Using ScanMatch Scores to Understand Differences in Eye Movements Between Correct and Incorrect Solvers on Physics Problems

Adrian Madsen
Kansas State University
adrianc@ksu.edu

Adam Larson
Kansas State University
adlarson@ksu.edu

Lester Loschky
Kansas State University
loschky@ksu.edu

N. Sanjay Rebello
Kansas State University
srebello@ksu.edu

Abstract

Using a ScanMatch algorithm we investigate scan path differences between subjects who answer physics problems correctly and incorrectly. This algorithm bins a saccade sequence spatially and temporally, recodes this information to create a sequence of letters representing fixation location, duration and order, and compares two sequences to generate a similarity score. We recorded eye movements of 24 individuals on six physics problems containing diagrams with areas consistent with a novice-like response and areas of high perceptual salience. We calculated average ScanMatch similarity scores comparing correct solvers to one another (C-C), incorrect solvers to one another (I-I), and correct solvers to incorrect solvers (C-I). We found statistically significant differences between the C-C and I-I comparisons on only one of the problems. This seems to imply that top down processes relying on incorrect domain knowledge, rather than bottom up processes driven by perceptual salience, determine the eye movements of incorrect solvers.

CR Categories: J.2 [Computer Applications]: Physical Science and Engineering – Physics

Keywords: eye movements, attention, scan path, ScanMatch, problem solving, physics

1 Introduction

Researchers have found consistent patterns of wrong answers to many simple conceptual physics questions [Trowbridge and McDermott 1980; McDermott et. al 1987]. Several cognitive top-down explanations have been provided, including misconceptions formed through interactions with the natural world or misapplication of conceptual resources [Docktor and Mestre 2011]. However, recent claims by Heckler [2011] have suggested a perceptual basis for students' incorrect answers, which are based on attention being directed to the most perceptually salient and plausibly relevant features in a problem. The most salient features capture attention through perceptual processes and less salient features have little opportunity to be considered. Heckler shows some evidence for perceptually-driven responses; however, no eye movement data supporting this hypothesis is provided. Further, he does not provide a specific definition of salience. Therefore, incorrect answers may be governed either by top-down processes relying on incorrectly learned or applied information, or by bottom-up perceptual processes resulting in certain elements capturing attention and leading to activation of reasoning resources based on these elements.

An eye-movement study was used to test these competing hypotheses. Introductory and graduate physics students answered conceptual physics problems regarding a diagram [Madsen et al., 2011]. Three areas of interest (AOIs) were defined for each diagram. First, thematically-relevant AOIs that contained information necessary to correctly answer the question were determined by experts in physics. Second, novice-like AOIs were defined based on coded interview data from novices [Madsen et al. 2011], and third, perceptually salient AOIs were defined as the area(s) on the diagram with the highest saliency rating according to the salience maps produced by a computational algorithm [Itti 2000]. For each problem, the percentage of time spent in each type of interest area was compared between students who answered the problem correctly and those who answered the problem incorrectly.

If top-down cognitive processes utilizing naïve theories or misapplied information were directing attention in physics problems, then those who answer the problems incorrectly should spend more time looking at the novice-like AOIs than those who answer correctly. If perceptual salience captures attention and leads students to an incorrect answer, then more time should be spent looking at perceptually salient AOIs. We found that in five of six problems, those who answered incorrectly spent significantly more time looking at the novice-like AOI than those who answered correctly. No differences were found between correct and incorrect solvers in the perceptually salient AOIs. However, it is important to note that Carmi and Itti [2006] studied the effects of perceptual saliency as a function of viewing time. They found that their model of perceptual salience performed best on the first six to seven fixations when viewing a scene. For the average viewer, this is equivalent to about the first two seconds of viewing. In light of this finding, we also compared the amount of time spent in the perceptually salient AOI during the first two seconds of viewing the diagram for those who answered correctly versus incorrectly. No significant differences were found between those who answered correctly versus incorrectly, although the data were in the predicted direction (i.e., the raw percentage of time spent in the perceptually salient AOI was higher for those who answered incorrectly on five of the six problems analyzed). Thus, it may be either that the small number of fixations observed in the first two seconds of diagram viewing lacked the statistical power to find an effect, or there may simply be no effect between those who answer correctly versus incorrectly on the viewing time of perceptually salient elements of the diagram.

In this paper, we will expand on our previous work [Madsen et al. 2011] to further investigate the role of perceptual salience in guiding the attention of those who incorrectly answer conceptual physics questions containing a diagram. A scan path analysis was performed using an algorithm called ScanMatch [Cristino et al. 2010], which is based on the Needleman-Wunsch algorithm used to compare DNA sequences. ScanMatch bins a saccade sequence both spatially and temporally and then recodes this information to

create a sequence of letters which represents the location, duration, and order of the fixations. The letter sequences of two sets of eye movements are then compared to each other to calculate a similarity score. A similarity score near one represents two sequences of eye movements that are very similar spatially and temporally. The ScanMatch analysis requires no decisions to be made about the data a priori, for example, one does not have to define AOIs based on an experimenter's definition or rating. Therefore, it is possible that differences exist in sets of eye movement data that are not detected by looking at fixation durations in AOIs.

We will compare the average ScanMatch scores produced by comparing the correct solvers to one another (C-C comparison), the incorrect solvers to one another (I-I comparison), and the correct solvers to the incorrect solvers (C-I comparison).

We hypothesize that if the incorrect solvers are being primarily led by the perceptual salience of the elements in the diagram, then it is likely that they will attend to the same elements in a similar order. For example, attention would be first guided to the most perceptually salient region, followed by the next most salient region, and so on [Itti 2000]. Thus, the I-I comparison would have higher ScanMatch scores than the C-C comparison, who might attend to perceptually salient areas early on in diagram viewing; however, the variable onset of top-down processes on eye movements would result in greater temporal and spatial variability of gaze towards thematically-relevant elements in the diagram, resulting in lower ScanMatch scores. The I-I and C-C groups would also have higher ScanMatch scores than the C-I group, since the correct solvers and incorrect solvers are known to spend different amounts of times looking at thematically-relevant and novice-like elements [Madsen et al. 2011; Carmichael et al. 2010].

Conversely, if top-down processes are directing the attention of incorrect solvers, namely some form of naïve theory, the ScanMatch score of the I-I comparison should be similar to that of the C-C comparison. The domain knowledge possessed by those in both comparison groups, whether correct or incorrect knowledge, guides their attention to look at certain elements of the problem, but not in a particular order. Once again, the I-I comparison and the C-C comparison should have higher ScanMatch scores than the C-I comparison.

In summary:

Hypothesis 1: If perceptual salience is primarily influencing the attention of incorrect solvers, the I-I comparison will have higher ScanMatch scores than the C-C comparison.

Hypothesis 2: If top-down processes utilizing naïve theories are primarily influencing the attention of incorrect solvers, the I-I comparison and the C-C comparison will have similar ScanMatch scores, and these will both be higher than the C-I comparison.

2 Methodology

There were 24 participants (three females), with two different levels of experience in physics. Ten participants were PhD students in physics and one was a postdoctoral researcher in physics; all had taught an introductory physics course. Thirteen participants were introductory psychology students who had taken at least one physics course in high school, though some had also taken a physics course at the university. The PhD students and post-doc voluntarily participated while the psychology students received course credit. Since we sought to compare those who answered the physics problems correctly versus incorrectly, we selected participants with a broad range of experience. We expected the PhD students to answer correctly, while the psychology students might answer incorrectly, though we know that this may not always be the case as it has been shown there is a wide distribution of expertise among introductory physics students and physics graduate students [Mason and Singh, 2011]. The materials consisted of 10 multiple-choice conceptual physics problems covering various topics in introductory physics. For an example, see Figure 1. Each problem contained a diagram with a thematically-relevant visual component that students needed to attend to in order to answer correctly. These problems also contained areas consistent with naïve conceptions documented in physics education literature [McDermott and Redish 1999].

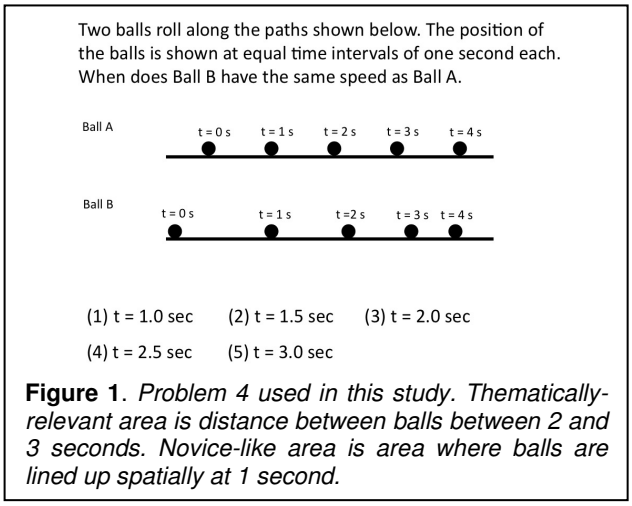

Figure 1. *Problem 4 used in this study. Thematically-relevant area is distance between balls between 2 and 3 seconds. Novice-like area is area where balls are lined up spatially at 1 second.*

The physics problems were presented to participants on a computer screen. Participants used a chin and forehead rest that was 24 inches from the screen. The screen had a resolution of 1024 by 768 pixels and a refresh rate of 85 Hz. Eye movements were recorded with an EyeLink 1000 desktop mounted eye-tracking system, which had an accuracy of less than 0.50° of visual angle. The images subtended 33.3° x 25.5° of visual angle. An eye movement was classified as a saccade if acceleration exceeded 8,500°/s^2 and velocity exceeded 30°/s. Participants' verbal explanations and gestures were recorded with a Flip video camera. Each participant took part in an individual session, lasting 20-40 minutes. At the beginning of the session, participants were given a short explanation of the experiment. The eye tracking system was calibrated to the individual using a nine-point calibration and validation procedure, with a threshold agreement of 0.50° visual angle required to begin the experiment. Next, the participant was instructed to silently answer 10 multiple-choice questions, with their head on a headrest, while their eye movements were recorded. Between questions, drift correction was carried out using the central fixation point to ensure proper calibration. Participants indicated their answer to each question using number keys on the keyboard. Finally, each participant was asked to provide a verbal cued retrospective report [Van Gog 2005] for which they were shown a replay of their eye movements on each problem and they were asked to explain their thought processes. This method has been found to produce more depth of explanation than a retrospective report without viewing one's eye movements. If a participant's explanation was unclear, follow-up questions were asked of him/her. Participants were not given any time limits.

3 Analysis and Results

We used the ScanMatch toolbox for MatLab [Cristino et. al 2010] to compare the scan paths of our participants based on the correctness of their answers given for each problem. The ScanMatch algorithm compares the sequence and durations of fixations in a pair-wise fashion and produces a numerical score representing the similarity of the scan paths both spatially and temporally. A score of one indicates that the scan paths being compared are identical while a score of zero represents no relationship between the scan paths. We calculated ScanMatch scores for three different comparisons of participants' scan paths. The correct-correct comparison (C-C) contained scores comparing each participant who answered a question correctly to one another. The incorrect-incorrect comparison (I-I) contained scores comparing each participant who answered a question incorrectly to one another. Finally, the correct-incorrect comparison (C-I) contained scores comparing those who answered correctly to those who answered incorrectly. We then completed a one-way ANOVA[1] comparing the ScanMatch scores of the C-C comparison, I-I comparison, and C-I comparison for each problem. When we obtained a significant result, we used post-hoc contrasts to determine which comparisons contained a significant difference. We then referenced the mean score values for each comparison to determine the direction of this difference. When homogeneity of variance was violated, we used the Games-Howell test for the post-hoc contrasts, otherwise we used Tukey's HSD test for the contrasts. In the previous study [Madsen et al. 2011] for which this analysis is a follow-up, the eye movements of only six of the 10 problems participants viewed were analyzed. This is because we found that four of the problems did not contain a consistent novice-like area of interest. On those four problems, participants who answered incorrectly reasoned from a wide variety of areas in the problem diagram. Without a precise definition for the novice-like area of interest, these problems could not be included in the original analysis. This scan path analysis is a follow-up on the previous analysis, so we analyze only those six problems included in the original study.

We found statistically significant main effects on three of the six problems tested (Table 1). On problem 1, the ANOVA showed a statistically significant main effect of comparison, $F(2,220)=7.324$, $p=.001$. The contrasts revealed that the I-I comparison had significantly higher ScanMatch scores than the C-I comparison ($p<.001$). Problem 2 also showed significant main effect of comparison, $F(2,250)=6.308$, $p=.002$. The contrasts showed that the I-I comparison ($p<.001$) had a higher ScanMatch score than the C-I comparison. Further, the I-I comparison had a significantly higher score than the C-C comparison ($p=.005$). A significant main effect was also found for problem 10, $F(2,273)=3.583$, $p=.029$. On this problem, the I-I comparison had a significantly higher ScanMatch score than the C-I comparison ($p=.05$). There were no differences found between comparisons on problems 3, 4 and 7.

[1] When using the one-way ANOVA, we recognize there may be issues with the homogeneity of variance because of the unequal sample sizes between correct and incorrect responders. For this reason, we used corrected post-hoc contrasts (Games-Howell test) when this assumption was violated. Further, we employed a non-parametric procedure [Feusner and Lukoff, 2008] to confirm the ANOVA results and found general agreement.

Problem	Comparison	Mean	SD (+/-)
1* (n=11 correct n=11 incorrect)	C-C (n=47)	.396	.068
	I-I (n=55)	.414	.056
	C-I (n=121)	.370	.080
2* (n=14 correct n=10 incorrect)	C-C (n=90)	.330	.151
	I-I (n=36)	.413	.047
	C-I (n=127)	.371	.119
3 (n=17 correct n=7 incorrect)	C-C (n=137)	.351	.093
	I-I (n=21)	.400	.108
	C-I (n=119)	.364	.100
4 (n=14 correct n=9 incorrect)	C-C (n=90)	.379	.088
	I-I (n=35)	.398	.055
	C-I (n=126)	.362	.088
7 (n=15 correct n=9 incorrect)	C-C (n=105)	.312	.125
	I-I (n=36)	.311	.119
	C-I (n=135)	.298	.112
10* (n=11 correct n=13 incorrect)	C-C (n=55)	.333	.086
	I-I (n=78)	.368	.091
	C-I (n=143)	.340	.078
*this indicates a significant difference at the p=.05 level			

Table 1. *Mean ScanMatch score for C-C, I-I, and C-I comparison for each problem used in the study.*

Figure 2 shows a box and whiskers plot comparing the ScanMatch scores of each group averaged over the problems in Table 1.

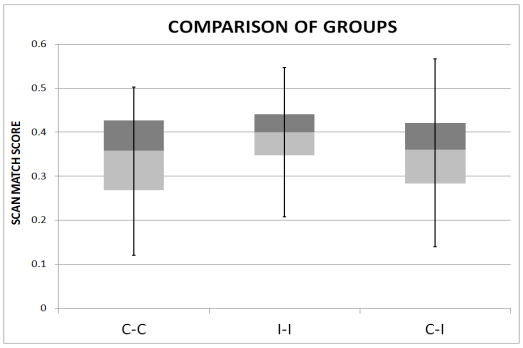

Figure 2. Box and whiskers plot showing the median, max, min and 1st and 3rd quartile of the ScanMatch scores for each group.

4 Conclusion

We did not find significant differences in ScanMatch scores between those in the C-C comparisons and those in the I-I comparisons on five of the six problems analyzed in this study. This evidence is consistent with the hypothesis that the attention of incorrect solvers is primarily directed by top-down naïve theories and not the relative perceptual salience of the elements. This finding aligns well with our previous findings [Madsen et al. 2011] that showed no significant difference in the percentage of fixation time in the perceptually salient areas of the diagram during the full problem period, or the first two seconds of viewing the diagram, when the effects of perceptual salience should be most pronounced. It also aligns well with the findings showing significant differences in the percentage of time incorrect solvers spent in the novice-like areas of the diagram and the percentage of time cor-

rect solvers spent in the thematically-relevant areas of the diagram.

We found significant differences between the I-I and C-I comparisons on three of the six problems. These differences were expected as we have previously seen that correct solvers and incorrect solvers spend different amounts of time looking at thematically-relevant and novice-like elements in the problem, so their scan paths scores are likely to be different. It is curious that we did not find that the I-I comparison and the C-C comparison had higher ScanMatch scores than the C-I comparison on all of the problems. The problems used in the study included a text problem statement, diagram, and multiple-choice answers. The hypotheses set forward in this study assumed a similar reading pattern of the problem statement and answer choices for all participants. The hypotheses were formed assuming only differences in how the participants looked at the diagram. Differences in reading the problem statement and answer choices may have overwhelmed small differences in diagram viewing, resulting in no difference in the ScanMatch scores of the C-C and I-I comparisons compared to the C-I comparison.

These findings may have implications for educational interventions aimed at helping novices learn to answer such conceptual questions correctly. Researchers in physics education have devoted much attention to addressing these consistent wrong answer patterns by changing the way students think about how the world works. If it were true that this problem had an underlying perceptual component, these interventions would need to instead help students learn how to ignore salient elements and focus instead on thematically-relevant elements. The results of this study suggest that wrong answers have roots in the incorrect ways students think about how the world works, not how a problem diagram looks. So it seems that the educational interventions used to improve student understanding are on the right track.

5 Limitations and Future Work

The manner in which participants read the problem statement and answer choices may be interfering with our goal of looking for differences in scan paths while viewing the diagram specifically. To address this issue, this work will be repeated with the text and diagram on two separate slides, which can be toggled between by pressing a button on the keyboard. In this new setup, the scan paths of the participants' first view the diagram can be compared to one another to look for influences of visual salience or naïve theories. Additionally, further studies will not use multiple-choice problems, as we have seen some participants rely on a strategy of eliminating distracter answer choices instead of reasoning through the problem on their own. Instead, participants will indicate when they are ready to answer and will give a verbal explanation of their answer and reasoning. Further, the physics topics covered in these problems are limited. It would be useful to expand the number of topics covered by using a larger variety of problems. This will allow us to determine if the conclusions drawn from this work are context-dependent or generalizable to a wider range of physics problems.

More importantly, follow-up studies will explore the hypothesis that cueing students' while they look at physics problems will improve their accuracy in solving them. Because our previous work [Madsen et al. 2011] has shown that those who answer such questions correctly look at the thematically-relevant AOIs more than the novice-like AOIs, we can test the hypothesis that cueing students to look at the thematically-relevant areas will improve the accuracy of their answers, and that doing this repeatedly will improve their accuracy on conceptually similar transfer problems.

References

CARMI, R. AND ITTI, L. 2006. Visual causes versus correlates of attentional selection in dynamic scenes. In *Vision Research,* vol. 46, 4333.

CARMICHAEL, A., LARSON, A., GIRE, E., LOSCHKY, L. AND REBELLO, N.S. 2010. How Does Visual Attention Differ Between Experts and Novices on Physics Problems? In *AIP Conference Proceedings* vol. 1289, 93-96.

CRISTINO, F., MATHOT, S., THEEUWES, J. AND GILCHRIST, I.D. 2010. ScanMatch: A novel method for comparing fixation sequences. In *Behavior Research Methods* vol. 42, 692.

DOCKTOR, J.L. AND MESTRE, J.P. 2011. A Synthesis of Discipline-Based Education Research in Physics. White paper commissioned by the *National Research Council.*

FEUSNER, M. & LUKOFF, B., 2008. Testing for statistically significantly differences between groups of scan patterns. In *Proceedings of the 2008 Symposium on Eye Tracking Research.* 43-46.

HECKLER, A.F. 2011, The Ubiquitous Patterns of Incorrect Answers to Science Questions: The Role of Automatic, Bottom-Up Processes. In *Psychology of Learning and Motivation: Cognition In Education*, J. P. Mestre and B. H. Ross Eds. Academic Press, Oxford, UK, 227-268.

ITTI, L. AND KOCH, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. In *Vision Research* vol. 40, 1489-1506.

MCDERMOTT, L.C., ROSENQUIST, M.L. AND VAN ZEE, E.H. 1987. Student difficulties in connecting graphs and physics: Examples from kinematics. In *American Journal of Physics* vol. 55, 503.

MCDERMOTT, L.C. AND REDISH, E.F. 1999. Resource letter: PER-1: Physics education research. In *American Journal of Physics* vol. 67, 755.

MADSEN, A., LARSON, A., LOSCHKY, L., AND REBELLO, N. S., 2011. Differences in visual attention between those who correctly and incorrectly answer physics problems. In *Physical Review Special Topics Physics Education Research,* submitted.

MASON, A. AND SINGH, C. 2011. Assessing expertise in introductory physics using categorization task. In *Physical Review Special Topics Physics Education Research* vol 7, 020110.

TROWBRIDGE, D. AND MCDERMOTT, L. 1980. Investigation of student understanding of the concept of velocity in one dimension. In *American Journal of Physics* vol. 48, 1020.

VAN GOG, T. 2005. Uncovering expertise-related differences in troubleshooting performance: Combining eye movement and concurrent verbal protocol data. In *Applied Cognitive Psychology* vol. 19, 205.