



Eye-movements reveal children's deliberative thinking and predict performance on arithmetic word problems

Chao-Jung Wu¹  · Chia-Yu Liu¹  · Chung-Hsuan Yang¹ · Yu-Cin Jian¹ 

Received: 17 April 2019 / Revised: 17 December 2019 / Accepted: 2 January 2020

Published online: 14 February 2020

© Instituto Superior de Psicologia Aplicada, Lisboa and Springer Nature B.V. 2020

Abstract

Despite decades of research on the close link between eye movements and human cognitive processes, the exact nature of the link between eye movements and deliberative thinking in problem-solving remains unknown. Thus, this study explored the critical eye-movement indicators of deliberative thinking and investigated whether visual behaviors could predict performance on arithmetic word problems of various difficulties. An eye tracker and test were employed to collect 69 sixth-graders' eye-movement behaviors and responses. No significant difference was found between the successful and unsuccessful groups on the simple problems, but on the difficult problems, the successful problem-solvers demonstrated significantly greater gaze aversion, longer fixations, and spontaneous reflections. Notably, the model incorporating RT-TFD, NOF of 500 ms, and pupil size indicators could best predict participants' performance, with an overall hit rate of 74%, rising to 80% when reading comprehension screening test scores were included. These results reveal the solvers' engagement strategies or show that successful problem-solvers were well aware of problem difficulty and could regulate their cognitive resources efficiently. This study sheds light on the development of an adapted learning system with embedded eye tracking to further predict students' visual behaviors, provide real-time feedback, and improve their problem-solving performance.

Keywords Arithmetic word problems · Deliberation · Eye movements · Problem-solving

✉ Chao-Jung Wu
cjwu@ntnu.edu.tw

Chia-Yu Liu
leave1756@gmail.com

Chung-Hsuan Yang
genieyang0214@gmail.com

Yu-Cin Jian
jianyucin@ntnu.edu.tw

Extended author information available on the last page of the article

Introduction

Problem-solving is a complex cognitive process that involves an interplay between external information input and retrieval of information from memory. It requires searching the external information presented in the problem and cues to find solutions. It also requires orienting to internal activities, such as retrieving knowledge from memory and, more importantly, constructing an appropriate mental representation of the problem. The limited cognitive resources of human beings make it difficult to simultaneously pay attention to the real environment and our mental activities. When confronting a high-workload problem, one must deliberate its mental representation and frequently switch attention between the environment and mental world.

Problem-solvers may close their eyes when constructing mental representations while engaged in a complicated cognitive activity (Doherty-Sneddon et al. 2002; Glenberg et al. 1998). Studies have also found that solving a high cognitive load problem can lead to behavioral or physiological changes in eye movements, e.g., long fixations, larger pupil size (Gredebäck and Melinder 2010; Knoblich et al. 2001; Paas et al. 2003). This suggests that measurements of the critical eye movements of problem-solvers with different abilities hold promise for research into their cognitive tasks, as shown in previous research (Chiou et al. 2019; Krstić et al. 2018; Lin et al. 2016; Wang et al. 2016), thus affording more accurate assistance. The present study identified three categories of eye-movement indicators, gaze aversion, long fixations, and spontaneous reflection. In addition, we reviewed the indicator of gaze duration, which reflects initial processing at the lexical level, and clarified the differences between the lexical level and deliberation in problem-solving.

Theoretical background

Gaze aversion: RT – TFD, TFD of blank, and NOF of blank

Empirical research has already found evidence demonstrating that gazing on areas without stimuli is not due to a lack of focus or inaccurate apparatus (Doherty-Sneddon et al. 2002; Glenberg et al. 1998; Walcher et al. 2017). These studies suggested that this systematic pattern reflects individuals' deep thinking at that moment and could be termed "gaze aversion." Doherty-Sneddon et al. (2002) investigated whether gaze aversion from environmental stimuli facilitates cognitive processing. They asked 5- and 8-year-old children to answer verbal reasoning and arithmetic questions of varying difficulty and recorded videos of their faces. The verbal questions required the children to define and spell words and repeat word lists (e.g., "dog, car, bed"). The arithmetic questions involved addition, subtraction, and multiplication (e.g., "count from 10 to 20"; " $2+8=?$ "; "If John has 8 sweets and Sarah has 9, who has more?"). The results showed that 8-year-old children averted their gaze more frequently when answering difficult questions than easy ones, regardless of whether the questions were verbal or arithmetic. However, 5-year-old children did not show this tendency. The researchers also trained 5-year-old children to use gaze aversion while solving problems and found that the trained group performed significantly better on difficult problems than the control group (Phelps et al. 2006). Doherty-Sneddon et al. (2002; Phelps et al. 2006) concluded that gaze aversion is a good indicator of thinking, and that it helped young children concentrate on

difficult questions. Their findings are consistent with the cognitive load explanation of gaze aversion.

Consistent with the findings of this research (Doherty-Sneddon et al. 2002; Phelps et al. 2006) involving children, Glenberg et al. (1998) used adults as participants and found similar results. They asked undergraduate students to answer general knowledge questions and videotaped their behaviors. The raters independently scored gaze aversion and found that participants' gaze aversion correlated with the accuracy of their answers and the difficulty of questions. They designed another experiment to manipulate whether participants averted their gaze or not with varying levels of difficulty in questions (easy, moderate, or difficult). The two independent variables were designed as within-subjects variables. In the "close eyes" condition, participants were instructed to close their eyes during the interval between reading the questions and the occurrence of an auditory signal to produce the answer. In the "look" condition, they were instructed to look at the experimenter's nose during the interval between question and answer. In addition to the general knowledge questions selected from their previous experiment, the researchers added mathematics questions (e.g., $x + y = z$; $x * y = z$, where the sizes of x and y were adjusted to modify the ease or difficulty of items) as experimental materials. The results found that for moderate to difficult questions, people had higher accuracy when they averted their gaze (i.e., closing their eyes) than when they did not, whether the questions concerned general knowledge or mathematics. Apparently, problem-solvers need to prevent sensory input from external stimuli to avoid interference with their thinking when they face a difficult problem-solving condition.

This literature review indicates that task difficulty is a critical factor in problem-solvers' use of the gaze aversion strategy while completing a problem-solving task. Additionally, problem-solvers' metacognitive ability also determined whether they would adopt the gaze aversion strategy to think longer on difficult tasks, which would ultimately bring the readers success in solving the problems.

All the abovementioned studies explored gaze aversion behavior with video recording (Doherty-Sneddon et al. 2002; Glenberg et al. 1998; Phelps et al. 2006), which retains the original data of the experiment procedure, but converting qualitative data to quantitative data is time-consuming and labor-intensive. For example, gaze aversion numbers in the learning episodes were calculated by two raters. Recently, Micic et al. (2010) began to use eye-tracking technology, which can record online reading behavior and analyze data using a computer, to investigate gaze aversion and cognitive activity. Their cognitive tasks were related to memory rather than problem-solving, with which this study is concerned. However, it shows that use of eye-tracking technology to determine indicators of thinking that reflect cognitive activity is a growing trend (e.g., Abeles and Yuval-Greenberg 2017; Ehrlichman and Micic 2012).

Thus, using eye-tracking technology, this study adopted several indicators of gaze aversion to examine whether they could reveal deliberation in problem-solving. First, we used the total time of closed eyes and gazing beyond the screen as an eye-movement indicator of deliberation, calculated by reaction time minus the total fixation duration (RT – TFD). The total fixation duration of blank areas that contained no stimulus (TFD of blank) is another key indicator of deliberation, supported by the studies of Doherty-Sneddon et al. (2002) and Glenberg et al. (1998). For this indicator, the total time of fixations on blank areas with no stimuli (e.g., symbols, words, numbers, pictures) was

calculated. The number of fixations of blank areas (NOF of blank) is the last key indicator of deliberation, since it is highly correlated with TFD of blank.

Long fixations: NOF of 500 ms

Long fixations are another indicator of deliberation in problem-solving. Ballard et al. (1997) suggested that short and long fixations serve to distinguish different functions. Following their approach, a short fixation is considered to reflect the incorporation of a surface analysis re-encoding problem elements (e.g., plus and equal signs, operands, and results in a mathematical formula) into working memory, and a long fixation is considered to indicate deep processing in completing a cognitive task. Single fixations longer than 500 ms were defined as long fixations. A study using probabilistic inference tasks that instructed participants to make decisions deliberately versus intuitively found that deliberate decision-making leads to more long fixations, whereas intuitive decision-making leads to short fixations (Horstmann et al. 2009).

Based on Ballard et al. (1997), Knoblich et al. (2001) predicted that people would demonstrate more long fixations during a cognitive impasse in a problem-solving situation. They asked undergraduate students to solve matchstick problems (one simpler and two difficult) on a computer screen as quickly as possible and recorded their eye movements. The researchers set three intervals of equal duration for each problem-solver and each task to track processing changes in problem-solving over time. The results showed that people tend to sit and stare at the problem during an impasse when solving a difficult problem. Therefore, they demonstrated a greater percentage of single fixations longer than 500 ms at the moment of impasse, but not for the easy questions. Additionally, the successful problem-solvers had a significantly higher percentage of fixations longer than 500 ms than the unsuccessful problem-solvers on the crucial problem element (e.g., the operators) during the final interval of the problem-solving episode. Based on these findings (Ballard et al. 1997; Horstmann et al. 2009; Knoblich et al. 2001), the total number of single fixation durations exceeding 500 ms (NOF of 500 ms) was considered a critical eye-movement indicator of deliberation in problem-solving in the present study.

Spontaneous reflection: pupil size

Pupillary responses are hard to control at will, being provoked by external or mental events (Siegle et al. 2008). The responsiveness of the indicator thus offers detailed insight into the time and magnitude of cognitive load and mental states (Lu et al. 2018; Piquado et al. 2010). In 1973, Kahneman developed the effort theory of attention, which with respect to pupillary dilations notes that the amount of pupillary dilation increases with difficulty across qualitatively different tasks and might reflect individuals' psychometric capacity. Laeng et al. (2012) suggested that pupillary dilations could be used to estimate the intensity of cognitive activity. Just and Carpenter's (1993) experiments found that comprehending complex sentences during a reading task led to a larger changes in pupil size. Subsequently, several empirical studies all found evidence that people demonstrated a larger mean pupil dilation in learning situations with high cognitive load (Gredebäck and Melinder 2010; Jackson and Sirois 2009; Klingner et al. 2011; Paas et al. 2003).

Siegle et al. (2008) measured pupil dilation with respect to cognitive tasks. They asked adult participants to complete a digit-sorting task and Stroop task, and used pupilometers to

track the location and size of the pupil. In the digit-sorting task, participants were first instructed to view a fixation mask (1 s), which was replaced by a set of three, four, or five digits (2 s), then another fixation mask (5 s). Finally, a new target digit appeared, and participants were required to answer whether the target digit appeared in the previous stage. The results revealed that pupil dilation occurred mainly during sustained information processing. Another study by Klingner et al. (2011) measured the pupillary dilations of 24 undergraduates during three cognitive tasks (i.e., mental multiplication, digit sequence recall, and vigilance) to evaluate the effects of aural and visual presentation modes. The results showed similar patterns of pupillary dilations in all tasks between the two modes, but the magnitude of dilations was significantly larger in the aural mode, perhaps because the dual code in working memory for the auditory presentations led to higher cognitive load than for visual presentations. Furthermore, the results also confirmed that the magnitude of dilations was variable according to the difficulty of tasks.

These studies revealed that the magnitude of pupil dilations could reliably and validly reflect the cognitive loading of tasks and the psychometric capacity of individuals. Thus, this study calculated the pupil sizes (Siegle et al. 2008) as a critical indicator of deliberation in problem-solving and examined the performance of successful and unsuccessful individuals in tasks of varying difficulty.

Rationale for the present study

In school, problem-solving is an essential competency in mathematical learning and is widely considered a highly desirable educational goal. Specifically, *compare word problems*, where the consistency of the relational term with the arithmetic operation determines the difficulty, is one of the critical problem types (Hegarty et al. 1995; Riley and Greeno 1988; Schumacher and Fuchs 2012). Empirical studies have found lower accuracy and longer response times on inconsistent than consistent problems, which is referred to as the consistency effect (Hegarty et al. 1992; Hegarty et al. 1995; Riley and Greeno 1988; Van der Schoot et al. 2009). Successful solvers took more time for inconsistent than consistent problems, and the extra time was localized in the integration stage of problem-solving. However, this response-time pattern was not obtained for unsuccessful solvers. In addition, presenting compare word problems with an illustration might be a way to improve individuals' problem-solving performance (Belenky and Schalk 2014; Booth and Koedinger 2012). Therefore, cognitive demands of different problems are dependent on the interaction of the problem structure with the solver's processes of problem representation.

The study addresses issues in two directions. First, we identify the eye-movement indicators that could reveal the solvers' regulation of cognitive load. These indicators also reveal the deliberation of successful and unsuccessful individuals when solving problems of different difficulty. For the three categories of indicators, we predicted that successful problem-solvers would demonstrate greater gaze aversion (i.e., RT – TFD, TFD of blank, and NOF of blank), more long fixations (i.e., NOF of 500 ms or greater), and greater pupil sizes than unsuccessful problem-solvers. In addition, we predicted that participants would demonstrate more gaze aversions, long fixations, and spontaneous reflections while solving difficult problems. However, we expected no significant differences between groups or difficulty levels of problems in the gaze duration, which is recognized as an indicator that reflects the initial reading processing in English texts (Hyönä et al. 2003; Yang et al. 2018) and Chinese texts (Jian et al. 2013), instead of deliberative thinking. We also examine the

effect of interactions between groups and problem difficulties on the four categories of eye-movement indicators.

Second, we investigate whether a combination of eye-movement indicators can roughly distinguish between successful and unsuccessful problem-solvers. If the patterns and quantity of eye-movement behaviors can adequately reflect individuals' processes of successfully or unsuccessfully solving problems at different levels of difficulty, then we should be able to predict individuals' problem-solving performance based on their visual behaviors.

Method

Participants and design

The valid participants were 69 sixth-grade students (aged 11–12) from two elite schools in Taipei. Prior to conducting this experiment, 179 sixth graders were assessed by the Reading Comprehension Screening Test (Ko 1999). Participants with normal reading ability who had parental consent for the eye-tracking experiment were invited to participate. All participants had already learned the testing material, two-step compare word problems, in their math classes over the preceding year. The eye-movement data of all participants were included since their valid ratio was above 85%, indicating good data quality. A mixed factorial design consisting of between- and within-subjects factors was adopted. The between-subjects factor was groups. Median splits on testing material scores defined the boundaries between two groups: successful ($n = 35$) and unsuccessful ($n = 34$) problem solvers. The median score on the test material was 12 out of a total score of 16 ($M = 10.84$, $SD = 3.26$). The within-subjects factors were related to the testing material, including the consistency of semantic structures and versions. All participants solved the testing material, which included 16 target problems generated with a 2×2 within-subjects design, so all comparisons involving problem types are within-subjects comparisons.

Materials

This study's testing material contained two-step compare word problems on various topics, including comparison, baseline amount, and cost and selling prices, from sixth graders' mathematics textbooks. These problems were adapted from Hegarty et al. (1992) and are similar to those used in participants' math classes. Four types of problems were used in this study, 16 problems in total. The types of problems were classified based on their semantic structures (consistent or inconsistent) and versions (text-only or illustrated). Each problem consisted of four sentences, each on a single line, presenting the background information, the relationship between known and unknown quantities, the known quantity, and a question. The consistent and inconsistent problems differed according to the consistency of their semantic structures with the arithmetic operations, presented in the third sentence, and illustrations were also adapted according to the texts. Moreover, the texts of problems in the text-only version were the same as in the illustrated version. The test format contained true/false questions with scores ranging from 0 to 16. Two experienced math teachers were invited to adjust the testing materials in this study to ensure that the wording and content of the problems were appropriate for the literacy level of elementary students. All testing materials were also adapted according

to the results of a preliminary study involving several elementary students. Furthermore, a pilot study was conducted to clarify the effects of consistency and problem version on 338 sixth-grade students' performance. Results indicated that the accuracy of consistent-illustrated, consistent-text, inconsistent-illustrated, and inconsistent-text problems was .84, .79, .61, and .57, respectively. The accuracy was higher for consistent problems than inconsistent ones, $F(1, 337) = 239.05, p < .001, \eta = .42$, accuracy was higher for illustrated problems than text-only ones, $F(1, 337) = 29.80, p < .001, \eta = .08$, and there was no interaction effect between consistency and problem version.

The reading comprehension screening test (Ko 1999) was adopted to confirm students' reading comprehension ability. The test contained the two dimensions of local processing and text processing. The dimension of local processing includes searching for the meaning of words as well as forming propositions and propositional integration, which corresponds to the test type of multi-meaning words and propositional integration. In addition, the dimension of text processing includes comprehending and reasoning about the meaning of a text, which corresponds to the test types of sentence comprehension and brief text comprehension. The test comprised 30 multiple-choice questions with a total score ranging 0–30. Based on the reading test results, students with scores lower than 14 may be diagnosed as having reading difficulties. The internal consistency reliability coefficient of this test was .80, and the retest reliability coefficient was .85.

Apparatus

An EyeLink 1000 video-based eye tracker (SR Research Ltd., Mississauga, Ontario, Canada) with a temporal resolution of 1000 Hz was used. Participants were seated with their heads stabilized on a chinrest and with a distance of approximately 60 cm between the monitor that displayed the tasks (19 in. diagonal with a resolution of 1280×1024 pixels) and their eyes. The screen covered 43° of horizontal visual angle and 32° of vertical visual angle.

Procedures

Data collection took place in two sessions. In the first session, all participants were collectively administered the reading comprehension screening test for approximately 25 min. The second session involved individual experiments with the eye tracker, administered in a quiet room. After instructions were provided, the eye tracker was calibrated for each participant using a nine-point procedure. Next, the participant was instructed to speak their problem-solving plans aloud as they read two practice problems on the screen, with no calculation or answer required. The formal experimental process was the same as the practice process. The sequence of the 16 problems was randomized for each participant. To create a natural reading condition, each participant was allowed to complete the tasks at his or her own pace. On average, participants finished the entire eye-tracking experiment in about 30–45 min.

Results

Table 1 shows the means and standard deviations of successful/unsuccessful problem-solvers by problem difficulty in terms of different eye-movement indicators. Six 2×4 (group \times problem difficulty) ANOVA tests on different eye-movement indicators and four

Table 1 Means and standard deviations for different eye-movement indicators as a function of group and difficulty

Indicator	Group	Consistent				Inconsistent			
		Illustrated		Text-only		Illustrated		Text-only	
		<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)
RT – TFD	1	4.84	(2.69)	6.54	(4.33)	7.25	(4.24)	8.89	(5.56)
	2	4.77	(2.36)	5.97	(4.11)	6.40	(3.91)	6.70	(4.20)
TFD of blank	1	1.00	(0.85)	1.17	(1.57)	1.39	(1.10)	1.71	(1.75)
	2	0.79	(0.46)	0.93	(0.90)	0.88	(0.48)	1.21	(1.09)
NOF of blank	1	4.17	(3.09)	4.65	(5.37)	5.33	(3.56)	6.73	(6.81)
	2	3.36	(1.80)	3.68	(3.32)	3.93	(1.99)	4.46	(3.54)
NOF of 500 ms	1	4.78	(2.36)	6.33	(3.37)	6.89	(4.87)	8.37	(4.99)
	2	4.38	(2.21)	5.49	(2.96)	5.77	(2.54)	5.74	(2.38)
Pupil sizes (z-score)	1	-0.28	(0.55)	-0.07	(0.63)	0.11	(0.60)	0.23	(0.52)
	2	-0.25	(0.64)	0.12	(0.59)	-0.27	(0.54)	0.29	(0.57)
Gaze duration	1	0.27	(0.04)	0.27	(0.03)	0.26	(0.04)	0.27	(0.03)
	2	0.26	(0.03)	0.27	(0.03)	0.26	(0.04)	0.26	(0.03)

Note. Group 1 comprises successful problem solvers, and Group 2 comprises unsuccessful problem solvers

multiple regression analyses were conducted. Table 2 summarizes the effects of ANOVAs for the six indicators.

Effect of group and problem difficulty on RT – TFD, TFD of blank, and NOF of blank

A two-way ANOVA on RT – TFD showed a significant effect for difficulty, $F(3, 201) = 17.11$, $p < .001$, $\eta^2 = .203$, but not for group, $F(1, 67) = 1.27$, $p > .10$. In addition, the results indicated that interaction between difficulty and group was marginally significant, $F(3, 201) = 2.26$, $p = .083$, $\eta^2 = .033$. In addition, a simple main-effect analysis was conducted. For the successful group, a significant difference in RT – TFD was found between the problem difficulties, $F(3, 201) = 15.22$, $p < .001$. Post hoc comparison indicated that RT – TFD was significantly greater for inconsistent-text problems ($M = 8.89$) than consistent-text and consistent-illustrated

Table 2 Summary of the effects of ANOVAs for different eye-movement indicators

Indicator	Effect of group	Effect of difficulty	Interaction
RT – TFD		✓	† (.083)
TFD of blank	† (.079)	✓	
NOF of blank	† (.080)	✓	
NOF of 500 ms	† (.070)	✓	
Pupil sizes		✓	✓
Gaze duration			† (.074)

Note. “✓” indicates a significant effect, and “†” indicates a marginally significant effect with the p value in the parentheses

problems ($M = 6.53, 4.84; ps < .001$) and marginally greater for inconsistent-text problems than inconsistent-illustrated problems ($M = 7.25, p = .063$). Furthermore, RT – TFD was significantly greater for inconsistent-illustrated and consistent-text problems than consistent-illustrated problems ($ps < .05$). For the unsuccessful group, a significant difference in RT – TFD was also found between the problem difficulties, $F(3, 201) = 3.99, p = .009$. Post hoc comparison showed that RT – TFD was significantly greater for inconsistent-text, inconsistent-illustrated, and consistent-text problems ($M = 6.70, 6.40, 5.97$) than consistent-illustrated problems ($M = 4.77; ps < .05$). For inconsistent-text problems, a significant difference in RT – TFD was observed between the two groups, $F(1, 268) = 5.12, p = .024$. Post hoc comparison revealed that the successful problem-solvers ($M = 8.98$) demonstrated marginally significantly greater RT – TFD than the unsuccessful problem-solvers ($M = 6.70; p = .068$).

We also conducted a two-way ANOVA on TFD of blank. The main effect was marginally significant on group, $F(1, 67) = 3.18, p = .079, \eta^2 = .045$, and significant on difficulty, $F(3, 201) = 5.90, p = .001, \eta^2 = .081$, but no interaction effect existed between the two factors, $F(3, 201) = .69, p > .10$. For group, post hoc comparison indicated that the successful group ($M = 1.32$) demonstrated marginally significantly greater TFD of blank than the unsuccessful group ($M = 0.95$). For difficulty, post hoc comparison showed that TFD of blank was significantly greater for inconsistent-text problems ($M = 1.46$) than consistent-illustrated problems and consistent-text problems ($M = 0.89, 1.05; ps < .01$), but only marginally significantly greater than for inconsistent-illustrated problems ($M = 1.13; p = .064$). TFD of blank was also significantly greater for inconsistent-illustrated problems than consistent-illustrated problems ($p = .043$).

A two-way ANOVA conducted on NOF of blank showed that the main effect was marginally significant on group, $F(1, 67) = 3.16, p = .080, \eta^2 = .05$, and significant on difficulty, $F(3, 201) = 5.54, p = .001, \eta^2 = .08$, but no interaction effect existed between the two factors, $F(3, 201) = .96, p > .10$. For group, post hoc comparison indicated that the successful problem-solvers ($M = 5.22$) demonstrated marginally significantly greater NOF of blank than unsuccessful problem-solvers ($M = 3.86$). For problem difficulty, post hoc comparison showed that NOF of blank was significantly greater for inconsistent-text problems ($M = 5.58$) than consistent-illustrated problems and consistent-text problems ($M = 3.76, 4.16; ps < .01$). NOF of blank was also significantly greater for inconsistent-illustrated problems than consistent-illustrated problems ($p = .034$).

Effect of group and problem difficulty on NOF of 500 ms

A two-way ANOVA conducted on NOF of 500 ms showed a significant effect for difficulty, $F(3, 201) = 16.72, p < .001, \eta^2 = .20$, and a marginally significant effect for group, $F(1, 67) = 3.36, p = .07, \eta^2 = .05$, as well as a significant interaction effect, $F(3, 201) = 3.64, p = .014, \eta^2 = .05$. A simple main-effect analysis was then conducted. For the successful group, a significant difference in NOF of 500 ms was found between problem difficulties, $F(3, 201) = 16.79, p < .001$. Post hoc comparison showed that NOF of 500 ms was significantly greater for inconsistent-text, consistent-text, and inconsistent-illustrated problems ($M = 8.37, 6.89, 6.33$) than for consistent-illustrated problems ($M = 4.78; ps < .001$). For the unsuccessful group, a significant difference in NOF of 500 ms was also found between problem difficulties, $F(3, 201) = 3.37, p = .022$. Post hoc comparison showed that NOF of 500 ms was significantly greater for inconsistent-text, inconsistent-illustrated, and consistent-text problems ($M = 5.74, 5.49, 5.77$) than for

consistent-illustrated problems ($M=4.38$; $ps < .001$). For inconsistent-text problems, a significant difference in NOF of 500 ms was observed between the two groups, $F(1, 268) = 10.54$, $p = .001$. Post hoc comparison showed that the successful problem-solvers ($M=8.37$) demonstrated significantly greater NOF of 500 ms than the unsuccessful problem-solvers ($M=5.74$; $p = .007$).

Effect of group and problem difficulty on pupil sizes

Given the tremendous individual differences in pupil sizes, each participant's pupil size was standardized. We subtracted the average score in the four conditions from the average score in a particular condition, and then divided it by the standard deviation of one's scores in the four conditions. We conducted a two-way ANOVA on the z -score of pupil sizes to explore the effect of group and problem difficulty on pupil sizes. A significant effect was observed for difficulty, $F(3, 201) = 7.60$, $p < .001$, $\eta^2 = .10$, but not for group, $F(1, 67) = 0.93$, $p = .34$. The interaction between the two factors was marginally significant, $F(3, 201) = 2.35$, $p = .074$, $\eta^2 = .03$. A simple main-effect analysis was then conducted. For the successful group, a significant difference in the z -scores of pupil sizes was found between problem difficulties, $F(3, 201) = 3.87$, $p = .001$. Post hoc comparison showed that the z -scores of pupil sizes were significantly greater for inconsistent-text problems ($M=0.23$) than for consistent-text and consistent-illustrated problems ($M = -0.07, -0.28$; $ps < .01$). Furthermore, the z -scores of pupil sizes were significantly greater for inconsistent-illustrated problems ($M=0.11$) than for consistent-illustrated problems ($p < .001$). For the unsuccessful group, a significant difference in the z -scores of pupil sizes was also found between problem difficulties, $F(3, 201) = 6.10$, $p = .001$. Post hoc comparison showed that the z -score of pupil sizes was significantly greater for inconsistent-text problems ($M=0.29$) than for inconsistent-illustrated and consistent-illustrated problems ($M = -0.27, -0.25$; $ps < .01$). Additionally, the z -scores of pupil sizes were significantly greater for consistent-text problems ($M=0.12$) than for consistent-illustrated problems ($p = .055$), but only marginally significantly greater than for inconsistent-illustrated problems ($p = .03$). For inconsistent-illustrated problems, a significant difference in the z -scores of pupil sizes was observed between the two groups, $F(1, 268) = 7.19$, $p = .008$. Post hoc comparison showed that the successful group ($M = 0.11$) demonstrated marginally significantly greater z -scores of pupil sizes than did the unsuccessful group ($M = -0.27$; $p = .008$).

Effect of group and problem difficulty on gaze duration

A two-way ANOVA conducted on gaze duration showed no main effect on group, $F(1, 67) = 0.32$, $p > .05$, or difficulty, $F(1, 201) = 1.12$, $p > .05$, nor did interaction effects exist between the two factors, $F(3, 201) = 0.87$, $p > .05$.

Figure 1 shows a successful problem-solver's eye-movement behaviors and indicator values while solving the easiest (consistent-illustrated) and hardest (inconsistent-text) types of problems. In the easiest problem condition, the problem-solver inspected the text and illustration evenly with only a little fixation on blank areas, few long fixations, and limited spontaneous reflections (Fig. 1a). In contrast, in the hardest problem condition, the problem-solver demonstrated much more fixation on areas without stimuli, more long fixations, and many spontaneous reflections (Fig. 1b).

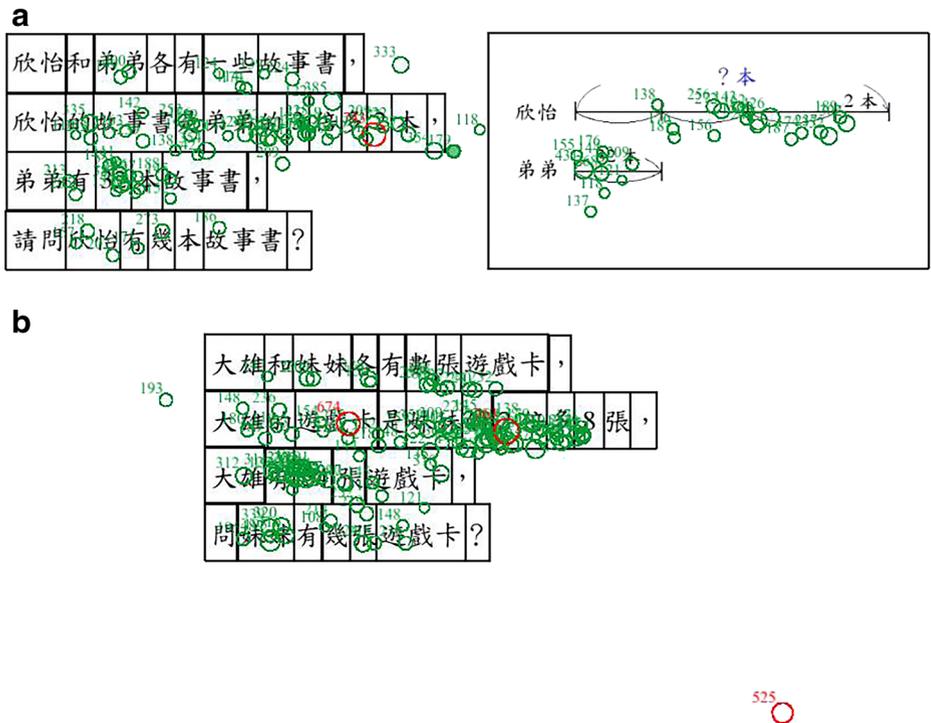


Fig. 1 The eye-movement behaviors of a successful problem solver in **(a)** a consistent-illustrated problem and **(b)** an inconsistent-text problem. The fixation time is represented by the size of the green circles, and the fixations over 500 ms are highlighted with red circles. The visual behaviors in **(a)** and **(b)** demonstrated great disparity in the values of RT – TFD (ms) (a: 5.30; b: 7.70), TFD of blank (ms) (a: 0.84; b: 1.98), NOF of blank (count) (a: 4; b: 9), NOF of 500 ms (count) (a: 1; b: 3), and z-score of pupil sizes (a: -0.08; b: 1.42)

Multiple regression analysis of problem-solving performance

To investigate how eye-movement indicators can predict problem-solvers' performance (i.e., success or failure) in the four types of problems, multiple regression analyses were conducted. Initially, all eye-movement indicators were included as predictors in a regression model. Although 49.9% of the variance in participants' performance was predicted by these indicators, severe multicollinearity occurred. The correlation coefficients indicated that some of the predictors were significantly correlated, even above the threshold of 0.70 recommended by Tabachnik and Fidell (2001; see Appendix for details). Thus, we reselected the predictors to include at least one indicator for each category and to avoid multicollinearity. Because the similarity of properties and correlation among RT – TFD, TFD of blank, and NOF of blank are high, these indicators were allocated in different models. Three regression models with three different sets of predictors were compared: (1) NOF of 500 ms, pupil sizes, and RT – TFD; (2) NOF of 500 ms, pupil sizes, and TFD of blank; and (3) NOF of 500 ms, pupil sizes, and NOF of blank. Table 3 presents the results of the individual multiple regression analyses. The three models explained 42%, 42%, and 44% of the total variance in participants' performance, $F(12, 56) = 3.62, 3.33, 3.45, ps = .001$. According to Cohen's f^2 , the effect sizes of three models were large (0.78, 0.71, and 0.74, respectively). Comparison of these three models with the initial model showed that the problem of multicollinearity was resolved. Furthermore, there were no significant differences among the R^2 values of these models, indicating that no

redundant variables were included. As shown in Table 3, model I demonstrated the highest R^2 , and pupil sizes^c, NOF of 500 ms^d, RT – TFD^d, and pupil sizes^a were the predictors with the largest effect on participants' performance, $t(68) = 3.18, 2.53, 2.43, 2.53, p = .002, .014, .014, .032$; $\beta = .62, .50, .45, -.40$. The results indicate that participants perform better on difficult tasks (i.e., inconsistent-text problems) if they tend to demonstrate greater pupil size, have greater numbers of long fixations, and spend more time fixated on blank areas. In contrast, participants perform better on easy tasks (i.e., consistent-illustrated problems) if they demonstrate smaller pupil dilation. Given this, the estimated multiple regression equation becomes:

$$\begin{aligned}
 Y = & 0.53 + 0.01\text{NOF of 500 ms}^a - 0.02\text{NOF of 500 ms}^b + 0.00\text{NOF of 500 ms}^c \\
 & + 0.03\text{NOF of 500 ms}^d - 0.14\text{pupil sizes}^a - 0.05\text{pupil sizes}^b \\
 & + 0.21\text{pupil sizes}^c - 0.01\text{pupil sizes}^d - 0.01\text{RT-TFD}^a - 0.01\text{RT-TFD}^b - 0.01\text{RT-TFD}^c \\
 & + 0.02\text{RT-TFD}^d.
 \end{aligned}$$

Additionally, we allocated participants to successful or unsuccessful groups by their expected performance values generated from the multiple regression equation, and then examined. Additionally, we allocated participants to successful or unsuccessful groups by their expected performance values generated from the multiple regression equations, and then examined the fit between the groups divided by the multiple regression equation and the groups divided by their scores on the two-step compare word problems (Table 4). The overall hit rate was 73.91%, indicating that this model could approximately predict problem-solvers' performance with the indicators of NOF of 500 ms, pupil sizes, and RT – TFD. Separate from this, when we included the t -scores of the reading comprehension screening test as a predictor (Tables 3, model IV), the overall hit rate was as much as 79.71%, with $R^2 = .52, F(12, 56) = 4.52, p = .001$, and Cohen's $f^2 = 1.07$, indicating the adequacy of the model (Table 4).

Discussion and conclusion

To address the first research question, this study identified five indicators of deliberation in three categories to reflect the difficulty of problems and the visual behaviors of successful and unsuccessful problem-solvers. The first category was gaze aversion, which included RT – TFD, TFD of blank, and NOF of blank. There was a marginally significant interaction between problem difficulty and participant group on RT – TFD. For TFD of blank and NOF of blank, significant main effects were observed for problem difficulty, but only marginally significant main effects were found for group. The second category was long fixations, indicated by NOF of 500 ms, and results indicated a significant interaction between problem difficulty and participant group. The third category was spontaneous reflection, indicated by pupil sizes, and the results showed a significant main effect only for problem difficulty. The indicator of gaze duration, which reflected preliminary lexical processing, showed no main effect and no interaction effect on difficulty and group, as predicted.

In general, as shown in Table 2, the main effects of problem difficulty were significant for all five eye-movement indicators, but the main effects of group demonstrated only marginal

Table 3 Multiple regression analyses for variables predicting participants' performances.

	Model I		Model II		Model III		Model IV	
	<i>B</i> (<i>SE</i>)	β						
(Constant)	0.53 (0.07)		0.51 (0.06)		0.52 (0.07)		-0.28 (0.28)	
NOF of 500 ms ^a	0.01 (0.02)	.12	0.02 (0.01)	.16	0.01 (0.02)	.13	0.00 (0.01)	.01
NOF of 500 ms ^b	-0.02 (0.01)	-.28	-0.02 (0.01)	-.37	-0.02 (0.01)	-.36	-0.02 (0.01)	-.24
NOF of 500 ms ^c	0.00 (0.01)	-.08	-0.01 (0.01)	-.15	-0.01 (0.01)	-.13	-0.01 (0.01)	-.19
NOF of 500 ms ^d	0.03 (0.01)	.50*	0.03 (0.01)	.62**	0.03 (0.01)	.62**	0.03 (0.01)	.69***
Pupil sizes ^a	-0.14 (0.06)	-.40*	-0.15 (0.06)	-.44*	-0.15 (0.06)	-.43*	-0.11 (0.06)	-.33
Pupil sizes ^b	0.05 (0.06)	.15	0.06 (0.06)	.18	0.06 (0.06)	.18	0.04 (0.06)	.11
Pupil sizes ^c	0.21 (0.06)	.62*	0.20 (0.06)	.57**	0.20 (0.06)	.58**	0.19 (0.06)	.55*
Pupil sizes ^d	-0.01 (0.06)	-.02	-0.03 (0.06)	-.07	-0.02 (0.06)	-.05	-0.05 (0.06)	-.12
RT - TFD ^a	0.01 (0.02)	.08					0.01 (0.02)	.08
RT - TFD ^b	-0.01 (0.01)	-.26					-0.01 (0.01)	-.14
RT - TFD ^c	-0.01 (0.01)	-.18					-0.01 (0.01)	-.13
RT - TFD ^d	0.02 (0.01)	.45*					0.01 (0.01)	.27
TFD of blank ^a			0.04 (0.03)	.00				
TFD of blank ^b			-0.02 (0.03)	-.13				
TFD of blank ^c			0.01 (0.03)	.05				
TFD of blank ^d			0.00 (0.05)	.31				
NOF of blank ^a					0.01 (0.01)	.01		
NOF of blank ^b					-0.01 (0.01)	-.19		
NOF of blank ^c					0.00 (0.01)	-.01		
NOF of blank ^d					0.00 (0.01)	.38		
Reading comprehension screening test							.013(.004)	.35*

Note. *B* = Unstandardized regression coefficients; *SE* = Standard error; β = Standardized regression coefficients
^a consistent-illustrated problems; ^b consistent-text problems; ^c inconsistent-illustrated problems; ^d inconsistent-text problems

* $p < .05$. ** $p < .01$. *** $p < .001$

significance for three indicators. This might be because the participants in this study, who were from elite schools, were all quite good at math, and thus the differences between the eye movements of successful and unsuccessful subjects were not that obvious. The interaction

Table 4 The classification accuracy of model I and IV

Model	Real performances of participants	
	Success	Failure
Expected values (model I)		
Success (<i>n</i>)	25	9
Failure (<i>n</i>)	9	26
Hit rate (%)	73.52	74.28
Expected values (model IV)		
Success (<i>n</i>)	27	7
Failure (<i>n</i>)	7	28
Hit rate (%)	79.41	80.00

effects of problem difficulty and group were significant or marginal for three indicators, and these results suggest that high ability solvers were more sensitive to problem structure and regulated their cognitive resources efficiently.

Our results of gaze aversion indicators are similar to the findings of Doherty-Sneddon et al. (2002) and Glenberg et al. (1998). This also expands on the eye-mind hypothesis (Just and Carpenter 1980) – both gaze distribution and gaze aversion could reflect an individual's attention. While processing demanding mental tasks, most people, especially those with high abilities, tend to avoid external stimuli to engage in the cognitive effort. We found that the successful participants demonstrated significantly greater NOF of 500 ms than unsuccessful ones while processing difficult problems. It indicated that the high ability individuals might be aware of problem's difficulty, so they fixated for longer durations to process the cognitive task deeply and then acquire higher accuracy. This is in keeping with previous research (Horstmann et al. 2009; Knoblich et al. 2001). For spontaneous reflection, we found that the pupillary sizes of successful problem-solvers were marginally larger than those of unsuccessful individuals, especially when completing a difficult task. This is consistent with previous findings that individuals demonstrated more pupillary dilation when processing tasks with higher workload level (Klingner et al. 2011; Siegle et al. 2008), reflecting their sustained processing of information. However, there were two eye-movement indicators that showed only main effects but no interaction effect between difficulty and group. The alternative explanation is that the solvers could regulate their engagement strategies for task completion.

Deliberative thinking of problem-solving is different from mind wandering which is related to the “looking at nothing” behavior. Mind wandering is a state of focus on internal information where one's attention has switched to private thoughts (Baird et al. 2012; Salvi and Bowden 2016). The absence of external demands increases the occurrence of mind wandering (e.g., daydreaming, mental simulations, looking for ideas, thinking creatively). Deliberative thinking for problem-solving involves allocating attention separately for external and internal information. However, this study cannot distinguish deliberative thinking from mind wandering only based on eye-movement indicators of gaze aversion. Although the effects of ability and problem difficulty on RT-TFD, TFD of blank, and NOF of blank were consistent with the hypothesis, these indicators of gaze aversion need to collaborate with other indicators to successfully predict the solver's performance.

The calibration drift (i.e., decreases in tracking accuracy over time) has potential to act as a confounder for eye movement indicators of gaze aversion because fixations on the visual stimulus might be categorized as gaze aversions incorrectly. However, the drift should not systematically covariate with participants' ability or problems' difficulty; therefore, it is not a confounding variable in this study.

Regarding pupil size, one might argue that sensitivity of pupil diameters to small physical stimulation is a potential confounder. For example, a sensory/perceptual difference with learning materials might also cause the change of pupil diameters. For this study, we developed 16 two-step compare word problems as the testing material, and each problem contained four sentences with similar structure ensuring sensory

uniformity. Thus, the significant difference of pupil sizes in this study is more likely to be the effect of the problem difficulty and the group.

To address the second research question, individuals' problem-solving performance on tasks with varying difficulties was predicted and examined. The results demonstrated that individuals' eye movements can successfully predict their problem-solving performance in arithmetic word problem tasks. Specifically, the model incorporating RT – TFD, NOF of 500 ms, and pupil sizes can best predict participants' performance with an overall hit rate of 74%. The overall hit rate of the model increased to nearly 80% when reading comprehension screening score was adopted as another predictor. This shows that individuals perform better when solving difficult problems if they spend more time fixating on blank areas, exhibit greater numbers of long fixations, and demonstrate greater pupil size. Meanwhile, individuals perform better when solving easy problems if they demonstrate smaller pupil size. One possible explanation for this is that individuals may not need a large number of external stimuli when thinking and integrating information in the mind. Additionally, the results also clarify that the successful problem-solvers could distinguish between the difficult and easy problems, and then take appropriate action to solve the problems accurately according to their varying difficulty.

As the studies conducted by Yang et al. (2018) and Liu et al. (2019), educators should be able to identify the characteristics of students and the patterns of eye movements associated with better learning outcomes first, and then make use of these findings to develop an adapted learning systems. Based on the evidence in this study, researchers can integrate eye-movement indicators of deliberation and diagnostic assessment into a real-time leaning system or digital scaffolding to provide students with individualized learning materials and feedback. For instance, if a student was found to fixate only on the stimulus area, exhibit few long fixations, and demonstrate smaller pupil size while solving difficult tasks, the system could offer real-time prompts or feedback to remind the student to reread and consider the materials thoroughly, decrease the difficulty of tasks by providing an easier version with illustrations, or allow educators or advisors to intervene and assist in students' learning directly. This individualized learning system would guide students in understanding how to solve arithmetic word problems step by step, helping them to become successful problem-solvers.

Furthermore, this study used an in vitro methodology to investigate learner's eye movement in a standard psychological lab setting. Thus, future research should adopt an in vivo methodology to explore the learner's deliberative thinking in a real-world context, and consequently open more avenues for new in vitro research. Broadly, this study makes it feasible to design an eye-tracking-embedded learning system incorporating the critical indicators to monitor and facilitate students' arithmetic problem-solving performance.

Funding information This work was financially supported by the Ministry of Science and Technology, Taiwan under grant number MOST 104-2511-S-003-013-MY3, MOST 108-2511-H-003-014-MY3, MOST 108-2636-H-003-003-, and by the "Institute for Research Excellence in Learning Sciences" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Appendix

construct	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1 RT - TFD ^a	—																			
2 RT - TFD ^b	.61**	—																		
3 RT - TFD ^c	.56**	.68**	—																	
4 RT - TFD ^d	.74**	.69**	.62**	—																
5 TFD of blank ^a	.24*	.35**	.30*	.30*	—															
6 TFD of blank ^b	.55**	.32**	.43**	.12	.71**	—														
7 TFD of blank ^c	.31**	.06	.31**	.34**	.21	.34**	—													
8 TFD of blank ^d	.41**	.41**	.63**	.13	.72**	.75**	.26*	—												
9 NOF of blank ^a	.29*	.38**	.35**	.01	.70**	.76**	.97**	.27*	—											
10 NOF of blank ^b	.61**	.32**	.47**	.17	.98**	.72**	.65**	.26*	.67**	—										
11 NOF of blank ^c	.44**	.13	.40**	.42**	.31*	.42**	.24	.92**	.39**	.47**	—									
12 NOF of blank ^d	.47**	.41**	.68**	.17	.74**	.74**	.74**	.38**	.75**	.75**	.27*	—								
13 NOF of 500 ms ^a	.03	.01	-.02	-.17	.06	.05	.02	-.10	.01	.03	-.02	-.16	—							
14 NOF of 500 ms ^b	.15	-.07	.08	-.13	.17	.10	-.01	.07	.14	.09	-.03	.08	.66**	—						
15 NOF of 500 ms ^c	-.01	-.12	.03	.07	.07	-.01	-.02	-.06	.14	-.04	-.04	-.06	.08	.56**	—					
16 NOF of 500 ms ^d	.04	-.09	.15	-.10	.11	.16	.08	.09	.08	.13	.05	.07	.74**	.59**	.68**	—				
17 Pupil sizes ^a	-.15	-.09	-.20	-.26*	-.05	.05	.17	-.08	-.07	.02	.16	-.12	-.32**	-.05	-.15	-.25*	—			
18 Pupil sizes ^b	.08	.03	.10	.09	.01	-.10	-.20	-.06	.05	-.07	-.16	.00	.18	.01	.00	.10	-.70**	—		
19 Pupil sizes ^c	-.03	-.15	-.11	.06	-.08	-.10	.02	.12	-.08	-.07	.01	.10	-.24*	-.16	-.12	-.06	-.72**	.51**	—	
20 Pupil sizes ^d	.11	.18	.24*	.14	.13	.20	-.02	.06	.15	.18	-.02	.08	.21	.17	.23	.17	.51**	-.63**	-.63**	—

Note. ^aConsistent-illustrated problems; ^bConsistent-text problems; ^cInconsistent-illustrated problems; ^dInconsistent-text problems

* $p < .05$. ** $p < .01$. *** $p < .001$

References

- Abeles, D., & Yuval-Greenberg, S. (2017). Just look away: Gaze aversions as an overt attentional disengagement mechanism. *Cognition*, *168*, 99–109. <https://doi.org/10.1016/j.cognition.2017.06.021>.
- Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science*, *23*(10), 1117–1122. <https://doi.org/10.1177/0956797612446024>.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–742. <https://doi.org/10.1017/S0140525X97001611>.
- Belenky, D. M., & Schalk, L. (2014). The effects of idealized and grounded materials on learning, transfer, and interest: An organizing framework for categorizing external knowledge representations. *Educational Psychology Review*, *26*(1), 27–50. <https://doi.org/10.1007/s10648-014-9251-9>.
- Booth, J. L., & Koedinger, K. R. (2012). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. *British Journal of Educational Psychology*, *82*(3), 492–511. <https://doi.org/10.1111/j.2044-8279.2011.02041.x>.
- Chiou, G. L., Hsu, C. Y., & Tsai, M. J. (2019). Exploring how students interact with guidance in a physics simulation: Evidence from eye-movement and log data analyses. *Interactive Learning Environments*. <https://doi.org/10.1080/10494820.2019.1664596>.
- Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., & Doyle, C. (2002). Development of gaze aversion as disengagement from visual information. *Developmental Psychology*, *38*(3), 438–445. <https://doi.org/10.1037//0012-1649.38.3.438>.
- Ehrlichman, H., & Micic, D. (2012). Why do people move their eyes when they think? *Current Directions in Psychological Science*, *21*(2), 96–100. <https://doi.org/10.1177/0963721412436810>.
- Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*(4), 651–658. <https://doi.org/10.3758/BF03211385>.
- Gredebäck, G., & Melinder, A. M. D. (2010). Infants understanding of everyday social interactions: A dual process account. *Cognition*, *114*, 197–206. <https://doi.org/10.3758/BF03211385>.
- Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology*, *84*(1), 76–84. <https://doi.org/10.1037/0022-0663.84.1.76>.
- Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, *87*(1), 18–32. <https://doi.org/10.1037//0022-0663.87.1.18>.
- Horstmann, N., Ahlgrimm, A., & Glöckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. *Judgment and Decision making*, *4*, 335–354. <https://doi.org/10.2139/ssrn.1393729>.
- Hyönä, J., Lorch, R. F., & Rinck, M. (2003). Eye movement measures to study global text processing. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye cognitive and applied aspects of eye movement research* (pp. 313–334). Elsevier Science BV.
- Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science*, *12*, 670–679. <https://doi.org/10.1111/j.1467-7687.2008.00805.x>.
- Jian, Y.-C., Chen, M. -L., & Ko, H. -W. (2013). Context effects in processing of Chinese academic words: An eye-tracking investigation. *Reading Research Quarterly*, *48*(4), 403–413. <https://doi.org/10.1002/rrq.56>.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–354. <https://doi.org/10.1037//0033-295X.87.4.329>.
- Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, *47*(2), 310–339. <https://doi.org/10.1037/h0078820>.
- Kahneman, D. (1973). *Attention and effort*. Prentice Hall.
- Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*(3), 323–332. <https://doi.org/10.1111/j.1469-8986.2010.01069.x>.
- Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, *29*(7), 1000–1009. <https://doi.org/10.3758/BF03195762>.
- Ko, H. W. (1999). Reading comprehension-screening test [in Chinese]. *Psychological Testing*, *46*, 1–11.
- Krstić, K., Šoškić, A., Ković, V., & Holmqvist, K. (2018). All good readers are the same, but every low-skilled reader is different: An eye-tracking study using PISA data. *European Journal of Psychology of Education*, *33*(3), 511–541. <https://doi.org/10.1007/s10212-018-0382-0>.
- Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. <https://doi.org/10.1177/1745691611427305>.

- Lin, Y.-T., Wu, C.-C., Hou, T.-Y., Lin, Y.-C., Yang, F.-Y., & Chang, C.-H. (2016). Tracking students' cognitive processes during program debugging-an eye-movement approach. *IEEE Transactions on Education*, 59(3), 175–186. <https://doi.org/10.1109/TE.2015.2487341>.
- Liu, T. S. W., Liu, Y. T., & Chen, C. Y. D. (2019). Meaningfulness is in the eye of the reader: Eye-tracking insights of L2 learners reading e-books and their pedagogical implications. *Interactive Learning Environments*, 27(2), 181–199. <https://doi.org/10.1080/10494820.2018.1451901>.
- Lu, H. I., Chan, Y. C., & Chen, H. C. (2018). Ambiguity and inference processing in verbal jokes: Analyses of eye movement (in Chinese). *Bulletin of Educational Psychology*, 50(4), 589–609. [https://doi.org/10.6251/BEP.201906_50\(4\).0002](https://doi.org/10.6251/BEP.201906_50(4).0002).
- Micic, D., Ehrlichman, H., & Chen, R. (2010). Why do we move our eyes while trying to remember? The relationship between non-visual gaze patterns and memory. *Brain & Cognition*, 74(3), 210–224. <https://doi.org/10.1016/j.bandc.2010.07.014>.
- Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8.
- Phelps, F. G., Doherty-Sneddon, G., & Warnock, H. (2006). Helping children think: Gaze aversion and teaching. *British Journal of Developmental Psychology*, 24(3), 577–588. <https://doi.org/10.1348/026151005X49872>.
- Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, 47, 1–10. <https://doi.org/10.1111/j.1469-8986.2009.00947.x>.
- Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition & Instruction*, 5, 49–101. https://doi.org/10.1207/s1532690xci0501_2.
- Salvi, C., & Bowden, E. M. (2016). Looking for creativity: Where do we look when we look for new ideas? *Frontiers in Psychology*, 7, 161. <https://doi.org/10.3389/fpsyg.2016.00161>.
- Schumacher, R. F., & Fuchs, L. S. (2012). Does understanding relational terminology mediate effects of intervention on compare word problems? *Journal of Experimental Child Psychology*, 111(4), 607–628. <https://doi.org/10.1016/j.jecp.2011.12.001>.
- Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, 45(5), 679–687. <https://doi.org/10.1111/j.1469-8986.2008.00681.x>.
- Tabachnick, B. G. & Fidell, L. S. (2001). *Using Multivariate Statistics*. Boston: Allyn and Bacon.
- Van der Schoot, M., Arkema, A. H. B., Horsley, T. M., & Van Lieshout, E. C. D. M. (2009). The consistency effect depends on markedness in less successful but not successful problem solvers: An eye movement study in primary school children. *Contemporary Educational Psychology*, 34, 58–66. <https://doi.org/10.1016/j.cedpsych.2008.07.002>.
- Walcher, S., Kömer, C., & Benedek, M. (2017). Looking for ideas: Eye behavior during goal-directed internally focused cognition. *Consciousness and Cognition*, 53, 165–175. <https://doi.org/10.1016/j.concog.2017.06.009>.
- Wang, C. Y., Tsai, M. J., & Tsai, C. C. (2016). Multimedia recipe reading: Predicting learning outcomes and diagnosing cooking interest using eye-tracking measures. *Computers in Human Behavior*, 62, 9–18. <https://doi.org/10.1016/j.chb.2016.03.064>.
- Yang, F.-Y., Tsai, M.-J., Chiou, G.-L., Lee, S. W.-Y., Chang, C.-C., & Chen, L.-L. (2018). Instructional suggestions supporting science learning in digital environments based on a review of eye tracking studies. *Educational Technology & Society*, 21(2), 28–45.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Chao-Jung Wu¹ · Chia-Yu Liu¹ · Chung-Hsuan Yang¹ · Yu-Cin Jian¹

¹ Department of Educational Psychology and Counseling/Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, No. 162, Sec. 1, Heping E. Rd, Taipei 106, Taiwan