cutoff value resulting in given Se and Sp values, Se = Sensitivity, Sp = Specificity.

Table 11

ROC Analysis of Proportion of Viewing Time Directed at Critical Objects During Hit Trials at the 250 ms Level

Timebin	AUC	SE	*p*	Cutoff	Se	Sp
0	0.392 [a]	0.093	0.244	--	--	--
250	0.474	0.085	0.757	--	--	--
500	0.696	0.077	<.05	.3086700	.333	0.917
				.3390033	.417	0.833
750	0.736	0.074	<.01	.3344586	.167	1.000
				.4055339	.292	0.917
				.4388091	.458	0.875
1000	0.715 [a]	0.074	0.011	--	--	--
1250	0.797	0.068	<.001	.3957232	.417	0.958
				.4395923	.667	0.917
				.4904247	.708	0.875
1500	0.757	0.071	<.01	.3919770	.458	1.000
				.4514972	.583	0.833
1750	0.691 [a]	0.078	<.05	--	--	--
2000	0.77 [a]	0.069	<.01	--	--	--
2250	0.744 [a]	0.071	<.01	--	--	--
2500	0.738	0.072	<.01	.3583186	.292	0.958
				.3779201	.333	0.917
				.4027470	.375	0.875
2750	0.653	0.08	0.07	--	--	--
3000	0.607 [a]	0.082	0.205	--	--	--
3250	0.569	0.084	0.409	--	--	--
3500	0.513 [a]	0.085	0.877	--	--	--
3750	0.539 [a]	0.085	0.643	--	--	--
4000	0.517	0.085	0.837	--	--	--
4250	0.493 [a]	0.085	0.934	--	--	--
4500	0.532 [a]	0.087	0.703	--	--	--
4750	0.5	0.086	1	--	--	--
5000	0.51 [a]	0.085	0.902	--	--	--
5250	0.48 [a]	0.085	0.813	--	--	--
5500	0.465 [a]	0.085	0.68	--	--	--
5750	0.456 [a]	0.085	0.599	--	--	--

Note. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset). Superscript [a] denotes ROC analyses that had multiple ties between positive and negative state variable groups and may have yielded biased statistics. AUC = Area Under the Curve, SE = Standard Error of the Mean, *p* = significance value, Cutoff = Test variable cutoff value resulting in given Se and Sp values, Se = Sensitivity, Sp = Specificity.

Table 12

ROC Analysis of Hit-Miss Difference in Proportion of Viewing Time Directed at Critical Objects at the 250 ms Level

Timebin	AUC	SE	*p*	Cutoff	Se	Sp
0	0.591 [a]	0.103	0.346	--	--	--
250	.634	.088	.112	--	--	--
500	.172	.061	<.001	--	--	--
750	.326	.081	<.05	--	--	--
1000	.448	.086	.536	--	--	--
1250	.352	.082	.080	--	--	--
1500	.339	.087	.055	--	--	--
1750	.326	.080	<.05	--	--	--
2000	.351	.081	.076	--	--	--
2250	.396	.083	.216	--	--	--
2500	.526	.085	.757	--	--	--
2750	.517	.085	.837	--	--	--
3000	.729	.072	<.01	.119090	.458	0.833
				.164362	.375	0.917
				.200573	.333	0.958
				.318416	.208	1.000
3250	.783	.069	<.001	.046147	.625	0.833
				.119298	.458	0.958
				.197254	.375	1.000
3500	.811	.063	<.001	.062629	.583	0.833
				.100052	.542	0.917
				.191564	.500	1.000
3750	.859	.053	<.001	.003750	.625	0.875
				.029785	.583	0.958
				.122434	.500	1.000
4000	.806	.065	<.001	.017538	.667	0.833
				.119790	.500	0.917
				.182975	.333	0.958
				.215064	.292	1.000
4250	.766	.071	<.01	.021827	.750	0.833
				.093456	.542	0.875
				.316745	.208	1.000
4500	.727	.079	<.01	.070382	.667	0.875
				.237619	.292	0.917
				.330460	.250	1.000
4750	.852	.061	<.001	.041709	.833	0.833
				.074043	.792	0.875
				.119997	.667	0.917
				.309963	.292	1.000
5000	.901	.043	<.001	-.01792	.792	0.833
				.028538	.75	0.875
				.105858	.542	0.917
				.170664	.500	1.000
5250	.891	.050	<.001	-.11144	.833	0.833
				.027199	.750	0.958
				.102521	.708	1.000

5500	.891	.046	<.001	-.05682	.792	0.833
				.013572	.667	0.917
				.046426	.625	0.958
				.168664	.583	1.000
5750	.939	.032	<.001	-.02968	.833	0.875
				-.012184	.792	0.917
				.102297	.583	0.958
				.165275	.542	1.000

Note. Timebin labels represent the lower-bound value of that timebin (i.e., "0" represents 0-250 ms following display onset). Superscript [a] denotes ROC analyses that had multiple ties between positive and negative state variable groups and may have yielded biased statistics. AUC = Area Under the Curve, SE = Standard Error of the Mean, *p* = significance value, Cutoff = Test variable cutoff value resulting in given Se and Sp values, Se = Sensitivity, Sp = Specificity.

Table 13

Diagnostic Utility Statistics for Study Variables Used to Categorize Controls and Simulators in the Present Study

					Controls		Simulators						
Type	Variable	Cutoff	Se	Sp	Correct	Incorrect	Correct	Incorrect	FPR	FNR	PPV	NPV	Complete
Behavior	CR	.8020833	0.958	1.0	24	0	23	1	0	0.041667	1	0.96	--
Hits-1000	1000	.3410504	.208	1.000	24	0	5	19	0	0.791667	1	0.55814	N
Hits-1000	1000	.4059858	.458	0.958	23	1	11	13	0.041667	0.541667	0.916667	0.638889	N
Hits-1000	1000	.4738923	.583	0.917	22	2	14	10	0.083333	0.416667	0.875	0.6875	N
Hits-1000	1000	.4819337	.625	0.875	21	3	15	9	0.125	0.375	0.833333	0.7	Y
Hits-1000	1000	.4904446	.667	0.833	20	4	18	6	0.166667	0.25	0.818182	0.769231	Y
Diff-1000	5000	.081228	.625	0.958	23	1	15	9	0.041667	0.375	0.9375	0.71875	Y
Diff-1000	5000	.147905	.542	1.000	24	0	13	11	0	0.458333	1	0.685714	Y
Hits-250	1250	.3957232	.417	0.958	23	1	18	6	0.041667	0.25	0.947368	0.793103	N
Hits-250	1250	.4395923	.667	0.917	22	2	21	3	0.083333	0.125	0.913043	0.88	N
Hits-250	1250	.4904247	.708	0.875	21	3	22	2	0.125	0.083333	0.88	0.913043	N
Diff-250	5750	-0.02968	0.833	0.875	21	3	20	4	0.125	0.166667	0.869565	0.84	Y
Diff-250	5750	-0.01218	0.792	0.917	22	2	19	5	0.083333	0.208333	0.904762	0.814815	Y
Diff-250	5750	0.102297	0.583	0.958	23	1	14	10	0.041667	0.416667	0.933333	0.69697	Y
Diff-250	5750	0.165275	0.542	1	24	0	13	11	0	0.458333	1	0.685714	Y
Diff-250	5250	-0.11144	0.833	0.833	20	4	20	4	0.166667	0.166667	0.833333	0.833333	Y
Diff-250	5250	0.027199	0.75	0.958	23	1	18	6	0.041667	0.25	0.947368	0.793103	Y
Diff-250	5250	0.102521	0.708	1	24	0	17	7	0	0.291667	1	0.774194	Y
Diff-250	4750	0.041709	0.833	0.833	20	4	20	4	0.166667	0.166667	0.833333	0.833333	Y
Diff-250	4750	0.074043	0.792	0.875	21	3	19	5	0.125	0.208333	0.863636	0.807692	Y
Diff-250	4750	0.119997	0.667	0.917	16	8	22	2	0.333333	0.083333	0.733333	0.888889	N
Diff-250	4750	0.309963	0.292	1	24	0	7	17	0	0.708333	1	0.585366	N

Note. Numbers in Type column represent level of analysis. Numbers in Variable column represent specific timebin. CR = Corrected Recognition, Cutoff = Test variable cutoff value resulting in given Se and Sp values, Se = Sensitivity, Sp = Specificity, FPR = False Positive Rate, FNR = False Negative Rate, PPV = Positive Predictive Value, NPV = Negative Predictive Value, Complete = whether using that variable's cutoff correctly identified the final simulator case.

Appendix A:

Stimulus Lists

List	Stimulus	Category	List	Stimulus	Category	List	Stimulus	Category
A	anchor	Inanimate	B	barrel	Inanimate	C	barn	Inanimate
A	bus	Inanimate	B	lamp	Inanimate	C	basket	Inanimate
A	chair	Inanimate	B	stool	Inanimate	C	desk	Inanimate
A	fence	Inanimate	B	truck	Inanimate	C	house	Inanimate
A	moon	Inanimate	B	well	Inanimate	C	shoebox	Inanimate
A	vase	Inanimate	B	bow	Manipulable	C	coat	Manipulable
A	bowl	Manipulable	B	comb	Manipulable	C	cup	Manipulable
A	brush	Manipulable	B	crown	Manipulable	C	flag	Manipulable
A	glove	Manipulable	B	glass	Manipulable	C	iron	Manipulable
A	harp	Manipulable	B	guitar	Manipulable	C	kite	Manipulable
A	key	Manipulable	B	lock	Manipulable	C	nut	Manipulable
A	nail	Manipulable	B	pot	Manipulable	C	skirt	Manipulable
A	pants	Manipulable	B	ruler	Manipulable	C	violin	Manipulable
A	pen	Manipulable	B	sock	Manipulable	C	wrench	Manipulable
A	saw	Manipulable	B	top	Manipulable	C	bear	Organic
A	celery	Organic	B	arm	Organic	C	bread	Organic
A	cherry	Organic	B	banana	Organic	C	clown	Organic
A	eagle	Organic	B	donkey	Organic	C	finger	Organic
A	eye	Organic	B	ear	Organic	C	fly	Organic
A	lemon	Organic	B	frog	Organic	C	heart	Organic
A	pear	Organic	B	horse	Organic	C	monkey	Organic
A	pig	Organic	B	lion	Organic	C	mouse	Organic
A	sheep	Organic	B	pepper	Organic	C	snail	Organic
A	tomato	Organic	B	seal	Organic	C	swan	Organic

List	Stimulus	Category	List	Stimulus	Category	List	Stimulus	Category
D	church	Inanimate	E	cannon	Inanimate	F	car	Inanimate
D	clock	Inanimate	E	couch	Inanimate	F	piano	Inanimate
D	door	Inanimate	E	ladder	Inanimate	F	sled	Inanimate
D	table	Inanimate	E	star	Inanimate	F	wagon	Inanimate
D	train	Inanimate	E	stove	Inanimate	F	wheel	Inanimate
D	window	Inanimate	E	ball	Manipulable	F	chain	Manipulable
D	axe	Manipulable	E	blouse	Manipulable	F	fork	Manipulable
D	button	Manipulable	E	boot	Manipulable	F	hammer	Manipulable
D	candle	Manipulable	E	bottle	Manipulable	F	hat	Manipulable
D	doll	Manipulable	E	broom	Manipulable	F	jacket	Manipulable
D	dress	Manipulable	E	drum	Manipulable	F	kettle	Manipulable
D	gun	Manipulable	E	mitten	Manipulable	F	knife	Manipulable
D	pencil	Manipulable	E	needle	Manipulable	F	shirt	Manipulable
D	pliers	Manipulable	E	screw	Manipulable	F	vest	Manipulable
D	watch	Manipulable	E	shoe	Manipulable	F	ant	Organic
D	carrot	Organic	E	bird	Organic	F	apple	Organic
D	cat	Organic	E	cow	Organic	F	cake	Organic
D	leaf	Organic	E	deer	Organic	F	corn	Organic
D	nose	Organic	E	flower	Organic	F	dog	Organic
D	rabbit	Organic	E	goat	Organic	F	duck	Organic
D	snake	Organic	E	lips	Organic	F	fish	Organic
D	spider	Organic	E	toe	Organic	F	fox	Organic
D	thumb	Organic	E	turtle	Organic	F	grapes	Organic
D	tree	Organic	E	zebra	Organic	F	onion	Organic

Appendix B:

Counterbalancing of Stimulus Lists

	Target- Present Displays			Target-Absent Displays		
Participant	Studied	Distractor 1	Distractor 2	Foil 1	Foil 2	Foil 3
1 / 25	A	B	C	D	E	F
2 / 26	B	C	A	E	F	D
3 / 27	C	A	B	F	D	E
4 / 28	D	E	F	A	B	C
5 / 29	E	F	D	B	C	A
6 / 30	F	D	E	C	A	B
7 / 31	A	B	C	D	E	F
8 / 32	B	C	A	E	F	D
9 / 33	C	A	B	F	D	E
10 / 34	D	E	F	A	B	C
11 / 35	E	F	D	B	C	A
12 / 36	F	D	E	C	A	B
13 / 37	A	B	C	D	E	F
14 / 38	B	C	A	E	F	D
15 / 39	C	A	B	F	D	E
16 / 40	D	E	F	A	B	C
17 / 41	E	F	D	B	C	A
18 / 42	F	D	E	C	A	B
19 / 43	A	B	C	D	E	F
20 / 44	B	C	A	E	F	D
21 / 45	C	A	B	F	D	E
22 / 46	D	E	F	A	B	C
23 / 47	E	F	D	B	C	A
24 / 48	F	D	E	C	A	B

Appendix C:

Experimental Instructions

Control Instructions:

"For this experiment, you are going to be completing a memory task. It is extremely important that you try your best during this task."

Malingering Instructions:

"For this experiment I would like you to imagine that you were in a car accident in which another driver hit your car. You were knocked unconscious, and woke up in the hospital. You were kept overnight for observation. The doctors told you that you experienced a concussion.

Try to imagine that a year after the accident, you are involved in a lawsuit against the driver of the other car. If you are found to have experienced significant injuries as a result of the accident, you are likely to receive a bigger settlement. You have decided to pretend that you are suffering from a memory disorder as a result of the accident.

As a part of the lawsuit, you are required to take a test to determine whether or not you actually have a memory problem. You are going to complete this test in a moment. If you can successfully convince the examiner that you have a memory deficit, you are likely to get a better settlement. However, it is important that you perform in a way so that the examiner believes that you truly have a memory problem, but that it is not obvious that you are faking.

For example, some strategies that would be too obvious and would alert the examiner that you are faking, would be to answer every question incorrectly or to not answer some of the items. Once the other experimenter enters the room, you won't be able to ask any questions about these instructions, so, do you have any questions about what you are trying to accomplish during this experiment?"

Appendix D:

Post-Test Questionnaire

1) Recall, in your own words, what the instructed objective was for this experiment.

2) How much effort did you put in to accomplish this objective?

0 1 2 3 4 5

(No effort) (Great effort)

3) How motivated were you to accomplish this objective?

0 1 2 3 4 5

(Not motivated (Very motivated)

at all)

4) How confident are you that you accomplished this objective?

0 1 2 3 4 5

(Not confident (Very confident)

at all)

5) What strategies did you use to accomplish the objective? (check all that apply)

_____ answered most/all items incorrectly

_____ answered in a pattern (e.g. alternated between “yes” and “no”)

_____ answered randomly

_____ looked purposely away from the task-relevant materials (e.g., looked at some pictures but not others)

_____ blurred vision so could not see stimulus during study or test phase

_____ attempted to get a certain percentage correct (what percentage? _____)

_____ did not respond to some/all test items

_____ took longer than was necessary to respond to test items

_____ other (please describe):

_____

_____