International Journal of Research & Method in Education
Vol. 29, No. 2, October 2006, pp. 185–208

An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments

Robert H. Tai,*[a], John F. Loehr[b] and Frederick J. Brigham[c]
[a]*University of Virginia, USA;* [b]*Chicago Public Schools, USA;* [c]*George Mason University, USA*

This pilot study investigated the capacity of eye-gaze tracking to identify differences in problem-solving behaviours within a group of individuals who possessed varying degrees of knowledge and expertise in three disciplines of science (biology, chemistry and physics). The six participants, all pre-service science teachers, completed an 18-item multiple-choice science assessment while having their eye-gaze tracked and recorded. Analysis of the data revealed differences in eye-gaze behaviour across different disciplines and similarities among participants with similar science backgrounds. This manuscript discusses various issues in eye-gaze tracking data analysis and suggests some analytical techniques for addressing these issues. The findings suggest that eye-gaze tracking may potentially be a useful approach to furthering our understanding of students' problem-solving behaviours.

Introduction

The purpose of this study is to explore the use of eye-gaze tracking devices to evaluate the behaviours of individuals as they solve standardized science assessment problems. In this exploratory study, individuals with varying degrees of expertise are asked to solve several science problems as their eye-gazes are recorded. The recorded eye-gaze information includes location of eye-gaze fixation on a computer screen, duration of fixation, the path of eye movements (saccades) and duration between fixations. We hypothesize that individuals with greater levels of expertise in a given domain would demonstrate patterns of eye-movements quantifiably different from individuals with less expertise in the domain. Should quantifiable differences between experts and non-experts exist, data documenting differences in the behaviours between these

*Corresponding author. Curry School of Education, University of Virginia, 405 Emmet Street South, Charlottesville, VA 22904, USA. Email: rht6h@virginia.edu

ISSN 1743-727X (print)/ISSN 1743-7288 (online)/06/020185–24

DOI: 10.1080/17437270600891614

individuals may be useful in several different ways. First, these differences may yield some insight into the tacit knowledge experts possess to which novices may not be privy. Identification of these differences would be especially advantageous, given that experts often lack the ability to verbalize or even isolate their knowledge in a manner that allows others to understand and learn from them (Sternberg & Horvath, 1995). Second, this data may provide insights into the existence of common expert problem-solving behaviours. Do gaze patterns differ for an individual as his or her level of expertise vary? Do patterns common among expert problem-solvers, but uncommon among novices exist? Should 'expert patterns' be found, this technique will allow them to be isolated and studied.

Eye-gaze tracking devices typically collect information about the location and duration of an eye fixation within a specific area on a computer monitor. When objects such as words and pictures are shown on the display, an individual's eye-gaze may be tracked as they look at these words and pictures. In addition, the location and timing of mouse clicks can also be recorded. Based on the location of eye fixations and mouse clicks, inferences may be drawn regarding the activity and intent of an individual. With respect to the scope of this manuscript, the authors considered the relative novelty of eye-gaze research in science education and concluded that providing a point of entry into the use of eye-gaze tracking in the study of science learning assessment by limiting the scope of this manuscript may be more useful than an exhaustive and complex review of eye-gaze tracking research.

This manuscript examines eye-gaze tracking as a means of identifying whether individuals with known differences in expertise differ in their eye-gaze patterns in a consistent fashion. Beginning with literature relevant to the understanding and interpretation of the data collected from this exploratory research study, the manuscript continues with design and details of the study. The researchers then discuss the form and function of the data collected from eye-gaze tracking. Next, a discussion of the collected data and various approaches to data analysis are investigated and explained. In particular, the researchers propose a means of displaying eye-gaze tracking data for analysis that the researchers have termed *zone graphs*. Finally, the manuscript concludes with a discussion of possibilities for further research into the uses of eye-gaze tracking in assessment.

Literature review

Eye-gaze tracking has been used for many years in the study of reading. In a comprehensive literature review, Rayner (1998) cites over 800 research articles spanning 20 years. However, the application of eye-gaze technology in terms of assessing the knowledge and expertise of individuals has been largely ignored. In fact, Pellegrino *et al.*, in their book on educational assessment, *Knowing what students know*, state:

> Eye-movement tracking, a specialized technique for studying reaction times and other key behaviors, has received virtually no attention in the assessment literature. By using what is now relatively inexpensive equipment capable of detecting the direction of a person's gaze while he or she is engaged in a task, psychologists can gather data about the sequence and

> duration of eye fixations. … Such analyses can yield insights into differences between experts and novices in a range of domains. (Pellegrino *et al.*, 2001, p. 98).

An important purpose of science education is the development of the ability to access and apply scientific knowledge in an appropriate manner to solve problems (National Research Council, 1995). When this ability stretches beyond the experiences of the learner to include the ability to solve new, but related problems, i.e., problems within a given domain, one may describe this behaviour as expertise, and in the case of science, scientific expertise (Glaser & Chi, 1988). When a learner begins to make connections and linkages that further facilitate their problem-solving ability, we may describe this behaviour as a higher level of expertise. The honing of problem-solving skills and development of tacit knowledge associated with a particular area of science may be viewed as a continuum of scientific expertise, with higher and lower levels. The development of techniques to gauge various levels of scientific expertise could prove useful in providing insight into the learning and teaching of science. For example, the ability to measure differences in developing expertise in secondary science students as they progress through a high school science course may provide valuable information to an instructor as he or she plans lessons and assignments. In this review, the authors will discuss the previous research in three research areas: expertise, eye-gaze tracking and assessment.

Expertise

While some research has shown that differences in problem-solving can be related to test-taking skills and strategies (Scruggs *et al.*, 1986; Mastropieri & Scruggs, 1999), of particular interest to educators are the studies that have shown differences in problem-solving behaviour related to individuals varying levels of expertise within a content domain (Chase & Simon, 1973; Chi *et al.*, 1981; Dillon, 1985b). Research indicates that the differences in problem-solving approaches used by two different test takers likely depends on their level of knowledge and experience. The central question of this research study lies within the identification and analysis of eye-gaze behaviours patterns and the association of these behaviour patterns with individuals of varying expertise.

Previous research has shown identifiable differences between the behaviours of novices and experts. With regard to their content knowledge, experts do possess a great deal of content knowledge, however the main difference from novices in this regard is that experts organize their knowledge in ways that are related to deep understanding of the subject matter. When it comes to problem-solving, experts notice features and patterns of information ignored by novices and are more sensitive to the context of a given situation. Unfortunately, experts are not always able to communicate to others all that they know (Chase & Simon, 1973; Chi *et al.*, 1981; Pellgrino *et al.*, 2001; Sternberg & Horvath, 1995). Only recently, a systematic investigation has begun to understand the ways in which expertise is developed. It is clear from this work on (a) the interaction of the learner's prior experience, culture and system of beliefs; (b) the amount of time and effort required to develop

understanding as opposed to memorization; and (c) the impact of context on the learning outcomes that education for the development of expertise and understanding is a far more difficult task than has faced educators in previous generations. Additionally, the task of educating for expertise and understanding will require new assessment methods as the current techniques relying on total scores fail to capture subtle, but potentially important differences among students with varying levels of expertise.

Eye-gaze tracking

Non-intrusive measures of behaviour concurrent with the performance of an assessment task such as eye-gaze tracking may meet the goal of differentiating individuals with substantial levels of expertise. The eye movements of a student can be captured and recorded in-obtrusively as the student goes about solving problems on an assessment. In this study, we limit our investigation to the examination of eye-gaze tracking in visually represented tasks. A robust body of research has indicated that eye-gaze may be considered an unbiased indicator of the focus of visual attention (see Dillon, 1985a; Pashler, 1998, 1999; Brigham *et al.*, 2001; Salvucci & Anderson, 2001). In fact, much existing research has used this connection between visual attention and eye-gaze to study human cognition in a wide range of topics from reading (Just & Carpenter, 1984) to high-speed train operation (Itoh *et al.*, 2002).

Previous research concerning eye movements in educational assessment has centered on describing the eye movement patterns of individuals with known levels of performance. For example, Hegarty *et al.* (1992) collected the eye-gaze data on students who were grouped as high and low accuracy performers with regard to their ability to solve arithmetic problems. The two groups were composed of the first and fourth quartiles of the population of college students available for their study. As predicted, Hegarty *et al.* found that students in the low accuracy group had quite different patterns of eye movement as compared to the students in the high accuracy group. Specifically, the low accuracy group demonstrated insensitivity to the structure of the problem when the problem was presented in a fashion inconsistent with the way the data was presented (e.g., fuel at Station X is $3.35 per gallon and five cents less at Station Y. How much is fuel at Station Y?).

In a later study, Hegarty *et al.* (1995) compared the eye movements of successful and unsuccessful problems-solvers to examine the extent to which relative levels of performance were affected by how the problem-solvers interpreted the problems. The researchers found that unsuccessful problem-solvers base their solution plan on numbers and keywords that they select from the problem, the direct translation strategy, whereas successful problem-solvers construct a model of the situation described in the problem and base their solution plan on this model, the problem–model strategy.

More recently, Olmeda (2002) found that students with attention deficit hyperactivity disorder (ADHD) could reliably be discriminated from similar age and reading ability peers who did not have ADHD on the basis of their eye movements while reading text on a computer screen. Interestingly, there were no differences in performance

between the groups of students on the initial text passages. The differences in reading behaviour emerged beyond the first 150 words that the students read. In later sections of the text the students with ADHD were less likely to read each word than were students without ADHD. Therefore, students with ADHD were more likely to skip individual words, loose their place in a line, and skip either substantial portions or entire lines of text.

The studies of eye movements related to academic performance described above examined groups with known and varying characteristics to demonstrate that eye movement patterns could be used to reliably discriminate the strategies used by different group members and to actually discriminate members of different groups with similar performance levels on other criterion tests. It therefore appears that eye movements may be useful in discriminating different levels of expertise on complex academics tasks within groups of students who have similar levels of performance.

Assessment

While the ubiquitous paper–pencil test, especially those incorporating a multiple-choice item format, has demonstrated its usefulness over the years, a number of limitations exist with this form of assessment. The ratio of correct to incorrect responses has long been the sole outcome of these tests; however developments in the field of assessment suggest that the total number of correct responses on a given test, while important, may be insufficient evidence for many decision-making tasks required of assessment (Dillon, 1997; Pellgrino *et al.*, 2001). Since most multiple-choice tests limited the access of the test administrator to only correct and incorrect question responses, much information about student thought-processes is lost.

However, within this testing context, there do exist some possible mechanisms for capturing information about student thought-processes. One method for obtaining information would require the test administrator to examine all tests and all items for any marks, such as erasures, that may indicate the student in question altered his or her response. Unfortunately, the absence of such marks does not indicate that the student did not consider alternative choices. Furthermore, the presence of such marks does not provide the test administrator with any information as to the thinking that causes the student to change his or her response. An alternative approach would have the test administrator use techniques such as stimulated recall, where student is interviewed after the test and asked to recall their thoughts while taking the test or assessment task (Ericsson & Simon, 1993), or protocol analysis, where the student is asked to 'think-aloud' during the assessment. These techniques provided insight into the thinking of the student prompted by the test but suffer from two major drawbacks. First, such techniques produce logistical difficulties, as it is inconceivable that for every test, every student taking the test could have one-on-one time with the test administrator to debrief. Second, research has shown that such techniques are vulnerable to expectancy effects and variability in outcome depending on when the verbal protocol is collected (Hayes *et al.*, 1998; Kusela &

Paul, 2000). For example, retrospective protocols may be incomplete because the student may forget, while verbal protocols obtained concurrent with the test may be distracting.

Due to the aforementioned drawbacks in obtaining information about student thinking, educators are reconceptualizing how item responses on a test can best be used (Sadler, 1998). Instead of treating the non-correct selections (distractors) of an item on a multiple-choice test as space fillers, these distractors may be used to understand student thinking. For example, on a mathematics test item assessing a student's knowledge of the addition of fractions (i.e., 3/4 + 5/6) one distractor could display correct addition of the fraction (19/12) but not the correct answer which would require the improper fraction to be reported as a mixed number (1 7/12). The presence of this and other types of 'errors' as distractors for each item would allow to the test administrator to make more informed judgments about the student's understanding of the material. Unfortunately, even this approach still has flaws that can ultimately degrade its usefulness. In particular, this approach to testing while providing more information about student understanding still does not permit the test administrator access to a student's thinking about a particular item. Under these conditions, the test administrator is unaware when two students take completely different approaches to a problem yet settle on the same answer.

One area in which levels of expertise may be useful to examine, even within adequate levels of performance is on minimum competency tests of the type now required by federal law in the US. For high school students, present forms of assessment appear to be aimed at tapping the body of declarative knowledge represented in various courses of study. It is, however, unclear that students with similar scores possess similar levels of declarative information. For example, students with substantial working memory capacity may be able to overcome deficits in declarative information by approaching the task as a problem-solving situation rather than a recall task. Students who derive correct responses via a problems-solving approach certainly can take pride in their accomplishment but they have not executed the same task as their classmates who knew and recalled the information. Among the factors other than possession and recall of the target information that influence performance on such measures are organization of long-term memory, working memory capacities, and language abilities (Pellgrino *et al.*, 2001).

The distinction between students who approach test items in a problem-solving mode versus a recall of declarative information model becomes more important when one considers the relation of expertise to performance within given domains. In short, research in expert performance suggests that experts not only possess greater stores of domain-specific information, but are more efficient and effective in their deployment of their information (Chase & Simon, 1973; Chi *et al.*, 1981). These observations suggest that even when individuals obtain similar scores on tests, they may possess widely different abilities related to the domain. Such ability differences are clearly related to educational assessment because the kind of performance observed in experts is the goal of most classroom instruction.

Therefore, the purpose of this present study is to determine the feasibility of using a subject's eye-movements to provide insight into elements of their performance beyond simple totals of correct and incorrect responses.

Methodology

Participants

This study included six participants, three males and three females, who were all pre-service secondary science teachers enrolled in education courses at the University of Virginia. With regard to science coursework, their backgrounds differed among three areas, biology, chemistry and physics. Each participant expressed varying degrees of ability, familiarity, and confidence in their understanding of the content of a subject area. Their ability, familiarity and confidence were largely based upon the amount of coursework that they had in a given area. A description of each individual's self-report of ability, familiarity, and confidence in their understanding of the content of a subject area as well as the amount of coursework taken in an area can be found in Table 1. For most of the individuals, their self-reports of ability, familiarity and confidence indicated that they would consider themselves 'expert' in only one area, most often associated with their undergraduate major as well as future teaching aspirations. However, two individuals, both male, felt that they had a background that allowed them to consider themselves 'expert' in more than one area. One participant felt that his background was strong in both chemistry and biology, while the other felt that his background was well distributed across all three areas. Not all participants possessed normal uncorrected vision, but those that did not were able to correct their vision by wearing glasses, ruling out visual acuity as a potential source of variability among the participants.

The limited sample size of this exploratory study was due in part to the depth of analysis we wished to pursue with regard to the science learning backgrounds of each of the participants. The intent of our study was to explore the feasibility of eye-gaze tracking as a means of gauging individual expertise and to offer some proof-of-concept regarding this analytical approach. A detailed exploration of the backgrounds and experiences of the participants was essential and contributed to the decision to limit the study sample.

Materials

All participants in the study completed an assessment that contained 18 multiple-choice questions, six in each of the topics of biology, chemistry and physics. Each item in the assessment was composed of four elements: (a) an image (a graph, an illustration, or a table/chart); (b) the text of a question or question stem; (c) the answer as well as three or four alternative responses to the question; and (d) a hyper-link to advance to the next question. Each element appeared in the same part of the screen in each item. See Figure 1 for an example test item used in the assessment.

Table 1. Study participants' demographics (undergraduate major, gender, and age) and topic expertise rating

Demographics	Rating	Comments
Bob Biology Male 22 yrs	Expert	Biology: 'I am very comfortable with the material and that I can answer questions and prepare for lessons without much difficulty' Chemistry: 'I have had over 30 hours of chemistry instruction…I feel that I have a decent hold on chemistry…' Physics: 'I often forget equations or feel that I do, but my grades in the area have been high (A's)…. I could do basic physics but that is where it stands'
	Non-expert	Not Applicable
Bonnie Biology Female 24 yrs	Expert	Biology: 'I am very comfortable with biology. I am most comfortable with the molecular aspect because my undergraduate education was more toward molecular genetics'
	Non-expert	Chemistry: 'Unless a topic in chemistry pertains to something in one of my biology courses or research, I don't know it' Physics: 'I never took physics in high school and maybe I should have so that I would have understood it better when I got to college. … I really do not understand it without some serious tutoring'
Calvin Chemistry Male 23 yrs	Expert	Chemistry: 'This has always been my favourite … the one in which I have had the most extensive coursework and training. As such I am the most comfortable with it' Biology: 'I have not [had] biology in any form since I was a junior in high school, though I had two excellent years of biology and really learned the material well. A good deal of vocabulary is fuzzy but most of the concepts remain'
	Non-expert	Physics: 'I did not take physics until college and it was a horrible experience…. As such, I am the least comfortable with it.'
Carrie Chemistry Female 23 yrs	Expert	Chemistry: ' I TAed general chemistry lab, CHEM 141L/142L & 151L/152L … I taught recitation for CHEM 142 …'
	Non-expert	Physics: 'Biology and physics are a toss up but I lean towards physics … I fell like it's more similar to chemistry than biology' Biology: 'I don't necessarily feel confident answering questions in [biology] without having access to something to back me up … the way I was taught biology was pretty fact based'

Table 1. *Continued*

Demographics	Rating	Comments
Patty Physics Female 29 yrs	Expert	Physics: 'My Father holds a Ph.D. in physics… I learned about single slit diffraction patterns as a toddler with Christmas lights … it's still a family tradition'
	Non-expert	Chemistry: 'My summer research experiences dealt with learning a lot of chemistry tied into physics' Biology: 'I think I know basic biology; I just don't perceive myself to be very inclined/interested in the field'
Paul Physics Male 26 yrs	Expert	Physics: 'Majored in physics as an undergrad, and did very well in high school with it'
	Non-expert	Biology: 'Had one class in college, but just a little bit easier for me to understand than chemistry' Chemistry: 'Had only high school level chem., avoided it like the plague in college'

The graphics were always in the right half of the screen. The question stem appeared in the upper left quadrant and the answer with alternative responses appeared in the lower left quadrant. The hyperlink to advance to the next question appeared at the lower right corner of each screen. The consistent layout of the screen was intended to minimize extraneous eye-movements.

This study used standardized science test questions from released items of the Virginia Standards of Learning end-of-course exams in biology and chemistry and the New York State Regents exam in physics between the years of 1998 and 2001.[1] The selected items were converted to HTML format to allow for electronic display. While every attempt was made to transfer each item faithfully from the paper format to the electronic format some wording of the question was altered to reflect the positioning of the image on the right side of each screen. In the original format, the image was placed in a variety of positions relative to the text and possible responses. Multiple-choice standardized exam questions were chosen for this study for several reasons. First, the task of solving science problems in a multiple-choice format is very common in assessment and familiar to our participants. Second, multiple-choice exams provide clear tasks and clear response options, simplifying the data collected as part of this exploratory study. Third, the test items used for this study had already been administered to high school students in both Virginia and New York and therefore were field-tested. Fourth, the format of multiple-choice questions allowed for the entire assessment item to be displayed on a single computer screen without scrolling, simplifying data analysis.

Experimental apparatus

All of the data for this study was collected on the same apparatus located in an office at the Curry School of Education, University of Virginia. The items for the assessment were displayed using Microsoft Internet Explorer version 6.0 on a DELL P991 17 inch Trinitron monitor interfaced with a Dell Optiplex GX110 computer. All words were displayed as black text against a white background with normal grammatical conventions. Images were inserted as JPEG digital pictures cropped from their original versions. A 'radio button' was provided next to each answer choice in order for participants to indicate their answer selection. Answer selections and any changes were registered by GazeTracker™ software.

Eye movement of the participants was monitored using an Eye Gaze Response Computer Interface Aid (ERICA) apparatus connected to the computer. Only a single computer was used in the set-up. GazeTracker™ software stored and managed the eye-movement data monitored by the ERICA apparatus. ERICA functions by monitoring reflections of infrared light off of the cornea and retina of one eye of a participant. The resolution is 0.5° of the visual angle, while the sampling rate in this study was 60 samples per second. The ERICA apparatus was located beneath the computer monitor with the headrest fastened to the front-edge of the desk to steady the head of the participant. The computer was positioned on the floor. A separate close-circuit television monitor was used by researchers to monitor participants' eye

and head position. A typical experimental trial including calibration lasted less than 20 minutes with no discernable change in participants' eye and head position.

Procedure

After completing a demographic survey, the participants were seated in front of the apparatus and positioned to permit data collection. The participant's seating and head position placed his or her eye 64.3 cm from the monitor, matched with the center of the computer display screen. To stabilize head position each participant rested his or her forehead on a crossbar headrest positioned just above the eyebrow ridge. The eye tracking system was then calibrated in a process that took approximately five minutes. The participants were told, prior to their participation, that they would be answering a series of 18 questions drawn from biology, chemistry and physics. The items were presented in the same order for each participant progressing from biology to chemistry to physics. Once a participant had answered an item and advanced to the next item, it was not possible to return to the previous item. However, if a participant felt it necessary to change his or her answer selection he or she could do so prior to leaving the question. The final instruction given to each participant was to answer each item before advancing to the next. The average time for completing all 18 items in the assessment was 10.9 minutes (SD = 2.5 minutes). Following completion of the assessment, each participant was debriefed. In addition, after initial data analysis, all participants responded to an email questionnaire to further clarifying their educational background.

Eye fixation parameters

The previously discussed procedures yielded data in the form of eye fixation locations, fixation durations, saccades, and saccadic durations. An eye fixation is the

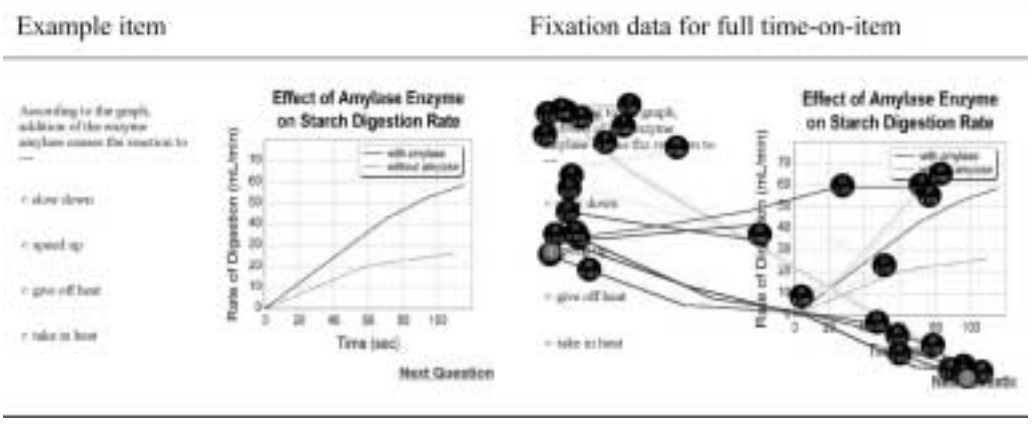


Figure 1. Example item and example of fixation data from Bonnie

period in which the eye remains relatively still allowing a person to collect visual information (Rayner, 1998). Fixation duration, typically measured in milliseconds (ms), represents the amount of time that an eye remains fixated. Saccades are gross eye-movements as an individual's gaze changes for example during reading or while examining an object or scene. Due to the speed of movement of the eye during a saccade, a person cannot recognize any visual information that his or her eye may collect (Rayner, 1998). In Figure 1, the left panel shows a blank example item, while the panel on the right shows the fixations on that item for an individual. Note: the black dots in the figure covering various elements represent the location of eye fixations while the lines connecting successive fixations represent saccades.

During a fixation, the human eye does not remain perfectly still; therefore, a fixational radius must be calculated to represent the region in which slight eye-movements will be measured as a single fixation. We have chosen to base our calculations of the fixational radius on the central 2° of the field of view associated with the region of highest visual acuity, the fovea (Rayner, 1998). This decision means that eye movement outside the set fixational radius will be recorded as a saccade. When an individual's gaze stops at a new point and remains within the set fixational radius about this new point, a new fixation will be registered. This process continues as long as a participant's gaze is monitored. As mentioned earlier, the distance measured from an individual's eye to the center of the computer monitor was found to be 64.3 cm, resulting in a calculated fixational radius of 1.1 cm on the monitor.

In addition to showing the locations of the eye fixations, the black dots shown in the figures contain information about the order in which each fixation occurred on an assessment item as well as the duration of each fixation. For the purposes of this analysis, 100 ms was set as the lower limit for fixation duration. This value was selected based on the work of McConkie *et al.* (1985), who examined the temporal characteristics of visual information processing during reading. These researchers noted that several temporal markers are associated with eye fixation during reading. Beginning with the termination of a saccade, 60 ms must pass before current visual information becomes available to the visual cortex for processing. At the end of a fixation, the time between when a command to move the eyes is sent and the onset of that saccade is reported to be 30 ms. If 10 ms are allowed for the processing of any currently observed text, we arrive on our lower limit of 100 ms for a fixation duration.

Before continuing on to the results and discussion, we wish to include a short note on the issue of multiple-choice test taking strategy. It seems appropriate to discuss whether our study would capture individuals' 'natural' behaviours versus test taking strategies. We have little doubt that most individuals have a strategy for approaching multiple-choice test questions. As a result we expect to see a 'baseline' pattern common to an individual across all test questions. In fact, a general baseline pattern may be common among several individuals. However, our study can also detect eye-gaze behaviours that depart from this baseline pattern and it is these variations we have set out to study. We believe that in order for test taking strategies to invalidate our study, an individual would need to concentrate solely on following a prescribed pattern of behaviour with little variation, while ignoring his or her thoughts and ideas

as he or she collects information to answer a test question. Our study will produce the information to allow us to make such a determination.

Results and discussion

Analysis of outcome scores and latent response times

While our study centers on the use of eye-gaze data to analyze problem-solving behaviour, we feel it is important to address two commonly used measures of problem-solving performance, outcome scores and latent response times. Though important, correct responses and latent response times are only gross measures of performance and offer little detail with respect to the problem-solving approaches taken by our participants. As a result, the following discussion provides some context to the eye-gaze analysis to follow.

The participants' responses to the various items used in the assessment were recorded via GazeTracker™ software. Their overall number of correct responses as well as their number of correct responses for each science sub domain is shown in Table 2. The overall mean for correct responses is 13.8 (SD = 2.3) out of a possible score of 18. Every participant responded incorrectly to at least one item and no item was incorrectly responded to by all participants. When the results are examined by sub domain, four of the six participants had very similar scores regardless of topic. Only Bonnie and Calvin showed markedly reduced performance in a particular topic, i.e., physics. Based on a comparison of correct responses, all participants performed comparably in biology and chemistry, and all but two performed comparably in physics.

Next consider latent response times, i.e. total time spent on the test questions. Here the data is less similar across topic and across participants. For example, Bonnie's latent response time of 3.15 minutes in biology is clearly shorter than her other latent response times in chemistry (4.22 minutes) and physics (4.17 minutes). This data suggests that she is most expert in the topic of biology. A comparison of latent response times within each individual participant yields fairly similar results, with only Patty displaying little variation across all topics. Difficulties arise when comparing latent response times across individuals within topics. In biology, five participants all show very similar latent response times, with only three claiming some level of expertise. In chemistry, four individuals have very similar latent response times, all within 30 seconds totaled over six questions. In physics, Paul, rated as more expert in physics than the other topics, reported a latent response time more comparable to the non-experts. In general, cross-participant comparisons of latent response times appear to be less robust than within-participant comparisons. This result suggests that an individual's work pace must be considered in an analysis comparing relative expertise.

Visual inspection of the data

When comparing the fixations of two different individuals on the same assessment item, clear qualitative differences between the number, density, and clustering of

Table 2. Outcome scores and latent response times for each participant

	Biology		Chemistry		Physics		Total (18 max.)	
Participant	Items correct	Time (minutes)	Items correct	Time (minutes)	Items correct	Time (minutes)	Items correct	Time (minutes)
Bob	6	3.54	6	2.96	5	2.77	17	9.28
Bonnie	5	3.15	4	4.22	1	4.17	10	11.55
Calvin	6	2.44	6	2.94	2	4.25	14	9.63
Carrie	5	3.67	6	2.50	4	6.22	15	12.40
Patty	4	3.05	5	2.55	4	2.25	13	7.84
Paul	5	5.23	5	5.35	4	4.04	14	14.62
Average (s.d.)	5.2 (0.8)	3.51 (0.94)	5.3 (0.8)	3.42 (1.13)	3.3 (1.5)	3.95 (1.13)	13.8 (2.3)	10.89 (2.45)

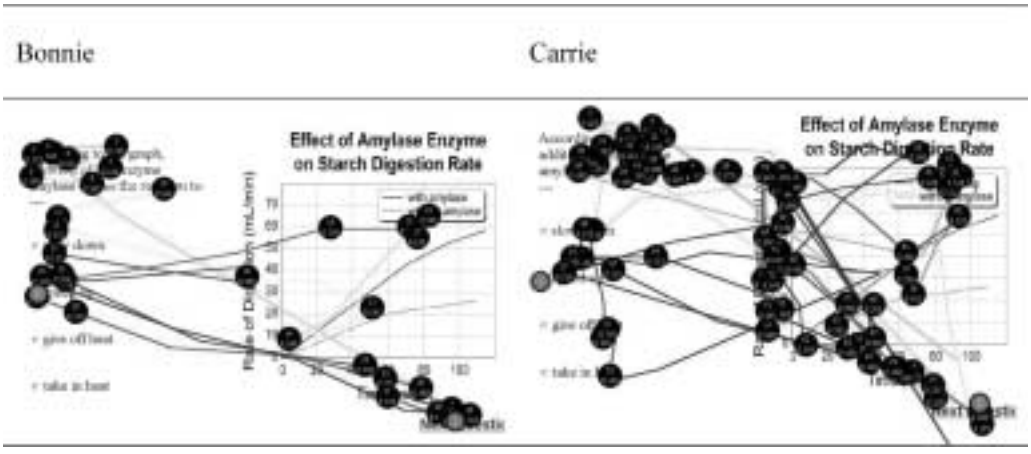


Figure 2. Comparison of fixation density and saccades on the same assessment item

fixations and the number of saccades connecting the fixations could be seen in several cases. Consider Figure 2, in which the data collected from two individuals, Bonnie and Carrie, from the same item, a biology question, is shown. Notice that the density of fixations on the image, a graph, on the right of each assessment item is much greater for Carrie as compared to Bonnie. It is important to keep in mind that fewer overall fixations suggest less time spent viewing specific areas of the assessment item, while fewer saccades suggests fewer movements among fixations. The difference in the density of fixations indicates that Carrie not only looked at the graph more than Bonnie, but that she was also examining different elements of the graph more closely. When we compare this visual inspection of the data with the information obtained from the participants regarding their knowledge of biology, a connection begins to emerge. In this instance, Bonnie's expertise in the topic of biology is much greater than Carrie's expertise. (See Table 1.) This initial inspection of the fixations and saccades suggests that a wider comparison might be important.

In Figure 3, the fixations and saccades of three individuals with expertise in three different topics are shown. The assessment items are grouped by individuals in the columns and by topics in the rows. An examination of the figure as a whole reveals differences in the density of fixations and the number of saccades across the assessment items shown. Turning to a more fine-grained examination by looking a particular individual's data also reveals differences. Bonnie has a greater density of fixations for the chemistry and physics assessment items as opposed to biology when compared to Calvin and Paul. For Calvin, the density of fixations on his biology and chemistry assessment items appears quite similar; however, there is a noticeable increase in fixations for his physics assessment item. In Paul's case, physics is the item on which he displays the fewest fixations as opposed to biology and chemistry. When we compared this data concerning amount of fixations and saccades we found that for the participants in this study fewer fixations and

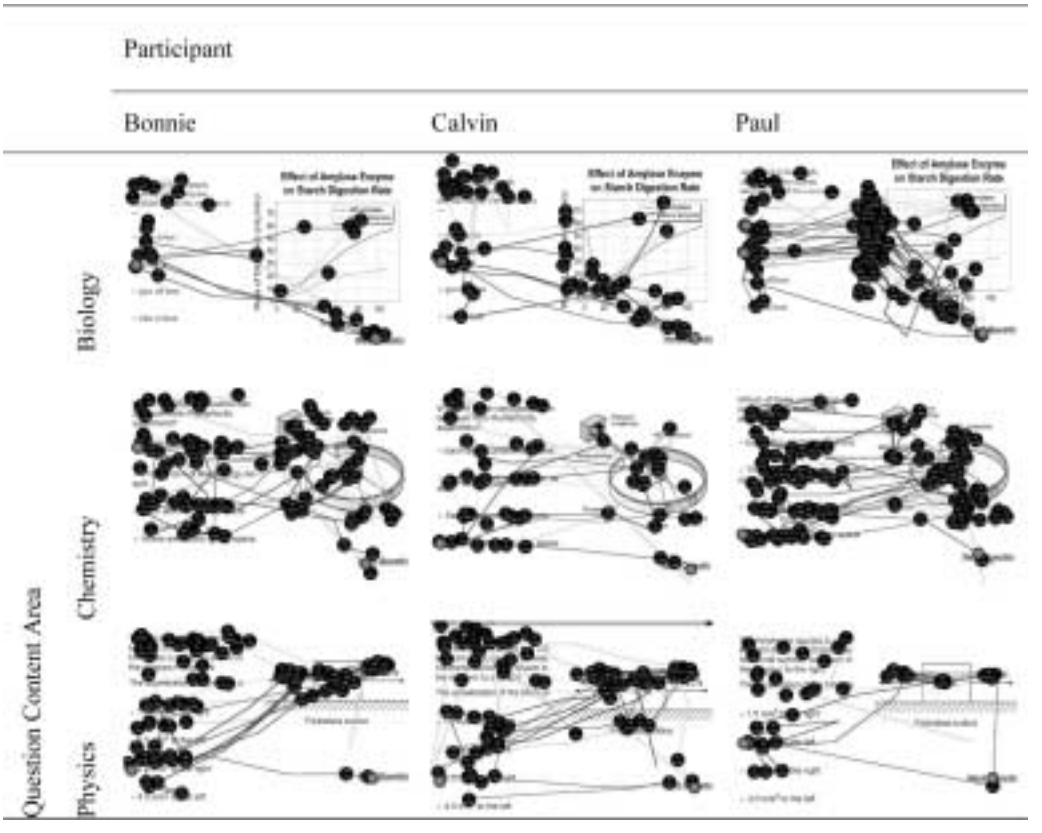


Figure 3. Selected fixation and saccade data for three participants

saccades were associated with an individual's expertise. Therefore, the data from a visual inspection and comparison of Bonnie's assessment items, both across individuals for the topic of biology and across her chemistry and physics assessment items, corresponds with her strong background in biology. This correspondence between fixation and saccade amount and expertise appears true for the other participants in this study as well.

The striking differences among the assessment items and the apparent association of these differences with expertise prompted us to further pursue a method of analysis to estimate relative expertise among the individuals. It should be noted that the assessment items shown here were selected for the clear differences among these individuals in order to make the point that fixation and saccade density and distribution appear to be associated with topic expertise. The differences in the density and distribution of fixations and saccades for some other assessment items are less striking. Therefore, rather than relying on a qualitative rating of relative expertise based on visual inspection, we chose to analyze the data in terms of quantifiable measures.

Quantitative analysis

In our analysis of the fixation durations, we considered total time within each science topic. This analysis did not produce clear connections between an individual's area(s) of expertise and total time within each topic. We discovered large degrees of variation within and across individuals with respect to the amount of time spent on each assessment item and within each science topic. A comparison of the fixation duration data did not produce clear and consistent differences corresponding to known levels of expertise among the six individuals in our study. These results agreed with those found by Chi *et al.* (1982). However, this finding does not suggest that there is no connection between the speed with which an individual solves a problem and his or her level of expertise. In fact, Glaser and Chi (1988, p. xviii) later write, 'Experts are fast; they are faster than novices at performing the skills of their domains …'. These findings suggest that individual differences among problem-solvers are an important consideration. Later in this discussion our analysis will provide an example linking speed and expertise for specific assessment items. The data collected in this pilot study serve as a single measure of problem-solving speed for each participant. It may well be that had measures been taken over time as an individual developed his or her current level of expertise, his or her time-on-item may very well have decreased, yet remained longer in duration than another person with less expertise. In short, some people take more time to do things even when they are experts.

Next, we turned to an analysis of fixation duration and fixation allocation. An initial step in this approach was to categorize the fixations according to their location on the assessment item. The collected data included start times, end times, and position coordinates for each fixation on each slide along with a categorization of each fixation according to *look zones* that we defined around the informational elements for each assessment item. The look zones were four rectangular areas, which overlaid on the question stem, the image, the answer and alternate responses, and the advance hyperlink. Since the size of these various elements differed depending upon the item, each item was assigned its own set of look zones that were used for the analysis across all of the participants. See Figure 4 for an example of look zone arrangement and size for a particular assessment item. All the fixations of each of the six participants on a given assessment item were categorized by the GazeTracker™ software using the look zone assignment for that particular assessment item. In some instances, positional shifts in the data were caused by an individual's head shift. The shifts discovered in the data were typically vertical. These shifts were very consistent and given the size of the look zones, typically an area of 24 cm^2,[2] only fixations falling along the boundaries between the question zone (Q-zone) and the answer zone (A-zone) or the image zone (I-zone) and the 'next question' hyperlink (H-zone) were of concern since within each pair the individual look zones are positioned directly above each other. In all instances, the clustering of fixations provided clear evidence of the proper categorization of the shifted fixations. Of the 10,489 fixations in the data set, the look zones for 436 fixations (4.2%) were manually identified by the researchers by examining each assessment item one fixation at a time. For the other 95.8% of the fixations,

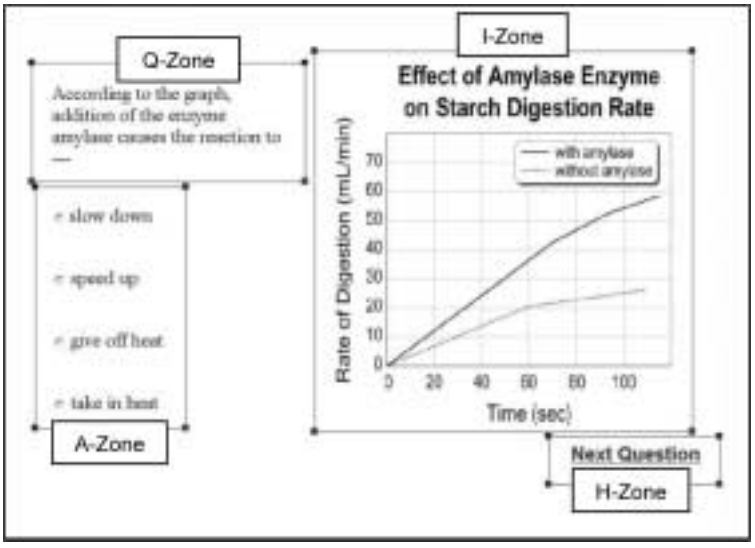
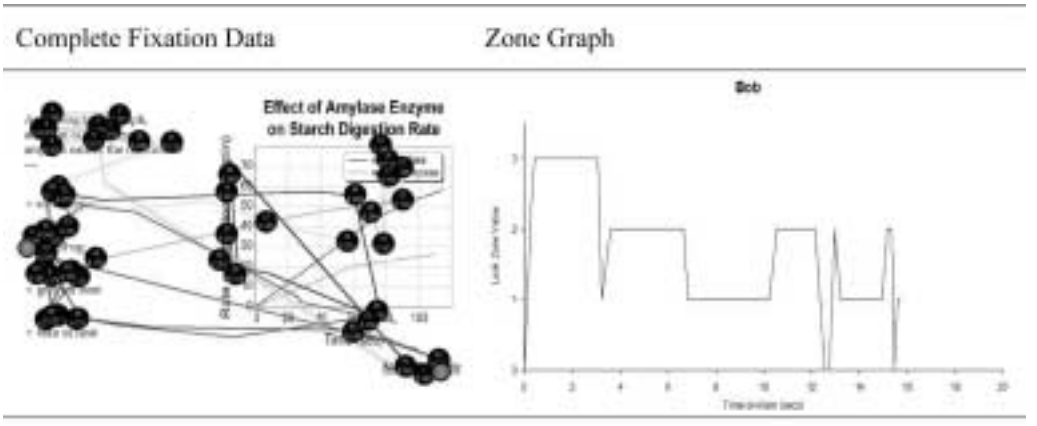


Figure 4. Look zone positions for enzyme question

GazeTracker™ unequivocally identified the fixation's location in one of the four look zones. Our challenge was to produce a representation of the fixation duration and fixation allocation information in a form more transparent for analysis. For this purpose, we assigned numerical values to each of the four look zones (Q-zone = 3, I-zone = 2, A-zone = 1 and H-zone = 0). Plotting the time-on-item along the horizontal axis, and the assigned look zone value along the vertical axis, we produced a graph that included both time-on-item and fixation location that we have termed a *zone graph*. An example of a zone graph along with the assessment item from which the data was obtained is shown in Figure 5.

Taking a look at the zone graph in Figure 5, we have a representation of Bob's fixations stretched across the temporal dimension of *time-on-item*. Upon viewing this assessment item, note that Bob immediately saccades to the Q-zone (= 3) then after approximately three seconds in this zone, he saccades to the A-zone (= 1) for a single fixation before making a saccade to the I-zone (= 2). After making several fixations in the I-zone, Bob saccades to the A-zone and fixates there for approximately another three seconds during which time he selects the correct answer before returning to the I-zone. After a second round of fixations in the I-zone, Bob fixates in the H-zone (= 0), but he does not activate the hyperlink to advance to the next assessment item. Instead, he briefly returns to the I-zone, fixates in the A-zone, and returns to the I-zone for a fourth time before fixating in the H-zone and then activating the hyperlink. Zone graphs translate fixation data into a quantitative format allowing for the comparison of participants eye-movements among various look zones.

To embark on these comparisons, it is necessary to examine sets of zone graphs together at one time. Figure 6 is a display in zone graph format of the fixation data shown in Figure 3. An examination of the graphs shown in Figure 6 reveals a similarity



[1]For the purpose of display, the horizontal time-on-item scale was set to a maximum of 20 seconds for the zone graph.

Figure 5. Complete Fixation Data & Zone Graph for Bob on the Enzyme Question.[1](Q- Zone Fixation = 3, I-Zone Fixation = 2, A-Zone Fixation = 1, and H-Zone Fixation = 0)

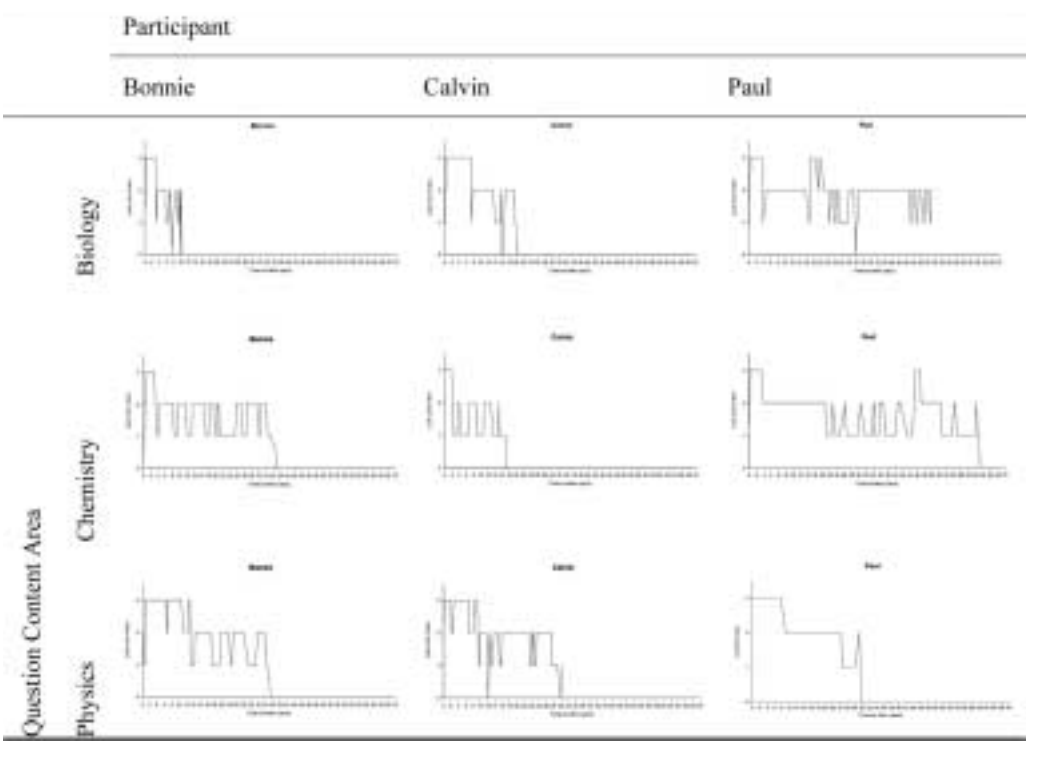


Figure 6. Zone Graphs Comparing Three Participants Eye-fixation Data from Figure 3 (On the vertical scale: Q-Zone Fixation = 3, I-Zone Fixation = 2, A-Zone Fixation = 1, and H-Zone Fixation = 0)

in problem-solving behaviour common across the three participants, and in fact, among all participants in this study. For each individual, regardless of the item's topic area, the initial action taken was to saccade to and to fixate in the Q-zone. This behaviour suggests that the participants begin problem-solving by looking at the question/question stem regardless of expertise in a given topic.

When we consider the data provided by the zone graphs (See Figure 6), we found several interesting characteristics. Consider the zone graphs in the topic of physics for Bonnie, Calvin and Paul. Note that Bonnie and Calvin are known to lack expert knowledge in physics while Paul is a physics expert. In fact, both Bonnie and Calvin report this science topic to be their weakest area. Note the number of saccades crossing between the I-zone and A-zone for these two physics non-experts. This result is in stark contrast to Paul's zone graph for the same physics item. An examination of this graph shows that Paul progresses from one zone to the next. He does not return to a particular look zone once he has exited it on this physics assessment item. When considering Paul's behaviour in topics outside of his expertise, we can see that the number of saccades crossing among the look zones is much greater. He made multiple interactions with the material in each of the look zones, before committing to a response and moving on to the next question. This behaviour is in stark contrast to the behaviour he exhibited on the physics item. Neither biology nor chemistry fell within Paul's scope of expertise. Paul serves as the clearest example of this difference in problem-solving behaviour.

Another important characteristic of the data displayed in Figure 6, is the clear link in these three assessment items with the topic of expertise of the three participants. Concentrating on Paul's chemistry zone graph, note that his eye-gaze was initially in the Q-zone for a period of time and then moved to the I-zone for a longer period, he then made several saccades between the I-zone and the A-zone before returning to the Q-zone and repeating the eye-gaze behaviour. Also, note that Paul's biology zone graph shows a modified version of this pattern. For the example physics problem in Paul's topic of expertise, his behaviour reveals much less saccadic activity among zones. In fact, his behaviour shows a progression across zones with few saccades across zones. This type of behaviour was exhibited in the zone graphs of several participants.

Finally, for all of the participants, we noted that the majority of saccades between look zones appeared between the I-zone (image) and the A-zone (answer). Figure 6 shows this activity. Saccades entering and exiting the Q-zone (question/question stem) are not as abundant. However, when the saccades to the Q-zone do occur, the assessment items eliciting this behaviour appear to fall outside of an individual's topic of expertise. Figure 6 shows this trait for all three participants except in one particular case. This exception occurs for Bonnie, a biology expert, on the chemistry item; she does not make a saccade back to the Q-zone once her gaze exits. In this case, after exiting the Q-zone, her saccades and fixations are confined between the I-zone and A-zone, much like Calvin. This finding has two interesting implications. First, while it is possible to discuss global domains of expertise (i.e., a biology expert, a chemistry expert, etc) with regard to an individual, these global domains of expertise may not

be an entirely accurate portrayal of an individual's understanding of specific topic knowledge (Glaser & Chi, 1988; Patel *et al.*, 1999). Therefore, analysis of relative areas of expertise using eye movements will require a comparison of a number of items both within and among topics. Second, the correspondence between the number of saccades returning to the Q-zone and participant expertise may provide a means for gauging an individual's expertise through eye-gaze tracking.

Conclusions and considerations for future research

The purpose of the present study was to determine the feasibility of using a subject's eye-movements to provide insight into elements of their performance beyond simple totals of correct and incorrect responses. Evaluating test scores on multiple-choice assessments provides an estimate of student science achievement. However, a test score is limited and provides little to no information about the problem-solving behaviours and strategies used by the test-taker. Thus, assessments of expertise based on test scores may at times be misleading and very often are limited in their capacity to gauge expertise. Indeed, previous research has suggested that individuals who obtain the same score on a given test may vary widely in their particular knowledge of the topic or level of expertise (Chase & Simon, 1973; Chi *et al.*, 1981; Dillon, 1985b). In addition to the selection of a correct response to a given test item, test takers can vary in their organization of background knowledge, insight into problem structure, and responsiveness to the context in which a given test item is embedded. Various methods have been suggested for tapping aspects of the test performance beyond total performance scores; however, many such methods are vulnerable to questions of intrusiveness and subject reactivity. Eye movements have been suggested as non-biased indicators of attentional allocation and attentional allocation is one indicator of cognitive activity.

Note that the participants in this study were university students who possessed different levels of knowledge of the science topics included in the study. Each participant viewed and responded to multiple choice assessment items in biology, chemistry, and physics, all topics with which they have had some exposure to in either high school or college. The participants included in this study were selected from a narrow range of scientific expertise at the post-secondary level. The results of the pilot study suggest that eye-gaze tracking was able, at least in this limited group of individuals, to provide evidence distinguishing the levels of expertise among our participants. This finding suggests that eye-gaze tracking may be relatively sensitive to small differences in scientific expertise.

Earlier work by Chi *et al.* (1982) reported a lack of differentiation between novices and experts when examining wholesale measures of time on particular problem-solving tasks. Analysis of the overall time spent by an individual on particular aspects of assessment items support Chi *et al*'s conclusions. However, a qualitative examination of the participants' eye movements on the various items suggested that differences existed, first, within individuals across different topics and, second, across individuals within the same topic. In order to transfer the data provided from the

qualitative examination of an individual's eye movement on a particular item into a form comparable across items and individuals in this study, we developed an analytic technique we have termed zone graphs. These zone graphs allowed eye movement data to be displayed concurrent with the time each participant spent on each assessment item. Using this technique, we have found that differences in eye-movements do exist across individuals within a particular scientific topic as well as within an individual across the three different scientific topics. Furthermore, some evidence from this study suggests that differences in eye movement may correspond to an individual's level of expertise within a given topic. While further work is still needed to better understand and quantify this observation, the evidence garnered from our pilot study suggests that this vein of research may provide some valuable insights in to an individual's problem-solving strategies and subsequently inform researchers of their level of expertise. Should these next steps prove to be fruitful, the possibility of using eye-gaze tracking as measures of developmental expertise in learners may lie in the future.

In summary, we have used eye-gaze tracking to tap into information not typically accessible in a standardized multiple-choice assessment format. It appears that this particular type of approach allows for the identification of when individuals express more and less efficient attentional allocation routines in problem-solving. This aspect of behaviour appears to be strongly linked with expertise and allows for the discrimination of highly developed expertise from emerging levels. Given the similarity of academic competence among the individuals participating in this study, it is encouraging that our eye-movement protocol was able to discriminate among their rather narrow bands of science expertise. However, as desirable as these levels of performance are, highly competent individuals such as our participants are rarely a concern to educators in general. Rather, much effort has been devoted to developing the academic abilities of less adept learners. Future research might examine the extent to which these protocols may be applied to assisting educators in the diagnosis of ineffective educational techniques and intervening in the teaching process to better tailor instructional methodology to student needs.

Notes

1. The researchers are based at the University of Virginia and therefore chose to use the standardized science exams for Virginia. However, Virginia did not have an exam in physics at the time of the study and so the researchers turned to the New York State Regents Exams, which have a long history in the development and administration of standardized exams.

References

Brigham, F. J., Zaimi, E., Matkins, J. J., Shields, J., McDonnough, J. & Jakubecy, J. (2001) The eyes may have it: reconsidering eye-movement research and human cognition. In: T. E. Scruggs & M. A. Mastropieri (Eds) *Advances in learning and behavioral disabilities: Technological Applications* (New York, Elsevier Science), Vol. 15, 39–59.

Chase, W. G. & Simon, H. A. (1973) Perception in chess, *Cognitive Psychology*, 4(1), 55–81.

Chi, M. T. H., Feltovitch, P. J. & Glaser, R. (1981) Categorization and representation in physics problems by experts and novices, *Cognitive Science,* 5, 121–152.

Chi, M. T. H., Glaser, R. & Rees, E. (1982) Expertise in problem-solving, in: R. J. Sternberg (Ed.) *Advances in the psychology of human intelligence. Volume 1* (Hillsdale, NJ, Lawrence Erlbaum Associates), 7–75.

Dillon, R. F. (1985a) Eye movement analysis of information processing under different testing conditions, *Contemporary Educational Psychology,* 10(4), 387–395.

Dillon, R. F. (1985b) Predicting academic achievement with models based on eye movement data, *Journal of Psychoeducational Assessment,* 3(2), 157–165.

Dillon, R. F. (1997) A new era in testing, in: R. F. Dillon (Ed.) *Handbook on testing* (Westport, CN, Greenwood Press) pp. 1–19.

Ericsson, K. A. & Simon, H. A. (1993) *Protocol analysis: verbal reports as data* (Cambridge, MA, MIT Press).

Glaser, R. & Chi, M. T. H. (1988) Overview, in: M. T. H. Chi, R. Glaser & M. J. Farr (Eds) *The nature of expertise* (Hillsdale, NJ, Lawrence Erlbaum Associates), xv–xxviii.

Hayes, S. C., White, D. & Bissett, R. T. (1998) Protocol analysis and the 'silent dog' method of analyzing the impact of self-generated rules, *Analysis of Verbal Behavior,* 15, 57–63.

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.

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.

Itoh, K., Arimoto, M. & Akachi, Y. (2002, September) Gaze relevance metrics for safe and effective operations of high-speed train: application to analysing train drivers' learning with new train interface, paper presented at the *11th European Conference on Cognitive Ergonomics*, 77–84, Catania, Italy. Available online at: www.ie.me.titech.ac.jp/lab/itoh/pdf/Itoh-ECCE11-rev.pdf (accessed 17 April 2006).

Just, M. A. & Carpenter, P. A. (1984) Using eye fixations to study reading comprehension, in: D. E. Kieras & M. A. Just (Eds) *New methods in reading comprehension research* (Hillsdale, NJ, Lawrence Erlbaum Associates), 151–182.

Kusela, H. & Paul, P. (2000) A comparison of concurrent and retrospective verbal protocol analysis, *American Journal of Psychology,* 113(3), 387–404.

Mastropieri, M. & Scruggs, T. (1999) *Teaching test taking skills: helping students show what they know* (Cambridge, MA, Brookline Books).

McConkie, G. W., Underwood, N. R., Zola, D. & Wolverton, G. S. (1985) Some temporal characteristics of processing during reading, *Journal of Experimental Psychology: Human Perception and Performance,* 11(2), 168–186.

National Research Council (1995) *National science education standards* (Washington, DC, National Academy Press).

Olmeda, R. A. (2002) *Using eye movements to differentiate students with and without ADHD in a simple reading task.* Unpublished manuscript, Charlottesville, VA, The University of Virginia.

Pashler, H. E. (Ed.) (1998) *Attention* (East Sussex, Psychology Press).

Pashler, H. E. (1999) *The psychology of attention* (Cambridge, MA, MIT Press).

Patel, V. L., Arocha, J. F. & Kaufman, D. R. (1999) Expertise and tacit knowledge in medicine, in: R. J. Sternberg & J. A. Horvath (Eds) *Tacit knowledge in professional practice: Researcher and practitioner perspectives* (Mahwah, NJ, Lawrence Erlbaum), 75–99.

Pellgrino, J., Chudowsky, N. & Glaser, R. (Eds) (2001) *Knowing what students know: the science and design of educational assessment* (Washington, DC, National Academy Press).

Rayner, K. (1998) Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin,* 124(3), 372–422.

Sadler, P. M. (1998) Psychometric models of student conceptions in science: reconciling qualitative studies and distractor-driven assessment instruments, *Journal of Research in Science Teaching,* 35(3), 265–296.

Salvucci, D. D. & Anderson, J. R. (2001) Automated eye-movement protocol analysis. *Human-Computer Interaction,* 16(1), 39–86.

Scruggs, T. E., White, K. R. & Bennion, K. (1986) Teaching test-taking skills to elementary-grade students: a meta-analysis, *Elementary School Journal,* 87(1), 69–82.

Sternberg, R. J. & Horvath, J. A. (1995) A prototype view of expert teaching, *Educational Researcher,* 24(6), 9–17.