



Published in final edited form as:

Neuropsychology. 2021 July ; 35(5): 472–485. doi:10.1037/neu0000747.

Detecting Simulated versus Bona Fide Traumatic Brain Injury Using Pupillometry

Sarah D. Patrick, Lisa J. Rapport, Robert J. Kanser, Robin A. Hanks, Jesse R. Bashem
Wayne State University

Abstract

Objective: Pupil dilation patterns are outside of conscious control and provide information regarding neuropsychological processes related to deception, cognitive effort, and familiarity. This study examined the incremental utility of pupillometry on the Test of Memory Malingering (TOMM) in classifying individuals with verified traumatic brain injury (TBI), individuals simulating TBI, and healthy comparisons.

Method: Participants were 177 adults across three groups: verified TBI ($n = 53$), feigned cognitive impairment due to TBI (SIM, $n = 52$), and healthy comparisons (HC, $n = 72$).

Results: Logistic regression and ROC curve analyses identified several pupil indices that discriminated the groups. Pupillometry discriminated best for the comparison of greatest clinical interest, verified TBI versus simulators, adding information beyond traditional accuracy scores. Simulators showed evidence of greater cognitive load than both groups instructed to perform at their best ability (HC and TBI). Additionally, the typically robust phenomenon of dilating to familiar stimuli was relatively diminished among TBI simulators compared to TBI and HC. This finding may reflect competing, interfering effects of cognitive effort that are frequently observed in pupillary reactivity during deception. However, the familiarity effect appeared on nearly half the trials for SIM participants. Among those trials evidencing the familiarity response, selection of the unfamiliar stimulus (i.e., dilation-response inconsistency) was associated with a sizeable increase in likelihood of being a simulator.

Conclusions: Taken together, these findings provide strong support for multimethod assessment: adding unique performance assessments such as biometrics to standard accuracy scores. Continued study of pupillometry will enhance identification of simulators who are not detected by traditional performance validity test scoring metrics.

Correspondence concerning this article should be addressed to Lisa J. Rapport, Department of Psychology, Wayne State University, Detroit, MI; rapport@wayne.edu.

Author Note

Sarah D. Patrick, Robert J. Kanser, Jesse R. Bashem, and Lisa J. Rapport, Department of Psychology, Wayne State University. Robin A. Hanks, Department of Physical Medicine and Rehabilitation, Wayne State University School of Medicine
Preliminary results of this study were presented at the 2020 Annual Meeting of the International Neuropsychological Society, Denver, CO., and the 2020 meeting of the American Psychological Association, Washington DC (virtual).
Robert J. Kanser is now at the James A. Haley Veteran's Hospital.
Jesse R. Bashem is now at Logical Behavioral Health.

Declaration of interest: The authors report that they have no conflicts of interest.

Keywords

Malingering; performance validity; pupillometry; traumatic brain injury

The validity of neuropsychological assessment is highly dependent on the quality of an examinee's engagement during the evaluation. Response styles account for as much as 50% of the outcome in test scores (Green, Rohling, Lees-Haley, & Allen, 2011; Meyers, Reinsch-Boothby, Miller, Rohling, & Axelrod, 2011). Unfortunately, purposeful attempts to feign brain injury are prevalent, particularly in the context of incentives for financial gain or mitigating responsibility (Flaro, Green, & Robertson, 2007; Rogers, 2008). Approximately 40% of compensation-seeking cases involve feigned cognitive impairment (Larrabee, 2003). Examinees must provide valid effort; otherwise, the assessment results may undermine the goals of accurate decisions regarding diagnosis, treatment planning, and resource allocation (Strauss, Sherman, & Spreen, 2006).

Psychologists cannot readily identify non-credible test performance based on personal interactions with the examinee (Boone, 2013). Performance validity tests (PVTs) used in traumatic brain injury (TBI) assessment generally capitalize on common misconceptions about TBI (e.g., that recognition memory and basic attention are typically impaired; Boone, 2013; Fuermaier et al., 2020; White et al., 2020). For example, many PVTs employ *floor effect* detection strategies, designed to be so easy that even individuals with significant cognitive impairments can perform well, given valid effort (Neudecker & Skeel, 2009). Failures on these tasks increase the likelihood that poor performance was purposeful.

Stand-alone PVTs such as the Test of Memory Malingering (TOMM; Tombaugh, 1996) are commonly structured as memory tasks in which examinees identify previously presented stimuli using a two-alternative forced-choice response format (Constantinou, Bauer, Ashendorf, Fisher, & McCaffrey, 2005). Unfortunately, stand-alone measures are readily identified and undermined with coaching (Brennan et al., 2009; Gunstad & Suhr, 2001; Kanser et al., 2017; Rose, Hall, Szalda-Petree, & Bach, 1998; Russeler, Brett, Klaue, Sailer, & Munte, 2008; Suhr & Gunstad, 2007; Tan, Slick, Strauss, & Hultsch, 2002). Examinees can easily obtain information about performance validity testing online, and knowledgeable attorneys can provide specific test-taking strategies to help clients identify PVTs and perform well enough to "pass" without detection (Bauer & McCaffrey, 2006). If examinees are aware of not only the presence of PVTs but also the specifics of how to perform in such a way to subvert them, these measures are rendered ineffective. There is a strong need for measures of malingering that cannot be consciously manipulated.

Biometrics yield psychophysiological markers that provide insight about effort and deception. Success in distinguishing true versus malingered memory test performance in TBI assessment has been demonstrated using biometrics such as reaction time (Kanser, Rapport, Bashem, & Hanks, 2019; Lupu, Elbaum, Wagner, & Braw, 2018; Patrick, Rapport, Kanser, Hanks, & Bashem, 2020; Rose et al., 1998) and oculomotor gaze patterns during PVTs (Kanser, Bashem, Patrick, Hanks, & Rapport, 2020; Tomer, Lupu, Golan, Wagner, & Braw, 2018). *Pupillary reactivity* during testing is another promising avenue for investigation in biometrics. Like pupillary responses to light and darkness, pupils dilate in response

to psychological and cognitive processes related to deception, cognitive effort, and the familiarity of stimuli (Beatty, 1982). Importantly, because pupil dilation is outside of conscious control (Loewenfeld & Lowenstein, 1999), pupillary patterns may provide insight about an individual's efforts regardless of their awareness of PVTs. If so, pupillometry could thwart coaching and other strategies to remain undetected among people who feign low functioning.

Biometrics have been used extensively in experimental and applied settings examining decision-making and detecting deception. A large body of research in the field of cognitive psychology has demonstrated that pupil dilation increases as a function of *cognitive load* and effort (Beatty, 1982; Kahneman & Beatty, 1966; Porter, Troscianko, & Gilchrist, 2007). In the context of PVT literature, the term "effort" is now considered dated because it evoked controversy regarding the interpretation of response styles. As Larrabee has noted, in the clinical setting, the terms "valid" and "invalid" most accurately capture the extent to which test performance reflects actual ability (Larrabee, 2012). However, in the literature on cognitive load, the term "effort" is used to reflect the *physiologic engagement and response* to task demands (Murphy, O'Connell, O'Sullivan, Robertson, & Balsters, 2014). Related to this robust phenomenon involving cognitive load and "effort" is the observation that pupils dilate during deceptive behavior such as lying (Dionisio, 2001). Research documenting the success of pupillometry in detecting deception is explained by the idea that deception requires substantially more mental effort than truth telling (Hu et al., 2015). These established findings in pupillometry could be applied to the context of PVTs: feigned impairment would be associated with greater load-related effort, and hence larger pupil dilations, than responding without the added load of purposeful deception. Like distinctions regarding the term "effort," evaluative terms such as "deceptive" responding are inappropriate applied in a clinical setting, because conscious and/or purposeful cognitive processes of examinees can rarely be known. In contrast, experimental designs can be established in which participants are explicitly assigned to conditions to perform deceptively or at their best. A well-established field of research focused on decision-making uses this paradigm to examine detection of deception (Heilbronner et al., 2009).

Pupil dilation also provides information about the familiarity of objects in recognition memory tasks. Vö and colleagues (2008) coined the *pupil old/new effect* to describe the automatic, unconscious dilation that occurs when individuals view (old) stimuli they have seen before as compared to novel (new) items. In the context of PVTs, examinees are typically asked to choose which of two stimuli has been previously viewed. The findings regarding familiarity suggest that individuals intentionally answering items incorrectly to feign brain injury would show a different pattern of pupil dilation than those who genuinely answer items incorrectly (Otero, Weekes, & Hutton, 2011). Furthermore, individuals engaging sincerely in a forced-choice memory task should respond behaviorally (select responses) in ways that align with their pupil dilation patterns. Regardless of response accuracy, individuals employing valid effort should select responses to which their pupils dilated most (the stimulus they recognized as familiar). Dilation-response inconsistency may indicate that an individual is actively avoiding selecting the stimuli they recognize as familiar. This prediction was partly tested by Heaver and Hutton (2010), who replicated the pupil old/new effect in an experimental task testing healthy undergraduates instructed to

perform validly, report all items as “new”, or feign amnesia. Across all three groups, pupils dilated more to familiar than unfamiliar stimuli.

The purpose of this study was to use a novel application of pupillometry to enhance diagnostic accuracy in the identification of bona fide TBI versus feigned cognitive impairment. Although pupillary reactivity has been well studied in experimental research on deception, its utility in the clinically-applied context of feigned cognitive impairment remains to be explored. The inclusion of a clinical group is especially important to the development of new indices designed to detect malingering. Laboratory designs that contrast only healthy adults instructed to be deceptive (TBI simulators) and healthy adults instructed to perform their best often yield results with much larger effects and different patterns than are observed for designs that include the clinical group of interest (Kanser et al., 2020; Kanser et al., 2019; Patrick et al., 2020).

Designs that lack a bona fide clinical group have low ecological validity and increase risk of false-positive classifications of invalid performance because the tests are not robust to effects of cognitive impairment (Bodner, Merten, & Benke, 2019; Patrick et al., 2020). This issue is especially important when examining pupillometry indices, because TBI is commonly associated with chronic oculomotor sequelae (Armstrong, 2018). Accordingly, this study compared patterns of pupil dilation during a computerized administration of the TOMM (TOMM-C) among adults with bona fide TBI and groups of healthy adults who were instructed either to perform at their best ability or to perform deceptively and simulate TBI. The main aim was to examine the extent to which pupillary characteristics provide incremental utility to the diagnostic accuracy of the TOMM-C. The central hypothesis was that persons with actual TBI and persons who feign TBI would exhibit distinct oculomotor patterns during cognitive evaluation, and that analysis of these patterns would enhance TOMM-C classification accuracy. Specifically, we expected that individuals engaging in purposeful feigned/deceptive responding would show physiologic indications of greater mental effort as indicated by pupillary dilation compared to healthy and brain-injured adults instructed to perform their best. Moreover, nuanced indices that combine tracking of automatic dilation to familiar stimuli with decision-making would show pathognomonic dilation-response inconsistency among individuals known to be responding deceptively.

Method

Participants

Participants were 177 adults across three groups. The traumatic brain injury group (TBI) consisted of 53 adults with bona fide moderate to severe TBI recruited from the Southeastern Michigan TBI Model System (SEMTBIS). SEMTBIS inclusion criteria included complicated-mild to severe TBI as indicated by post-traumatic amnesia of at least 24 hours, loss of consciousness of at least 30 minutes, and a Glasgow Coma Scale score of less than 13 at emergency department admission, or abnormal neuroimaging. All TBI participants had sustained injuries severe enough to warrant inpatient rehabilitation treatment, were > 16 years of age at the time of injury, used English as their primary language, and were at least 1 year post injury.

Neurologically healthy adults were recruited from the metropolitan Detroit area via advertisements and flyers. The healthy comparison group (HC) included 72 adults instructed to give full effort during testing. A TBI simulator group (SIM) consisted of 52 healthy adults instructed to simulate brain injury during testing. The participants used English as their primary language and had no history of significant neurological conditions (e.g., seizures, TBI, stroke), serious psychiatric illness (e.g., psychotic disorder, major depression), or current substance use disorder. Although some participants wore corrective lenses, all participants had sufficient visual acuity required to complete calibration procedures for eye tracking, as well as reading small print and identifying figures included in a neuropsychological test battery.

Participants ranged in age from 18 to 78 years. The SIM group ($M = 34.8$, $SD = 16.4$) was significantly younger than both the HC ($M = 46.7$, $SD = 16.5$) and TBI groups ($M = 46.3$, $SD = 12.4$), and the HC and TBI groups did not differ from each other, $F(2, 177) = 10.84$, $p < .001$. Highest level of education achieved ranged from 7 to 20 years and differed between groups, $F(2, 177) = 19.69$, $p < .001$. HC ($M = 13.9$, $SD = 2.3$) and SIM ($M = 14.8$, $SD = 2.3$) had higher levels of education than TBI ($M = 12.1$, $SD = 2.2$). The present sample excludes 2 SIM group participants who explicitly noted that they did not follow the study instructions (i.e., forgot to simulate TBI during testing) during a post-test debriefing and manipulation check.

Measures

Participants completed a computerized version of the *Test of Memory Malingering* (TOMM-C; Tombaugh, 1996) in the context of a comprehensive neuropsychological battery. Responses were provided using a Cedrus RB-530 response box with two buttons matching the orientation of test stimuli. The TOMM-C was administered with standard instructions aside from references to the response box. The TOMM is among the most widely-used and well-researched PVTs (Sharland & Gfeller, 2007; Slick, Tan, Strauss, & Hultsch, 2004). It is a 50-item, forced-choice, visual memory test that consists of two learning trials and two recognition trials.

The Tobii TX-300 Eye Tracking System (TX-300) was used for pupillometry. The TX-300 uses dark-pupil monitoring with multiple infrared cameras at a sampling rate of 300 Hz. Informed consent procedures stated explicitly that eye movement would be recorded, and participants completed an eye-tracking calibration protocol prior to the start of the experimental portion. Pupil and gaze data were processed by Tobii Studio software, and posttest data reduction using Structured Query Language (SQL). The apparatus was arranged according to guidelines in the Tobii technical manual. Participants were seated approximately 65 cm from the TX-300 monitor. Participants were oriented to the stationary eye-tracking apparatus and instructed to maintain their position facing the equipment to ensure accurate capture of eye-tracking information. A separate computer was linked to the TX-300 via E-Prime Extensions for Tobii (EET) to present the task. In addition to recording accuracy, response time, and oculomotor behaviors of interest, EET provided validity estimates for each 3 ms sampling of oculomotor behaviors for each eye.

Pupillary indices were calculated separately for Trial 1 and Trial 2. Maximum pupil dilation for each item was determined by the three contiguous data points (i.e., 9 ms) with the largest average dilation. Similarly, pupillary baseline was calculated using the three contiguous data points with the smallest average dilation. Additional indices included the average maximum and baseline dilations when: viewing all items, viewing correct stimuli, viewing incorrect stimuli, answering correctly, and answering incorrectly. Variability in effort was assessed by the difference between average maximum and baseline dilation. Familiarity was assessed via the difference between the average maximum dilations while viewing correct versus incorrect stimuli. Difference scores were calculated by subtracting the average minimum dilation from the average maximum dilation, and the average maximum dilation while viewing incorrect versus correct stimuli. Table 1 presents the operational definitions for the pupillary variables and links them to the constructs they reflect.

Procedure

Following confirmation of eligibility and completion of informed consent, testing was completed in a single session lasting approximately 2 hours. Prior to completing the assessment battery, participants in the TBI Simulator Group (SIM) were presented with a scenario in which they are described to have experienced a minor car accident. This scenario has been used successfully in TBI simulation studies with similar research designs (Coleman, Rapport, Millis, Ricker, & Farchione, 1998; Kanser et al., 2020; Patrick et al., 2020; Rapport, Farchione, Coleman, & Axelrod, 1998; Tombaugh & Tombaugh, 1997). Participants in the TBI Simulator group were read a list of symptoms that commonly occur following TBI, such as slowed thinking, memory dysfunction, and other behaviors relevant to cognitive testing. They were warned about the presence of PVTs within the battery and informed of additional financial incentives (\$30 bonus and a raffle entry for \$200) for successfully remaining undetected by PVTs. Lastly, participants in this group were encouraged to use resources such as the internet to prepare for the evaluation, which was scheduled 1 – 2 weeks from the initial screening. This information was also emailed to the participants for their review during the preparation period.

To obtain an estimate of general intelligence, all participants completed the Wechsler Test of Adult Reading (WTAR; The Psychological Corporation, 2001) under standard conditions, with instruction to perform to the best of their abilities. Following completion of the test battery, participants in the SIM group were administered a questionnaire regarding their preparation for the session as well as their strategies for simulating TBI.

Statistical Analysis

The data were screened prior to analysis per recommendations by Tabachnick, Fidell, and Ullman (2018), including assumptions of relevant statistical models. Participants missing valid pupillary data for greater than 50% of items during a TOMM-C trial were excluded for that trial ($n = 14$ on either Trial 1 or Trial 2). Descriptive statistics and univariate tests, nonparametric and parametric as appropriate, were used to compare groups. Logistic regression and receiver operating characteristics (ROC) curve analyses tested the predictive value of pupillary indices individually and incremental to accuracy scoring. Area under the curve (AUC) values above .70 are considered “acceptable,” values above .80 are considered

“excellent,” and AUC greater than .90 is “outstanding” (Hosmer, Lemeshow, & Sturdivant, 2013). Sensitivities (Sn) were calculated with specificity (Sp) set at .90 to reflect modern standards for clinical application. Youden’s J was calculated as a summary of overall diagnostic efficiency using the sensitivity and specificity from the logistic regressions ($Sn + Sp - 1$; Youden, 1950). Classification statistics positive predictive power (PPP) and negative predictive power (NPP) were calculated at two theoretical base rates of interest, 10% and 40%, for the statistically derived sensitivity and specificity. The 40% base rate represents the estimated rate of feigned impairment in cases involving compensation or litigation (Larrabee, Millis, & Meyers, 2009). The 10% base rate is more challenging psychometrically and represents an estimate of what might be encountered in a general, non-forensic clinical setting.

Results

Table 2 provides descriptive statistics and univariate comparisons of TOMM-C indices. On TOMM-C Accuracy (total number correct), 14.9% of the total sample received a perfect score on Trial 1, and 53.2% of the sample received a perfect score on Trial 2. Kruskal-Wallis test indicated that TOMM-C Accuracy differed significantly among the groups, with large effect sizes for both trials (Trial 1 $\eta^2 = .38$; Trial 2 $\eta^2 = .43$). Post hoc Mann-Whitney tests ($p < .05$ criterion) indicated that across both trials, TOMM-C Accuracy for SIM was significantly lower than for TBI and HC, and TOMM-C Accuracy for TBI was significantly lower than HC. Kruskal-Wallis tests also indicated that all of the pupillary indices differed significantly across groups, and nearly all of the pupillary indices showed medium ($\eta^2 = .06$) to large ($\eta^2 = .14$) effect sizes (Cohen, 1988). The difference-score Effort-Range showed a large effect across both trials ($\eta^2 = .16$), whereas Familiarity-Difference did not (Trial 1 $\eta^2 = .06$, Trial 2 $\eta^2 = .03$).

Of the pupillary indices, Response-Unfamiliar (Max) had the largest effect size for Trials 1 and 2 ($\eta^2 = .20 - .23$). In Trial 1, SIM showed larger dilation than HC and TBI, and HC showed larger dilation than TBI. In Trial 2, SIM showed larger dilation than both HC and TBI; however, HC and TBI did not significantly differ from each other. Post hoc tests indicated that all pupillary indices differed significantly between TBI and SIM except Effort-Range. SIM showed significantly larger pupil dilation compared to TBI for all indices except difference scores. Of note, for all indices involving maximum dilation, SIM showed significantly greater dilation than HC. However, SIM and HC were equivalent on nearly all indices involving minimum dilation, except View-Familiar (Baseline) for Trial 1. Both SIM and HC showed significantly larger dilation than TBI for all indices calculated using maximum and minimum dilation.

Descriptive Correlations

Correlations examined the extent to which demographic characteristics might affect pupillary reactions associated with cognitive performance (e.g., age, cognitive ability; see Supplemental Table S1). Age showed inverse correlation with all maximum- and minimum-dilation indices, with medium effects ranging from $-.34$ to $-.46$. Education showed positive correlation to all maximum- and minimum-dilation indices on both trials, except Response-

Unfamiliar (Max) on Trial 2; however, the effects were generally small (Trial 1 ρ .20 to .25; Trial 2 ρ s < .20). Across both trials, WTAR was positively related to all maximum and minimum pupillary indices, ranging from .28 to .37 for Trial 1, and .23 to .30 for Trial 2.

For Trial 1, the standard Accuracy score was generally unrelated to pupil indices reflecting minimum (baseline) or maximum dilation (ρ .00 to $-.12$); however, it showed inverse relation to range of cognitive effort (Effort-Range $\rho = -.42$) and positive relation to viewing familiar versus unfamiliar items (Familiarity-Difference $\rho = .24$). Trial 2 showed similar pattern for the difference scores, albeit weaker (ρ .20 to $-.38$); inverse correlations were also observed for each of the indices tapping maximum dilation ($\rho = -.22$ to $-.38$), whereas minimum-dilation indices were unrelated to accuracy ($\rho = .07$ to $-.15$). Also noteworthy is that the set of indices reflecting maximum dilation were essentially redundant, with intercorrelations ρ .98 (Trial 1) and .94 (Trial 2). A similar pattern was observed for the indices reflecting minimum (baseline) dilation showing intercorrelations ρ .97 (Trial 1) and .93 (Trial 2). In contrast, the difference-score variables (Effort-Range, Familiarity-Difference) showed only small to medium relation to the maximum- and minimum-dilation variables from which they were calculated (ρ $-.02$ to .45).

Classification Accuracy SIM versus TBI

Tables 3a (Trial 1) and 3b (Trial 2) present classification accuracy statistics, logistic regressions, and ROC curve analyses for predictions of *SIM versus TBI* group membership by pupillary indices individually and combined with accuracy in two-variable models. For parsimony, only significant single-variable models were tested as two-variable models. Individually, pupillary indices showed modest classification accuracy, albeit generally stronger than typical PVTs, which average approximately .56 sensitivity (Vickery, Berry, Inman, Harris, & Orey, 2001). Average sensitivities of the pupillary indices were .67 and .70 for Trials 1 and 2, respectively. Sensitivities at 90% specificity (Sn_{Sp90}) ranged from low for Familiarity-Difference Trial 2 (.12) to modest for Response-Unfamiliar (Max) Trial 2 (.65). When the 90% specificity cutpoint is associated with multiple sensitivity values, a range of values is presented in the tables. Of note, positive predictive power (PPP) for all of the pupillary indices exceeds the base rates.

Discriminability was “excellent” (AUC > .80) for Trial 2 accuracy (.85) and “acceptable” (AUC > .70) for Trial 1 accuracy (.78). Several indices showed excellent discriminability in Trial 1 (.80 to .83) as well as View-Unfamiliar (Max) in Trial 2 (.80). TOMM-C Trial 2 Accuracy outperformed pupillary indices and Trial 1 Accuracy, with Sn_{Sp90} of .65, Youden’s J .57, and AUC .85. For Trial 1, Response-Unfamiliar (Max) and View-Familiar (Max) performed best, both in the excellent range. For Trial 2, Response-Familiar (Max) and Effort-Max performed best (acceptable). To determine the extent to which pupillary indices improved discrimination over TOMM-C Accuracy, two-variable models combined Accuracy (Block 1) with one pupillary index that significantly predicted group membership (Block 2). In general, the two-variable models performed strongly: all pupillary indices provided incremental predictive value over TOMM-C Accuracy (Trial 1 all $p < .001$; Trial 2 all $p < .05$), with the exception of Effort-Range for both trials. Several two-variable models showed outstanding discriminability, including Effort-Baseline and Response-Familiar (Baseline)

for Trial 1 (.90), Response-Unfamiliar (Max) and View-Familiar (Baseline) for Trial 2 (.90), and Response-Unfamiliar (Baseline) for both Trials 1 and 2 (.90–.91). All other two-variable models showed excellent discriminability (.88–.89), except Trial 1 Effort-Range (.79, acceptable).

Classification Accuracy HC versus SIM

Tables 4a (Trial 1) and 4b (Trial 2) present the analyses for SIM versus HC. Numerous pupillary indices were significant predictors of group membership ($p < .05$). Average sensitivities were lower than those observed for SIM-TBI contrasts: .38 Trial 1 and .37 Trial 2 (range .10 to .95). Sn_{Sp90} ranged from .12 for Response-Unfamiliar (Baseline) Trial 1 to .55 for Effort-Range Trial 2. Nonetheless, PPP exceeded the base rates for the majority of the indices.

All two-variable models were significant ($p < .001$). For Trial 1, all pupillary indices added significant predictive value to TOMM-C Accuracy scores ($p < .05$) except Familiarity-Difference; however, no Trial 2 pupil indices added significant predictive value. For both trials, TOMM-C Accuracy performed best, with Trial 1 showing outstanding discriminability and Trial 2 excellent discriminability. Of the pupillary indices, Effort-Range performed best on both trials (acceptable). Individually, Trial 2 Response-Unfamiliar (Max) and Effort-Range showed acceptable discriminability on both trials, whereas all other indices fell below the acceptable range. All two-variable models for both trials were in the outstanding range for discriminability except for Familiarity-Difference for Trial 2 (.89, excellent).

Classification HC versus TBI

Tables 5a (Trial 1) and 5b (Trial 2) present the analyses for HC and TBI. Again, several individual indices were significant predictors, and all two-variable models were significant for Trial 1 ($p < .001$) and Trial 2 ($p < .01$). For both trials, all pupillary indices added significant predictive value to TOMM-C Accuracy ($p < .05$) except Trial 1 Effort-Range. Average sensitivities were .50 and .40 for Trials 1 and 2, respectively. Sn_{Sp90} was generally low (.08–.26, Trial 1; .08–.30, Trial 2). With few exceptions, PPP exceeded the base rates.

For ROC analyses, TOMM-C Accuracy was acceptable (.71) for Trial 1 and below acceptable for Trial 2 (.59). Overall, several individual pupil indices performed in the acceptable range, but none were excellent. For Trial 1, Response-Unfamiliar (Baseline) and Response-Familiar (Baseline) performed best. For Trial 2, Response-Familiar (Baseline) and Effort-Range performed best. For two-variable models, in Trial 1, Response-Unfamiliar (Baseline) and Response-Unfamiliar (Max) yielded excellent classification (.80), with all other indices acceptable (.73–.79). For Trial 2, two-variable models with Effort-Baseline, View-Unfamiliar (Baseline), Response-Familiar (Baseline), View-Familiar (Baseline), View-Unfamiliar (Max), Effort-Range were acceptable (.70–.73), with all others falling below the acceptable range.

Old/New Effect and Dilation-Response Inconsistency

To confirm the presence of the old/new effect, in which participants' pupils dilate more to old, previously-seen stimulus items (i.e., the correct answer) than to new, unfamiliar stimulus items (i.e., the incorrect answer), Response-Familiar (Max) and Response-Unfamiliar (Max) were compared. Paired-samples *t* tests revealed that participants' pupils dilated more when viewing correct (old) response options than incorrect (new) response options for both Trial 1, $t(173) = 8.31, p < .001$, Cohen's $d = 1.3$ (very large), and Trial 2, $t(171) = 3.44, p = .001$, Cohen's $d = 0.5$ (medium). This pattern was consistent within groups, except for the SIM group during Trial 2, for which pupil dilation did not significantly differ for correct (old) and incorrect (new) response options, $t(50) = 0.37, p = .713$, Cohen's $d = 0.10$.

To investigate whether the SIM group was more likely to choose stimuli to which they did *not* dilate most (i.e., stimuli that were not familiar per the old/new effect), groups were compared on the number of instances in which the eye dilated to the correct stimulus, yet the participant chose the incorrect stimulus (dilation-response inconsistency; DRI). In Trial 1, 50.0% of HC, 75.5% of TBI, and 94.2% of SIM answered incorrectly despite showing greater dilation to the correct stimulus. In Trial 2, 8.3% of HC, 16.0% of TBI, and 74.5% of SIM examinees exhibited DRI. Kruskal-Wallis tests revealed that groups differed across both trials of the TOMM-C, Trial 1 $H = 55.22, p < .001, d = 0.9$ (large); Trial 2 $H = 76.58, p = .001, d = 0.5$ (medium). Mann-Whitney tests showed that SIM exhibited DRI more often than both TBI (Trial 1 $U = 751.50, p < .001$; Trial 2 $U = 447.50, p < .001$) and HC, Trial 1 $U = 478.00, p < .001$; Trial 2 $U = 543.00, p < .001$. TBI showed DRI more often than HC in Trial 1 ($U = 1154.00, p < .001$) but did not differ from HC in Trial 2, $U = 1657.00, p = .178$. For Trial 1, it was rare for HC to exhibit DRI even once ($M = 0.94, SD = 1.19$) compared to TBI ($M = 2.23, SD = 2.07$) and SIM ($M = 5.17, SD = 3.97$). For Trial 2, it was rare for either HC ($M = 0.11, SD = 0.40$) or TBI ($M = .32, SD = 0.98$) to exhibit DRI compared to SIM ($M = 3.71, SD = 3.83$).

Discussion

The findings support the value of biometric indicators of cognitive effort in the context of performance validity assessment. Several pupillary indices discriminated feigned impairment from healthy adults and adults with verified TBI who were instructed to perform their best. Moreover, several indices improved the diagnostic accuracy of the traditional TOMM accuracy score in identifying feigned cognitive impairment. In the context of pupil dilation as an indicator of cognitive effort, TBI simulators appeared to work harder (i.e., experience more cognitive load) than adults who did not experience the added burden of responding deceptively, regardless of TBI status. Additionally, because pupillary response divulges familiarity, test behaviors inconsistent with that sign (i.e., choosing an incorrect response having shown pupillary recognition of the correct one) informed identification of persons feigning TBI.

As expected, healthy adults feigning TBI showed greater pupil dilation than both healthy comparisons putting forth best effort and individuals with bona fide TBI, indicating that they experienced greater cognitive load during the task. Indices reflecting both chronic

effort and range of effort indicated that feigners experienced the greatest cognitive load, followed by adults with TBI, who experienced greater load than healthy adults who were not responding deceptively. This pattern was consistent across both trials of the TOMM. Engaging in deception is more difficult than truth telling (Hu et al., 2015), and pupil dilation is a robust indicator of cognitive load (Kahneman & Beatty, 1966; Kahneman & Peavler, 1969; Porter et al., 2007). Simulators must track their number of incorrect answers so their deception is not conspicuous, and weigh the options of answering correctly or deliberately answering incorrectly on every trial. Individuals who are trying to perform at their best, regardless of TBI status, must merely pay attention to the correct answers, using substantially fewer cognitive resources than are required by deception. In contrast, *baseline* dilation for healthy adults – whether feigning TBI or not – was equivalent, and both groups of healthy adults showed greater baseline dilation than adults with TBI. This finding may reflect that general arousal or engagement in the task is diminished for adults with TBI. Differences in baseline dilation likely reflect attenuated responsiveness of brain areas associated with mediation of attention, cognitive load, and arousal. For example, pupil size reflects locus coeruleus-norepinephrine arousal and activity during attentional demands (Oliva, 2019), a brain area commonly susceptible to damage in TBI (Valko et al., 2016).

Several pupillary indices added unique information beyond the discriminative ability accounted for by the traditional TOMM accuracy score and were particularly helpful for comparisons of *simulators versus TBI* – the comparison of most ecological importance. This is a key finding because it highlights the utility of multimethod assessment: Even modest predictive power in an individual index becomes especially valuable if it provides unique information. Most test batteries load up on monomethod tasks, each of which assess performance validity via behavioral responses that can be consciously manipulated (i.e., accuracy). Although multiple measures provide opportunity for convergent validity and enhance reliable assessment of a single construct, measurement error associated with the method is carried throughout the set of tasks, and the value of redundancy reaches an asymptote. Notably, in this study, multiple pupil indices were stronger indicators of feigned TBI than traditional accuracy during Trial 1 (average maximum dilation while answering incorrectly performed best during both trials of the task). Individually, most indices showed “acceptable” to “excellent” discriminability. Using an important clinical criterion – sensitivity of the indices at specificity set to 90% or better – the pupillary indices were generally stronger for Trial 1 than Trial 2, whereas Trial 1 Accuracy was less sensitive and had poorer discriminability than Trial 2 Accuracy. Overall, sensitivities at 90% specificity were modest, but multiple indices exceeded the average traditional PVT (56%; Vickery et al., 2001). Furthermore, positive predictive powers exceeded base rate predictions at 40% and even 10%. When combined with the traditional accuracy scores, nearly all the pupil indices showed excellent or outstanding discriminability in detecting feigned versus actual TBI.

The finding that pupil dilation distinguished simulators from bona fide TBI *more effectively* than simulators versus healthy adults who were not responding deceptively is an important distinction from prior PVT research. Notably, fewer indices were viable predictors of group membership for the two groups of healthy adults. The pattern generally indicated that healthy adults who were not responding deceptively experienced a wider range of cognitive

effort during the task than adults feigning TBI, who maintained a higher cognitive load consistently throughout the task. This pattern was evidenced in multiple ways of assessing cognitive load (e.g., average baseline dilation, difference between maximum and minimum dilation, etc.). Sensitivities at specificity 90% were generally low for individual pupillary indices in distinguishing simulators and healthy comparisons. Nonetheless, when combined with accuracy, several indices added unique predictive value to traditional accuracy, showing excellent to outstanding discriminability. Additionally, sensitivity at specificity 90% for combined models were strong across both trials.

For comparisons of *nondeceptive healthy responders and TBI*, all pupillary indices during Trial 1 distinguished the groups except the difference for viewing correct versus incorrect stimuli. For Trial 2, in addition to difference viewing correct versus incorrect, the maximum and minimum dilations for incorrectly answered trials were also similar for the two groups. These findings also support the hypotheses that pupil dilation reflects cognitive load. Additionally, they highlight that using absolute pupil size as an index of deceptive responding would pose challenges, because TBI differed from healthy comparisons and simulators. In this regard, nuanced indices that reflect relative process (e.g., difference scores) have more promise because they indicate pupil dilation relative to baseline dilation. Patterns of dilation-response inconsistency may be the most promising avenue because it is independent of absolute pupil size.

Consistent with Vö et al. (2008), the present study provided evidence for the old/new effect in terms of pupil dilation as an index of familiarity. Generally, participants' pupils dilated more to old (correct) response options than new (incorrect) response options, signaling recognition of familiar (correct) response choices. Importantly, this pattern held for adults with TBI, indicating that the presence of TBI did not alter the expected, automatic dilation for familiar versus novel stimuli. In contrast, the typically robust phenomenon of the old/new effect was relatively diminished among TBI simulators. This finding may reflect competing, interfering effects of cognitive effort that are frequently observed in pupillary reactivity during deception (Hu et al., 2015). However, the old/new effect appeared on nearly half the trials for TBI simulators. Among those trials in which the familiarity response was present, selection of the unfamiliar stimulus (i.e., dilation-response inconsistency) was associated with a sizeable increase in likelihood of being a simulator. Across both trials, adults feigning TBI more often selected responses contrary to their recognition of familiarity based on pupil dilation patterns, whereas adults with TBI and healthy comparisons rarely selected against their familiarity even once.

The inclusion of a group with bona fide TBI proved to be a critical aspect of these findings. Typically, analogue studies involving only healthy adults (simulators and full-effort comparisons) greatly overestimate the utility of performance validity assessment measures (Kanser et al., 2019). Of note, however, if this study had included solely healthy adults (no TBI group), the findings may have *underestimated* the usefulness of pupillometry. The findings comparing adults feigning TBI versus adults with actual TBI were *stronger* than for comparisons of feigners versus healthy adults who were not instructed to feign. The finding that adults with TBI showed the smallest pupil dilations for both maximum and minimum pupillary indices indicate that these indices are sensitive to the presence of TBI. Oculomotor

impairments are common after TBI, and pupil reactivity is a longstanding indicator of brain integrity (Armstrong, 2018; Ciuffreda, Joshi, & Truong, 2017). Thus, in addition to identifying distinctive patterns indicating greater cognitive effort among adults feigning TBI, pupillometry can distinguish characteristic dysfunctions associated with neural injuries sustained in TBI. The combined effects from comparisons among these groups explain why predicting group membership in this scenario was more powerful as compared to designs including only healthy adults in a classic analogue simulation paradigm.

For some people, biometrics such as reaction times, visual tracking, and pupillometry can seem to be a more intrusive form of assessment as compared to traditional forms of cognitive assessment (Farah, Hutchinson, Phelps, & Wagner, 2014). This might be especially true for pupillometry, which monitors and records behaviors that are not consciously controlled. The fact that eye tracking is historically linked to lie detection research also can evoke reflexive emotional reactions when considered in the context of everyday clinical neuropsychology evaluations. Related issues raised in legal and forensic circles concern fourth amendment rights to privacy, which protect against searches without a warrant, including physical tests such as neuroimaging (Farah et al., 2014). Additionally, there must be important concern regarding whether this technology would ever achieve the kind of precision required to meet Daubert “known error rate” standards (Daubert v. Merrell Dow Pharmaceuticals Inc., 1993) for forensic applications. The subjective aspects of these issues are important to consider in the context of public trust. In that regard, because it seems especially sensitive to real brain injury (i.e., identifying oculomotor phenomena that are pathognomonic physiologic indicators of brain damage), it seems relevant to highlight the promising effectiveness of pupillometry as an adjunctive tool to reduce false-positive diagnoses of malingering. Like all tools, the information garnered from biometric indices must be used responsibly and ethically, in service to the best welfare of the client. Carefully designed, rigorous clinical trials will be the ultimate manner to address whether these kinds of biometric technologies can be responsibly generalized from laboratory to clinical application (Langleben & Moriarty, 2013).

Limitations

At present, eye-tracking remains specialized, technical equipment to use for this purpose, and it may be cost prohibitive for many practicing clinicians when added to the investment of standard test kits and forms. However, as popularity of eye tracking increases and the technology progresses, future versions of standard equipment and software will be increasingly affordable, user-friendly, and capable of the nuanced indices created in this study. In terms of return on investment, it seems noteworthy that the cost of covering one patient for damages associated with TBI can far exceed the cost of the equipment. Nonetheless, it is important to emphasize that this study provides evidence for proof-of-concept only, and actual application of it in a clinical setting will require considerable refinement. Related to this point, although the effect sizes for several pupillary indices met the criteria for “medium” and “large” based on the benchmarks suggested by Cohen (1988), the interpretation of effect size depends heavily on the context in which it is interpreted. To aid in evaluating the current findings, detailed information regarding the sensitivity and specificity as well as applied predictive powers in the context of clinical

and nonclinical samples were provided. Even so, according to Zakzanis (2001), effect sizes larger than Cohen's d 3.0 should be demonstrated for establishing test indexes to be applied in neuropsychological diagnostic contexts. Although this rigorous criterion is ideal for the clinical setting, adopting such high criteria may prematurely quash the development of ideas in a proof-of-concept stage. In their current form, these new pupillary indexes do not meet those criteria for clinical application, and it will take improvement and replications in diverse samples before they could be adopted in clinical batteries.

Although this computerized version of the TOMM used the same instructions and stimuli as the paper version, it was a non-standard version of the test. A computer delivers stimuli more precisely than human examiners, but the presence of a human examiner may elicit unique effects from examinees. Additionally, there were demographic differences between groups such that the simulator group was younger than both the TBI and healthy comparison groups, and both healthy adult groups had more years of education than the TBI group. Ultimately, the differences in age and level of education created a more rigorous test of the pupillometry indices, because these factors are positively related to successfully feigning impairment (Rapport et al., 1998). Lastly, these were adults with moderate to severe TBI, and the findings cannot be generalized to mild TBI, which is likely of greatest interest to neuropsychologists in forensic settings, or other disorders characterized by cognitive impairment. Replication with independent samples is needed to examine the generalizability of these findings.

Conclusions and Future Directions

This study supports the use of biometric indicators to enhance the diagnostic accuracy of performance validity tests. A constellation of findings indicated that pupillary indices were tapping cognitive processes as predicted by theory: modest but not redundant relation to accuracy, effectiveness in identifying persons known to be feigning, and especially the nuanced patterns regarding consistency (or inconsistency) between automatic physiological response and behavior. Although performance validity assessment can view feigned impairment as “poor effort,” these findings indicate that intentional poor performance on PVTs requires substantially more cognitive resources and effort than best-effort performance. Of note, potential sensitivity of pupillometry to the physiological sequelae of TBI appeared to enhance its capacity to identify feigning rather than limit it. Adults with TBI showed smaller pupil size during PVT performance than their healthy counterparts who feigned TBI or performed at their best, which may reflect compromise of the integrity of the visual system (Armstrong, 2018; Ciuffreda & Ludlam, 2011) or indirect sequelae, such as diminished engagement or apathy commonly observed after TBI (Worthington & Wood, 2018). Nonetheless, adults with TBI showed normal and expected pupillary patterns of recognition for familiar stimuli, similar to healthy adults performing at their best, and they responded consistent with their physiologic indicators of familiarity.

The utility of dilation indices for items answered incorrectly was strong despite the fact that few participants answered items incorrectly. Low frequency of incorrect responses is common for the TOMM, which typically shows a marked ceiling effect. Examination

of these pupillary indices on PVTs or clinical tests with a greater range of scores may strengthen the utility of indices related to incorrect responses.

Taken together, these findings provide strong support for multimethod assessment, such as adding unique performance assessments like biometrics to standard accuracy scoring metrics. These additions can enhance testing data that are considered in the context of comprehensive information needed for diagnosis and treatment planning (Sherman, Slick, & Iverson, 2020), which include extratest information such as history and behavioral observations, as well as embedded PVTs. This study establishes that experimental paradigms in pupillometry can be applied to common clinical tests, but considerable research is needed before applied criteria can be created and formally incorporated into the diagnostic process. Continued study of pupillometry in the clinical setting will provide useful, unique information to help identify individuals feigning impairment who are not detected by traditional performance validity test scoring metrics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Carole Koviak and Robert Kotasek of the Southeastern Michigan Traumatic Brain Injury Model System; Shoshana Krohner, Barret Vermilion and Monica De Iorio, Wayne State University Psychology Department, for their invaluable help and support in data collection. Thanks to the NIH Initiative for Maximizing Student Development biomedical training program for their support (S. D. Patrick).

This research was supported by grants from the National Institute on Disability, Independent Living and Rehabilitation Research (NIDILRR 90IF0092, Rapport), Wayne State University Graduate School (Kanser), NIGMS/NIH grant R25 GM 058905 – 21 (Patrick), and Blue Cross Blue Shield of Michigan Foundation (Patrick, 002859.SAP), and collaboration with NIDILRR 90DPTB0006 (Hanks).

References

- Armstrong RA (2018). Visual problems associated with traumatic brain injury. *Clinical and Experimental Optometry*, 101(6), 716–726. doi:10.1111/cxo.12670 [PubMed: 29488253]
- Bauer L, & McCaffrey RJ (2006). Coverage of the Test of Memory Malingering, Victoria Symptom Validity Test, and Word Memory Test on the Internet: Is test security threatened? *Archives of Clinical Neuropsychology*, 21(1), 121–126. doi:10.1016/j.acn.2005.06.010 [PubMed: 16122901]
- Beatty J (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2), 276–292. doi:10.1037/0033-2909.91.2.276 [PubMed: 7071262]
- Bodner T, Merten T, & Benke T (2019). Performance validity measures in clinical patients with aphasia. *Journal of Clinical and Experimental Neuropsychology*, 41(5), 476–483. doi:10.1080/13803395.2019.1579783 [PubMed: 30794037]
- Boone KB (2013). *Clinical practice of forensic neuropsychology: an evidence-based approach* New York: Guildford Press.
- Brennan AM, Meyer S, David E, Pella R, Hill BD, & Gouvier WD (2009). The vulnerability to coaching across measures of effort. *The Clinical Neuropsychologist*, 23(2), 314–328. doi:10.1080/13854040802054151 [PubMed: 18609324]
- Ciuffreda KJ, Joshi NR, & Truong JQ (2017). Understanding the effects of mild traumatic brain injury on the pupillary light reflex. *Concussion*, 2(3), Cnc36. doi:10.2217/cnc-2016-0029 [PubMed: 30202579]

- Ciuffreda KJ, & Ludlam DP (2011). Objective diagnostic and interventional vision test protocol for the mild traumatic brain injury population. *Optometry-Journal of the American Optometric Association*, 82(6), 337–339. doi:10.1016/j.optm.2011.03.006
- Cohen J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Routledge.
- Coleman RD, Rapport LJ, Millis SR, Ricker JH, & Farchione TJ (1998). Effects of coaching on detection of malingering on the California Verbal Learning Test. *Journal of Clinical and Experimental Neuropsychology*, 20(2), 201–210. doi:10.1076/jcen.20.2.201.1164 [PubMed: 9777474]
- Constantinou M, Bauer L, Ashendorf L, Fisher JM, & McCaffrey RJ (2005). Is poor performance on recognition memory effort measures indicative of generalized poor performance on neuropsychological tests? *Archives of Clinical Neuropsychology*, 20(2), 191–198. doi:10.1016/j.acn.2004.06.002 [PubMed: 15708729]
- Daubert v. Merrell Dow Pharmaceuticals Inc. (1993). In (pp. 509 U.S. 579, 113 S. Ct. 2786, 2125 L. Ed. 2782d 2469).
- Dionisio DP (2001). Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology*, 38(2), 205–211. doi:10.1111/1469-8986.3820205 [PubMed: 11347866]
- Farah MJ, Hutchinson JB, Phelps EA, & Wagner AD (2014). Functional MRI-based lie detection: scientific and societal challenges. *Nature Reviews Neuroscience*, 15(2), 123–131. doi:10.1038/nrn3665
- Flaro L, Green P, & Robertson E (2007). Word Memory Test failure 23 times higher in mild brain injury than in parents seeking custody: The power of external incentives. *Brain Injury*, 21(4), 373–383. doi:10.1080/02699050701311133 [PubMed: 17487635]
- Fuermaier ABM, Tucha O, Russ D, Ehrenstein JK, Stanke M, Heindorf R, . . . Tucha L (2020). Utility of an attention-based performance validity test for the detection of feigned cognitive dysfunction after acquired brain injury. *Journal of Clinical and Experimental Neuropsychology*, 42(3), 285–297. doi:10.1080/13803395.2019.1710468 [PubMed: 32041477]
- Green P, Rohling ML, Lees-Haley PR, & Allen LM 3rd. (2011). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury*, 15(12), 1045–1060. doi:10.1080/02699050110088254
- Gunstad J, & Suhr JA (2001). Efficacy of the full and abbreviated forms of the Portland Digit Recognition Test: Vulnerability to coaching. *The Clinical Neuropsychologist*, 15(3), 397–404. doi:10.1076/clin.15.3.397.10271 [PubMed: 11778778]
- Heaver B, & Hutton SB (2010). Keeping an eye on the truth: Pupil size, recognition memory and malingering. *International Journal of Psychophysiology*, 77(3), 306–306. doi:10.1016/j.ijpsycho.2010.06.206
- Heilbronner RL, Sweet JJ, Morgan JE, Larrabee GJ, Millis SR, & Conference P (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. doi:10.1080/13854040903155063 [PubMed: 19735055]
- Hosmer D, Lemeshow S, & Sturdivant R (2013). *Applied logistic regression* (3rd ed.). Hoboken, NJ: Wiley.
- Hu C, Huang K, Hu X, Liu Y, Yuan F, Wang Q, & Fu G (2015). Measuring the cognitive resources consumed per second for real-time lie-production and recollection: a dual-tasking paradigm. *Frontiers in Psychology*, 6. doi:10.3389/fpsyg.2015.00596
- Kahneman D, & Beatty J (1966). Pupil diameter and load on memory. *Science*, 154(3756), 1583–1585. Retrieved from <http://www.jstor.org/stable/1720478> [PubMed: 5924930]
- Kahneman D, & Peavler WS (1969). Incentive effects and pupillary changes in association learning. *Journal of Experimental Psychology*, 79(2, pt.1), 312–318. doi:10.1037/h0026912 [PubMed: 5785645]
- Kanser RJ, Bashem JR, Patrick SD, Hanks RA, & Rapport LJ (2020). Detecting feigned traumatic brain injury with eye tracking during a test of performance validity. *Neuropsychology*.

- Kanser RJ, Rapport LJ, Bashem JR, Billings NM, Hanks RA, Axelrod BN, & Miller JB (2017). Strategies of successful and unsuccessful simulators coached to feign traumatic brain injury. *The Clinical Neuropsychologist*, 31(3), 644–653. doi:10.1080/13854046.2016.1278040 [PubMed: 28084893]
- Kanser RJ, Rapport LJ, Bashem JR, & Hanks RA (2019). Detecting malingering in traumatic brain injury: Combining response time with performance validity test accuracy. *The Clinical Neuropsychologist*, 33(1), 90–107. doi:10.1080/13854046.2018.1440006 [PubMed: 29469640]
- Langleben DD, & Moriarty JC (2013). Using brain imaging for lie detection: Where science, law and research policy collide. *Psychol Public Policy Law*, 19(2), 222–234. doi:10.1037/a0028841 [PubMed: 23772173]
- Larrabee GJ (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist*, 17(3), 410–425. doi:10.1076/clin.17.3.410.18089 [PubMed: 14704892]
- Larrabee GJ (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, 18(4), 625–631. doi:10.1017/s1355617712000240 [PubMed: 23057079]
- Larrabee GJ, Millis SR, & Meyers JE (2009). 40 plus or minus 10, a new magical number: reply to Russell. *The Clinical Neuropsychologist*, 23(5), 841–849. doi:10.1080/13854040902796735 [PubMed: 19343592]
- Loewenfeld IE, & Lowenstein O (1999). *The pupil: Anatomy, physiology, and clinical applications*: Butterworth-Heinemann.
- Lupu T, Elbaum T, Wagner M, & Braw Y (2018). Enhanced detection of feigned cognitive impairment using per item response time measurements in the Word Memory Test. *Applied Neuropsychology. Adult*, 25(6), 532–542. doi:10.1080/23279095.2017.1341410 [PubMed: 28696772]
- Meyers J, Reinsch-Boothby L, Miller R, Rohling M, & Axelrod B (2011). Does the source of a forensic referral affect neuropsychological test performance on a standardized battery of tests? *The Clinical Neuropsychologist*, 25(3), 477–487. doi:10.1080/13854046.2011.554442 [PubMed: 21491350]
- Murphy PR, O'Connell RG, O'Sullivan M, Robertson IH, & Balsters JH (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping*, 35(8), 4140–4154. doi:10.1002/hbm.22466 [PubMed: 24510607]
- Neudecker JJ, & Skeel RL (2009). Development of a novel malingering detection method involving multiple detection strategies. *Archives of Clinical Neuropsychology*, 24(1), 59–70. doi:10.1093/arclin/acp008 [PubMed: 19395357]
- Oliva M (2019). Pupil size and search performance in low and high perceptual load. *Cognitive Affective & Behavioral Neuroscience*, 19(2), 366–376. doi:10.3758/s13415-018-00677-w
- Otero SC, Weekes BS, & Hutton SB (2011). Pupil size changes during recognition memory. *Psychophysiology*, 48(10), 1346–1353. doi:10.1111/j.1469-8986.2011.01217.x [PubMed: 21575007]
- Patrick SD, Rapport LJ, Kanser RJ, Hanks RA, & Bashem JR (2020). Performance validity assessment using response time on the Warrington Recognition Memory Test. *The Clinical Neuropsychologist*, 1–20. doi:10.1080/13854046.2020.1716997
- Porter G, Troscianko T, & Gilchrist ID (2007). Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal of Experimental Psychology*, 60(2), 211–229. doi:10.1080/17470210600673818 [PubMed: 17455055]
- Rapport LJ, Farchione TJ, Coleman RD, & Axelrod BN (1998). Effects of coaching on malingered motor function profiles. *Journal of Clinical and Experimental Neuropsychology*, 20(1), 89–97. Retrieved from <Go to ISI>://WOS:000074688100008 [PubMed: 9672822]
- Rogers R (2008). An introduction to response styles. In Rogers R (Ed.), *Clinical assessment of malingering and deception* (Third ed., pp. 3–13). New York: Guilford Press.
- Rose FE, Hall S, Szalda-Petree AD, & Bach PJ (1998). A comparison of four tests of malingering and the effects of coaching. *Archives of Clinical Neuropsychology*, 13(4), 349–363. doi:10.1016/S0887-6177(97)00025-5 [PubMed: 14590614]

- Russeler J, Brett A, Klaue U, Sailer M, & Munte TF (2008). The effect of coaching on the simulated malingering of memory impairment. *BMC Neurology*, 8, 37. doi:10.1186/1471-2377-8-37 [PubMed: 18838010]
- Sharland MJ, & Gfeller JD (2007). A survey of neuropsychologists' beliefs and practices with respect to the assessment of effort. *Archives of Clinical Neuropsychology*, 22(2), 213–223. doi:10.1016/j.acn.2006.12.004 [PubMed: 17284353]
- Sherman EMS, Slick DJ, & Iverson GL (2020). Multidimensional Malingering Criteria for Neuropsychological Assessment: A 20-Year Update of the Malingered Neuropsychological Dysfunction Criteria. *Archives of Clinical Neuropsychology*, 35(6), 735–764. doi:10.1093/arclin/aaa019 [PubMed: 32377667]
- Slick DJ, Tan JE, Strauss EH, & Hultsch DF (2004). Detecting malingering: a survey of experts' practices. *Archives of Clinical Neuropsychology*, 19(4), 465–473. doi:10.1016/j.acn.2003.04.001 [PubMed: 15163448]
- Strauss E, Sherman EMS, & Spreen O (2006). *A compendium of neuropsychological tests: administration, norms, and commentary*: Oxford University Press.
- Suhr JA, & Gunstad J (2007). Coaching and malingering: A review. In Larrabee GJ (Ed.), *Assessment of malingered neurocognitive deficits* (pp. 287–311). New York: Oxford University Press.
- Tabachnick BG, Fidell LS, & Ullman JB (2018). *Using multivariate statistics* (7th ed.): Pearson.
- Tan JE, Slick DJ, Strauss E, & Hultsch DF (2002). How'd they do it? Malingering strategies on symptom validity tests. *The Clinical Neuropsychologist*, 16(4), 495–505. doi:10.1076/clin.16.4.495.13909 [PubMed: 12822058]
- The Psychological Corporation. (2001). *Wechsler Test of Adult Reading (WTAR)* San Antonio, TX: The Psychological Corporation.
- Tombaugh TN (1996). *Test of Memory Malingering (TOMM)* New York, NY: Multi Health Systems.
- Tombaugh TN, & Tombaugh TN (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, 9(3), 260–268. doi:10.1037/1040-3590.9.3.260
- Tomer E, Lupu T, Golan L, Wagner M, & Braw Y (2018). Eye tracking as a mean to detect feigned cognitive impairment in the word memory test. *Applied Neuropsychology-Adult*, 1–13. doi:10.1080/23279095.2018.1480483
- Valko PO, Gavrilov YV, Yamamoto M, Noain D, Reddy H, Haybaeck J, . . . Scammell TE (2016). Damage to Arousal-Promoting Brainstem Neurons with Traumatic Brain Injury. *Sleep*, 39(6), 1249–1252. doi:10.5665/sleep.5844 [PubMed: 27091531]
- Vickery CD, Berry DTR, Inman TH, Harris MJ, & Orey SA (2001). Detection of inadequate effort on neuropsychological testing: A meta-analytic review of selected procedures. *Archives of Clinical Neuropsychology*, 16(1), 45–73. doi:10.1016/S0887-6177(99)00058-X [PubMed: 14590192]
- Võ MLH, Jacobs AM, Kuchinke L, Hofmann M, Conrad M, Schacht A, & Hutzler F (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, 45(1), 130–140. doi:10.1111/j.1469-8986.2007.00606.x [PubMed: 17910733]
- White DJ, Korinek D, Bernstein MT, Ovsiew GP, Resch ZJ, & Soble JR (2020). Cross-validation of non-memory-based embedded performance validity tests for detecting invalid performance among patients with and without neurocognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, 1–14. doi:10.1080/13803395.2020.1758634
- Worthington A, & Wood RL (2018). Apathy following traumatic brain injury: A review. *Neuropsychologia*, 118, 40–47. doi:10.1016/j.neuropsychologia.2018.04.012
- Youden WJ (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. Retrieved from 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3 [PubMed: 15405679]
- Zakzanis KK (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology*, 16(7), 653–667. [PubMed: 14589784]

Key Points

Question:

Is pupillometry a useful biometric measure for identifying feigned cognitive impairment?

Findings:

Several pupillary indices discriminated feigned impairment from healthy adults and adults with TBI instructed to perform their best, both as independent indicators and beyond traditional TOMM accuracy scores.

Importance:

The findings support the utility of biometric measures, such as pupillometry, in the context of performance validity assessment.

Next Steps:

The examination of pupillometry may enhance the detection of individuals feigning cognitive impairment who are not identified by traditional performance validity tests.

Table 1.

Pupillary Indices: Operational Definitions of Constructs Assessed

Pupil Variable	Construct	Description
<i>Effort-Max</i>	Cognitive effort	Average maximum pupil dilation
<i>Effort-Baseline</i>		Average minimum pupil dilation
<i>View-Familiar (Max)</i>	Familiarity	Average maximum pupil dilation when viewing correct response option
<i>View-Familiar (Baseline)</i>		Average minimum pupil dilation when viewing correct response option
<i>View-Unfamiliar (Max)</i>	Unfamiliarity	Average maximum pupil dilation when viewing incorrect response option
<i>View-Unfamiliar (Baseline)</i>		Average minimum pupil dilation when viewing incorrect response option
<i>Response-Familiar (Max)</i>	Cognitive effort when answering correctly (truthfully)	Average maximum pupil dilation on items answered correctly
<i>Response-Familiar (Baseline)</i>		Average minimum pupil dilation on items answered correctly
<i>Response-Unfamiliar (Max)</i>	Cognitive effort when answering incorrectly (untruthfully if simulator)	Average maximum pupil dilation on items answered incorrectly
<i>Response-Unfamiliar (Baseline)</i>		Average minimum pupil dilation on items answered incorrectly
<i>Effort-Range</i>	Range in cognitive effort	Average minimum pupil dilation subtracted from the average maximum pupil dilation
<i>Familiarity-Difference</i>	Difference between familiar items and unfamiliar items	Average maximum pupil dilation when viewing incorrect response option subtracted from correct response option.

Table 2. Descriptive Statistics and Group Comparisons¹ of Test of Memory Malingering (TOMM) Trial 1 Performance for Healthy comparison (HC, n = 72), Traumatic Brain Injury (TBI, n = 53) Groups, and Simulator (SIM, n = 52)

Variable	HC			TBI			SIM			Significant Contrasts
	M	SD	η^2	M	SD	η^2	M	SD	η^2	
<u>Trial 1 Accuracy</u>	47.29	3.20	0.38	44.00	4.92	0.38	36.48	7.92	0.43	HC > TBI > SIM
Effort-Max	2.70	0.31	.18	2.53	0.25	.18	2.95	0.44	.18	SIM > HC > TBI
View-Familiar (Max)	2.66	0.31	.19	2.47	0.24	.19	2.89	0.43	.19	SIM > HC > TBI
View-Unfamiliar (Max)	2.64	0.31	.18	2.46	0.25	.18	2.88	0.43	.18	SIM > HC > TBI
Response-Familiar (Max)	2.71	0.30	.18	2.53	0.25	.18	2.95	0.44	.18	SIM > HC > TBI
<u>Response-Unfamiliar (Max)</u>	2.76	0.32	.23	2.52	0.22	.23	2.98	0.44	.23	SIM > HC > TBI
Effort-Baseline	2.43	0.29	.15	2.22	0.21	.15	2.52	0.38	.15	SIM = HC > TBI
View-Familiar (Baseline)	2.48	0.29	.17	2.27	0.22	.17	2.61	0.38	.17	SIM > HC > TBI
View-Unfamiliar (Baseline)	2.49	0.29	.15	2.29	0.23	.15	2.61	0.38	.15	SIM = HC > TBI
Response-Familiar (Baseline)	2.44	0.28	.16	2.22	0.21	.16	2.52	0.38	.16	SIM = HC > TBI
<u>Response-Unfamiliar (Baseline)</u>	2.40	0.32	.18	2.19	0.21	.18	2.53	0.39	.18	SIM = HC > TBI
Effort-Range	0.27	0.09	.16	0.31	0.10	.16	0.43	0.20	.16	SIM > TBI > HC
Familiarity-Difference	0.02	0.02	.06	0.01	0.02	.06	0.01	0.02	.06	HC = TBI > SIM
<u>Trial 2 Accuracy</u>	49.61	0.98	0.43	48.94	2.24	0.43	38.76	9.73	0.43	HC > TBI > SIM
Effort-Max	2.66	0.34	.15	2.51	0.25	.15	2.90	0.42	.15	SIM > HC > TBI
View-Familiar (Max)	2.62	0.34	.15	2.47	0.27	.15	2.83	0.40	.15	SIM > HC > TBI
View-Unfamiliar (Max)	2.61	0.34	.16	2.45	0.24	.16	2.83	0.40	.16	SIM > HC > TBI
Response-Familiar (Max)	2.67	0.34	.16	2.51	0.25	.16	2.90	0.41	.16	SIM > HC > TBI
<u>Response-Unfamiliar (Max)</u>	2.63	0.34	.20	2.49	0.29	.20	2.94	0.46	.20	SIM > HC = TBI
Effort-Baseline	2.43	0.32	.12	2.23	0.22	.12	2.50	0.36	.12	SIM = HC > TBI
View-Familiar (Baseline)	2.48	0.33	.12	2.30	0.26	.12	2.58	0.36	.12	SIM = HC > TBI
View-Unfamiliar (Baseline)	2.48	0.33	.13	2.28	0.22	.13	2.58	0.37	.13	SIM = HC > TBI
Response-Familiar (Baseline)	2.44	0.31	.13	2.23	0.22	.13	2.50	0.35	.13	SIM = HC > TBI
<u>Response-Unfamiliar (Baseline)</u>	2.34	0.29	.13	2.21	0.22	.13	2.50	0.39	.13	SIM = HC > TBI
Effort-Range	0.23	0.08	.18	0.29	0.10	.18	0.40	0.21	.18	SIM > TBI > HC

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Variable	HC		TBI		SIM		H(2)	p	η^2	Contrasts
	M	SD	M	SD	M	SD				
Familiarity-Difference	0.01	0.02	0.00	0.02	0.00	0.02	4.52	< .001	.03	TBI = SIM, HC; HC > SIM

Note. Kruskal-Wallis and Mann-Whitney tests. Effect size η^2 estimated from Kruskal-Wallis H. See Table 1 for variable abbreviation key.

Classification Statistics: TOMM Trial 1 Performance for Single and Two-variable Models Predicting Simulator (SIM) and Traumatic Brain Injury (TBI) Group Membership

Table 3a.

	S_{Sp0}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ^2	Predictor p
One-Variable Models:										
Trial 1 Accuracy	.46-.54	0.40	.63	.76	.23	.95	.78	[.69, .87]	29.83***	
Effort-Max	.58	0.48	.69	.79	.26	.96	.80	[.71, .88]	33.23***	
View-Familiar (Max)	.56-.58	0.50	.70	.80	.26	.96	.81	[.72, .89]	34.45***	
View-Unfamiliar (Max)	.56-.58	0.46	.67	.79	.24	.96	.80	[.72, .89]	33.97***	
Response-Familiar (Max)	.58	0.48	.69	.79	.26	.96	.80	[.71, .88]	33.74***	
Response-Unfamiliar (Max)	.60	0.50	.68	.80	.27	.96	.83	[.74, .91]	36.94***	
Effort-Baseline	.44	0.39	.61	.76	.21	.96	.77	[.68, .86]	25.50***	
View-Familiar (Baseline)	.44-.48	0.40	.62	.77	.21	.96	.80	[.70, .88]	29.84***	
View-Unfamiliar (Baseline)	.40-.44	0.39	.63	.76	.22	.95	.77	[.68, .86]	26.53***	
Response-Familiar (Baseline)	.44-.46	0.37	.60	.75	.21	.94	.77	[.68, .86]	26.21***	
Response-Unfamiliar (Baseline)	.46-.48	0.43	.61	.80	.21	.97	.79	[.70, .88]	27.78***	
Effort-Range	.33-.35	0.33	.63	.73	.21	.94	.69	[.58, .79]	15.18***	
Familiarity-Difference	.19-.21	0.19	.48	.69	.14	.93	.61	[.50, .72]	2.25	
Two-Variable Models: Accuracy +										
Effort-Max	.69	0.57	.73	.84	.31	.97	.88	[.82, .95]	58.90***	<.001
View-Familiar (Max)	.71-.73	0.61	.74	.85	.33	.97	.89	[.83, .95]	60.49***	<.001
View-Unfamiliar (Max)	.75-.77	0.59	.73	.85	.31	.97	.89	[.83, .95]	60.63***	<.001
Response-Familiar (Max)	.69	0.57	.73	.84	.31	.97	.88	[.82, .95]	59.18***	<.001
Response-Unfamiliar (Max)	.74	0.62	.71	.88	.30	.97	.89	[.83, .95]	54.34***	<.001
Effort-Baseline	.78	0.63	.75	.86	.33	.97	.90	[.83, .96]	63.91***	<.001
View-Familiar (Baseline)	.80	0.63	.75	.86	.33	.97	.89	[.83, .95]	60.00***	<.001
View-Unfamiliar (Baseline)	.78	0.63	.75	.87	.33	.98	.89	[.82, .95]	58.92***	<.001
Response-Familiar (Baseline)	.78	0.61	.74	.85	.33	.97	.90	[.84, .96]	64.44***	<.001
Response-Unfamiliar (Baseline)	.78-.80	0.62	.74	.87	.32	.97	.90	[.84, .96]	56.03***	<.001

	Sh_{Sp90}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ^2	Predictor p
Effort-Range	.50-.52	0.37	.62	.75	.22	.95	.79	[.71, .88]	32.87***	.092

Note: Sh_{Sp90} = Sensitivity (detection of simulated TB1) when Sp Specificity (bona fide TB1) is set at cut point .90 (range presented when cut point is associated with multiple Sn values); PPP = Positive Predictive Power, NPP = Negative Predictive Power (each presented for 40% and 10% base rate); AUC = ROC area under the curve. See Table 1 for variable abbreviation key.

Classification Statistics: TOMM Trial 2 Performance for Single and Two-variable Models Predicting Simulator (SIM) and Traumatic Brain Injury (TBI) Group Membership.

Table 3b.

	Sh _{SP00}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ ²	Predictor p
One-Variable Models:										
Trial 2 Accuracy	.65	0.57	.84	.80	.50	.95	.85	[.78, .93]	50.66***	
Effort-Max	.59-.61	0.41	.62	.77	.22	.96	.78	[.69, .87]	21.22***	
View-Familiar (Max)	.53	0.37	.60	.76	.21	.96	.78	[.69, .87]	26.14***	
View-Unfamiliar (Max)	.51-.55	0.38	.61	.77	.21	.96	.80	[.71, .88]	30.20***	
Response-Familiar (Max)	.57-.59	0.45	.64	.79	.23	.96	.79	[.70, .88]	29.40***	
Response-Unfamiliar (Max)	.56-.65	0.24	.49	.77	.14	.96	.78	[.67, .90]	15.83***	
Effort-Baseline	.35	0.39	.61	.76	.21	.96	.75	[.65, .84]	21.49***	
View-Familiar (Baseline)	.45-.47	0.41	.62	.77	.22	.96	.75	[.66, .85]	20.08***	
View-Unfamiliar (Baseline)	.43	0.42	.64	.79	.23	.96	.77	[.67, .86]	24.31***	
Response-Familiar (Baseline)	.35	0.39	.61	.76	.21	.96	.75	[.66, .85]	21.81***	
Response-Unfamiliar (Baseline)	.40	0.21	.47	.85	.13	.94	.75	[.63, .87]	10.80**	
Effort-Range	.33-.39	0.29	.58	.71	.20	.93	.65	[.54, .76]	12.62***	
Familiarity-Difference	.12	0.05	.41	.62	.11	.92	.55	[.43, .66]	0.87	
Two-Variable Models: Accuracy +										
Effort-Max	.69-.75	0.63	.83	.83	.44	.96	.89	[.83, .96]	60.10***	.006
View-Familiar (Max)	.67-.73	0.63	.83	.83	.44	.96	.89	[.83, .96]	59.32***	.008
View-Unfamiliar (Max)	.71-.75	0.63	.83	.83	.44	.96	.89	[.83, .95]	60.41***	.004
Response-Familiar (Max)	.69-.75	0.63	.83	.83	.44	.96	.89	[.83, .96]	60.25***	.006
Response-Unfamiliar (Max)	.74-.77	0.59	.70	.88	.26	.98	.90	[.83, .97]	33.97***	.041
Effort-Baseline	.75-.77	0.67	.86	.85	.53	.97	.89	[.83, .95]	61.58***	.004
View-Familiar (Baseline)	.65-.69	0.57	.80	.80	.39	.96	.90	[.83, .95]	59.18***	.009
View-Unfamiliar (Baseline)	.77	0.64	.81	.84	.42	.98	.89	[.83, .96]	61.59***	.003
Response-Familiar (Baseline)	.75-.77	0.67	.86	.85	.53	.97	.89	[.83, .95]	61.66**	.004
Response-Unfamiliar (Baseline)	.70-.81	0.61	.71	.86	.31	.98	.91	[.84, .98]	35.55***	.041

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	Sh_{Sp90}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ^2	Predictor p
Effort-Range	.65-.71	0.57	.84	.80	.50	.95	.85	[.77, .93]	50.79***	.719

Note: Sh_{Sp90} = Sensitivity (detection of simulated TBI) when Sp Specificity (bona fide TBI) is set at cut point .90 (range presented when cut point is associated with multiple Sn values); PPP = Positive Predictive Power, NPP = Negative Predictive Power (each presented for 40% and 10% base rate); AUC = ROC area under the curve. See Table 1 for pupil variable abbreviation key.

Table 4a. Classification Statistics: TOMM Trial 1 Performance for Single and Two-variable Models Predicting Simulator (SIM) and Healthy Comparison (HC) Group Membership.

	Sh _{Sp90}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ^2	Predictor p
One-Variable Models:										
Trial 1 Accuracy	.67	0.64	.80	.85	.39	.97	.90	[.84, .96]	74.28***	
Effort-Max	.33-.35	0.25	.59	.69	.21	.93	.67	[.57, .77]	13.29***	
View-Familiar (Max)	.31	0.24	.58	.69	.19	.93	.66	[.56, .76]	11.40**	
View-Unfamiliar (Max)	.33	0.24	.58	.69	.19	.93	.66	[.56, .76]	12.33***	
Response-Familiar (Max)	.33-.35	0.24	.58	.69	.20	.92	.66	[.56, .76]	12.81***	
Response-Unfamiliar (Max)	.32-.34	0.24	.53	.71	.15	.94	.64	[.53, .75]	8.56**	
Effort-Baseline	.17	0.04	.46	.61	.17	.90	.55	[.45, .65]	2.12	
View-Familiar (Baseline)	.14-.15	0.05	.47	.61	.12	.90	.59	[.49, .69]	4.37*	
View-Unfamiliar (Baseline)	.15	0.04	.44	.61	.13	.91	.58	[.48, .69]	3.74	
Response-Familiar (Baseline)	.15-.17	0.04	.46	.61	.17	.90	.54	[.44, .65]	1.86	
Response-Unfamiliar (Baseline)	.12	0.08	.45	.63	.13	.90	.58	[.47, .70]	3.25	
Effort-Range	.54	0.45	.74	.76	.33	.95	.77	[.68, .86]	34.37***	
Familiarity-Difference	.19-.21	0.16	.55	.65	.16	.92	.66	[.56, .76]	7.50**	
Two-Variable Models: Accuracy +										
Effort-Max	.76-.82	0.68	.84	.86	.45	.97	.94	[.90, .98]	90.76***	.011
View-Familiar (Max)	.76-.80	0.66	.84	.85	.45	.97	.94	[.90, .98]	90.49***	.012
View-Unfamiliar (Max)	.76-.82	0.66	.84	.85	.45	.97	.94	[.90, .98]	90.90***	.010
Response-Familiar (Max)	.76-.82	0.68	.84	.86	.45	.97	.94	[.90, .98]	90.90***	.010
Response-Unfamiliar (Max)	.66-.76	0.63	.79	.84	.40	.98	.92	[.87, .97]	72.19***	.016
View-Familiar (Baseline)	.76-.78	0.66	.84	.85	.45	.97	.94	[.90, .98]	88.14***	.027
Effort-Range	.71-.75	0.62	.80	.84	.39	.97	.93	[.89, .97]	83.76***	.023
Familiarity-Difference	.69-.83	0.69	.84	.87	.48	.97	.90	[.84, .96]	77.49***	.081

Note. Sh_{Sp90} = Sensitivity (detection of simulated TBI) when Sp Specificity (bona fide TBI) is set at cut point .90 (range presented when cut point is associated with multiple Sn values); PPP = Positive Predictive Power, NPP = Negative Predictive Power (each presented for 40% and 10% base rate); AUC = ROC area under the curve. See Table 1 for pupil variable abbreviation key.

Classification Statistics: TOMM Trial 2 Performance for Single and Two-variable Models Predicting Simulator (SIM) and Healthy Comparison (HC) Group Membership.

Table 4b.

	Sh _{Sp90}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ^2	Predictor p
One-Variable Models:										
Trial 2 Accuracy	.84	0.71	.93	.85	.69	.97	.89	[.82, .96]	82.65***	
Effort-Max	.33	0.28	.66	.69	.23	.93	.68	[.58, .78]	12.02**	
View-Familiar (Max)	.29	0.16	.58	.65	.19	.92	.66	[.56, .76]	9.69**	
View-Unfamiliar (Max)	.29	0.20	.61	.66	.19	.92	.66	[.56, .76]	10.07**	
Response-Familiar (Max)	.33	0.26	.65	.68	.24	.93	.67	[.57, .77]	10.26**	
Response-Unfamiliar (Max)	.37	0.07	.42	.80	.11	1.00	.71	[.58, .85]	6.58*	
Effort-Baseline	.16-.18	0.04	.57	.61	.20	.91	.54	[.44, .65]	1.28	
View-Familiar (Baseline)	.16	0.04	.50	.61	.10	.90	.58	[.48, .68]	2.75	
View-Unfamiliar (Baseline)	.16	0.03	.50	.61	.11	.90	.57	[.47, .67]	2.43	
Response-Familiar (Baseline)	.16-.18	0.04	.57	.61	.20	.91	.54	[.44, .64]	1.08	
Response-Unfamiliar (Baseline)	.26	0.06	.41	1.00	.11	1.00	.60	[.44, .76]	2.62	
Effort-Range	.55	0.44	.77	.75	.37	.94	.78	[.70, .86]	38.34***	
Familiarity-Difference	.24	0.12	.61	.63	.21	.91	.61	[.51, .71]	4.65*	
Two-Variable Models: Accuracy +										
Effort-Max	.82-.84	0.70	.95	.85	.75	.97	.90	[.83, .96]	83.96***	.258
View-Familiar (Max)	.82-.84	0.70	.95	.85	.75	.97	.90	[.83, .96]	83.88***	.272
View-Unfamiliar (Max)	.82-.84	0.70	.95	.85	.75	.97	.90	[.83, .96]	83.92***	.265
Response-Familiar (Max)	.82	0.70	.95	.85	.75	.97	.90	[.83, .96]	84.02***	.247
Response-Unfamiliar (Max)	.79	0.72	.83	.89	.45	.98	.95	[.89, 1.00]	37.01***	.351
Effort-Range	.82	0.67	.90	.84	.56	.97	.92	[.86, .97]	86.52***	.053
Familiarity-Difference	.82	0.71	.93	.85	.69	.97	.89	[.82, .96]	87.72***	.783

Note. Sh_{Sp90} = Sensitivity (detection of simulated TBI) when Sp Specificity (bona fide TBI) is set at cut point .90 (range presented when cut point is associated with multiple Sn values); PPP = Positive Predictive Power, NPP = Negative Predictive Power (each presented for 40% and 10% base rate); AUC = ROC area under the curve. See Table 1 for pupil variable abbreviation key.

Classification Statistics: TOMM Trial 1 Performance for Single and Two-variable Models Predicting Healthy Comparison (HC) and Traumatic Brain Injury (TBI) Group Membership.

Table 5a.

	Sh _{SP00}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ^2	Predictor p
One-Variable Models:										
Trial 1 Accuracy	.28	0.37	.71	.72	.29	.94	.71	[.61, .80]	18.21***	
Effort-Max	.08-.23	0.26	.59	.70	.19	.93	.67	[.58, .77]	10.57**	
View-Familiar (Max)	.12-.19	0.30	.59	.72	.19	.94	.70	[.60, .79]	13.46***	
View-Unfamiliar (Max)	.15-.17	0.29	.59	.71	.21	.93	.68	[.58, .77]	11.66**	
Response-Familiar (Max)	.23-.26	0.30	.60	.71	.21	.94	.68	[.59, .78]	11.89**	
Response-Unfamiliar (Max)	.18-.20	0.32	.57	.73	.19	.94	.74	[.64, .84]	16.93***	
Effort-Baseline	.15-.17	0.34	.62	.73	.21	.94	.72	[.63, .81]	19.38***	
View-Familiar (Baseline)	.15-.23	0.32	.60	.73	.21	.94	.71	[.62, .81]	17.78***	
View-Unfamiliar (Baseline)	.15-.25	0.31	.60	.72	.21	.93	.71	[.62, .80]	16.58***	
Response-Familiar (Baseline)	.23	0.35	.62	.74	.22	.94	.74	[.65, .82]	21.36***	
Response-Unfamiliar (Baseline)	.18-.24	0.36	.61	.75	.19	.94	.72	[.62, .82]	14.03***	
Effort-Range	.17-.21	0.16	.57	.65	.18	.92	.62	[.52, .72]	5.87*	
Familiarity-Difference	.10	-0.01	.38	.60	.08	.90	.57	[.47, .67]	1.96	
Two-Variable Models: Accuracy ±										
Effort-Max	.28-.32	0.41	.67	.75	.25	.95	.77	[.68, .86]	26.56***	.006
View-Familiar (Max)	.29-.37	0.40	.66	.75	.24	.95	.78	[.69, .86]	28.81***	.002
View-Unfamiliar (Max)	.25-.38	0.42	.67	.76	.27	.95	.76	[.68, .85]	26.93***	.005
Response-Familiar (Max)	.30-.32	0.39	.67	.75	.25	.95	.77	[.68, .86]	26.45***	.006
Response-Unfamiliar (Max)	.27-.29	0.41	.64	.76	.23	.94	.81	[.73, .90]	31.40***	.001
Effort-Baseline	.36-.40	0.42	.66	.76	.25	.96	.79	[.71, .87]	32.27***	.001
View-Familiar (Baseline)	.35-.37	0.41	.65	.76	.25	.96	.78	[.70, .87]	31.36***	.001
View-Unfamiliar (Baseline)	.36	0.40	.65	.75	.25	.95	.78	[.69, .86]	29.92***	.001
Response-Familiar (Baseline)	.36-.40	0.42	.66	.76	.25	.96	.79	[.71, .87]	32.55***	<.001
Response-Unfamiliar (Baseline)	.24	0.46	.65	.80	.24	.96	.80	[.71, .88]	28.81***	.003

	Sh_{Sp90}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ^2	Predictor p
Effort-Range	.30	0.34	.68	.72	.24	.94	.73	[.64, .82]	28.81***	.109

Note: Sh_{Sp90} = Sensitivity (detection of simulated TBI) when Sp Specificity (bona fide TBI) is set at cut point .90 (range presented when cut point is associated with multiple Sn values); PPP = Positive Predictive Power, NPP = Negative Predictive Power (each presented for 40% and 10% base rate); AUC = ROC area under the curve. See Table 1 for pupil variable abbreviation key.

Table 5b.

Classification Statistics: TOMM Trial 2 Performance for Single and Two-variable Models Predicting Healthy Comparison (HC) and Traumatic Brain Injury (TBI) Group Membership.

	Sh _{Sp90}	Youden J	PPP BR 40%	NPP BR 40%	PPP BR 10%	NPP BR 10%	AUC	AUC 95% CI	χ^2	Predictor p
One-Variable Models:										
Trial 2 Accuracy	.18	0.12	.69	.63	.22	.91	.59	[.49, .70]	5.14*	
Effort-Max	.08-.12	0.07	.48	.62	.13	.91	.63	[.54, .73]	6.56*	
View-Familiar (Max)	.08-.12	0.13	.56	.64	.17	.91	.64	[.55, .74]	6.84**	
View-Unfamiliar (Max)	.08-.12	0.17	.57	.66	.17	.92	.66	[.56, .76]	8.52**	
Response-Familiar (Max)	.12-.18	0.08	.47	.63	.13	.91	.64	[.54, .74]	7.59**	
Response-Unfamiliar (Max)	.10	0.22	.48	.71	.14	.94	.64	[.45, .83]	1.68	
Effort-Baseline	.14	0.19	.54	.67	.16	.92	.70	[.60, .79]	15.18***	
View-Familiar (Baseline)	.14	0.21	.57	.67	.18	.93	.68	[.58, .77]	10.69**	
View-Unfamiliar (Baseline)	.14	0.25	.59	.69	.20	.93	.70	[.60, .79]	14.76***	
Response-Familiar (Baseline)	.14-.24	0.26	.59	.70	.19	.93	.71	[.62, .80]	16.70***	
Response-Unfamiliar (Baseline)	.10	0.22	.48	.71	.14	.94	.65	[.46, .83]	2.29	
Effort-Range	.24-.30	0.24	.66	.68	.25	.92	.68	[.58, .77]	11.90**	
Familiarity-Difference	.16-.18	-0.01	.33	.60	.00	.90	.57	[.46, .67]	1.07	
Two-Variable Models: Accuracy +										
Effort-Max	.16-.18	0.11	.54	.63	.14	.91	.69	[.59, .78]	12.23**	.012
View-Familiar (Max)	.16-.18	0.20	.58	.67	.20	.92	.69	[.59, .79]	12.43**	.011
View-Unfamiliar (Max)	.16-.18	0.19	.58	.66	.17	.92	.70	[.61, .79]	14.09**	.005
Response-Familiar (Max)	.16-.18	0.13	.56	.64	.16	.91	.69	[.59, .78]	12.10**	.013
Effort-Baseline	.20-.22	0.33	.63	.72	.23	.93	.73	[.64, .82]	20.35***	<.001
View-Familiar (Baseline)	.16	0.30	.62	.71	.21	.94	.71	[.62, .81]	15.89***	.002
View-Unfamiliar (Baseline)	.18-.20	0.32	.63	.71	.21	.93	.73	[.64, .82]	19.50***	.001
Response-Familiar (Baseline)	.20-.22	0.33	.63	.72	.23	.93	.73	[.64, .82]	20.23***	<.001
Effort-Range	.30-.32	0.19	.59	.66	.17	.92	.70	[.60, .79]	14.51**	.004

Note. Sh_{Sp90} = Sensitivity (detection of simulated TBI) when Sp Specificity (bona fide TBI) is set at cut point .90 (range presented when cut point is associated with multiple Sn values); PPP = Positive Predictive Power, NPP = Negative Predictive Power (each presented for 40% and 10% base rate); AUC = ROC area under the curve. See Table 1 for pupil variable abbreviation key