Domain-specific video game play as training for decision making

Lucy Parrington\textsuperscript{a}, Lisa Wise\textsuperscript{b}, and Clare MacMahon\textsuperscript{a}

\textsuperscript{a}Department of Biomedical and Health Science, Swinburne University of Technology, Hawthorn, AUSTRALIA; \textsuperscript{b}Department of Psychological Sciences, Swinburne University of Technology, Hawthorn, AUSTRALIA

Decision making in team sports can be used as a proxy for decision making in other time-constrained and pressure-based tasks involving goal-oriented, team-based behaviour. One method for training and evaluating decision-making skills involves the use of video-based decision-making tasks in which video clips are used to simulate the decision space, and participants are required to indicate the form of action they would take (e.g., kick, pass to player x, hold on to the ball). Both the production of the testing protocol and the analysis of decision quality inherent in these paradigms require human factors considerations. For example, use of off-the-shelf sports-specific video games in lieu of video footage may offer a strategy to reduce production labour for training tools, while also offering expanded opportunities to study the development of decision-making skills. The main focus of our research was to evaluate the potential utility of a short-term (30 minute) domain-specific video-game intervention on decision-making performance. While our results provided preliminary support for improved sport-specific decision-making, this paper highlights methodological issues pertaining to the assessment of decision quality in video-based decision-making tasks. We discuss the role of specific research questions and future use of data in driving the selection of data analysis procedures.

Practitioner Summary: This paper highlights issues relating to methods of automating the process of quantitatively evaluating decision-making performance. We promote to the practitioner that determining the purpose of the evaluation should be the most important factor in determining the analysis strategy. Automated methods produce prompt results but potentially exclude viable data, which can be perceptually linked with correct decisions. Conversely, qualitative visual inspection can provide a valuable interpretation of the data, but may introduce too much subjectivity when used alone. In addition to methodological issues, we discuss the use of domain specific video games as a method of training, which may be applied in situations where the use of decision-training tools cannot be directly implemented. Whilst the focus of this research is within the sports domain, this video-based decision-making paradigm in conjunction with video-game play has potential to be used in training and performance evaluation in other areas of expertise. Of note for the practitioner is that quantitative analyses do not always provide definitive solutions in evaluating “correct decisions”, and should not be used in place of expert analyses in new data sets and new contexts.

Keywords: video-game training, expertise, decision making, decision analysis

1. Introduction

Decision making is a key perceptual-cognitive skill in many complex tasks and domains of functioning. The cognitive challenges faced in fast-paced team sports such as hockey, football and basketball can be used as a proxy for decision making in other time-constrained and pressure-based tasks involving goal-oriented, team-based behaviour. The evaluation of decision making in team sports (e.g. Australian football, soccer, netball) has often involved video-based protocols, with researchers attempting to quantify decisions using measurements such as the speed at which the decisions are made (decision latency) and the quality of the decisions (decision accuracy) through varying methods from mouse click selections, to verbal reports, to action-based responses (e.g. Farrow, 2010; Lorains, Ball, & MacMahon, 2013). Prior research has provided a valuable contribution to the sports decision-making assessment and training literature and, findings in this area demonstrate that video-based decision training paired with feedback and error-correction, improve decision-making accuracy (Lorains et al., 2013).

Despite considerable advances in the field, the quantification of decision accuracy still poses a range of methodological questions with regards to extracting and representing expert decision options for different
scenarios, which are then used to code the accuracy of player decisions. For example, the use of sets of rectangular coordinates to outline valid ‘scoring-areas’, such as a box around a player identified as the best option to pass to, can provide a basic representation of an expert’s choice of action, but finer grained options such as whether to place the pass in front of the player, or directly into their hands would require more sophisticated action representation. However, even when designated scoring areas do not fully reflect the expert’s choice of action, it is likely that they will still provide sufficient cues to help a less experienced, yet knowledgeable person score decisions. Being able to automatically score selections that fall within defined spatial locations can be very helpful in promptly processing results and can reduce the need for an expert to independently analyse each decision. On the other hand, using only an automated procedure may overlook potential decisions that anticipate the upcoming play, rather than reflect the action at the point of the freeze frame, or that sit slightly outside of the defined response area but nevertheless may still be reflective of an appropriate choice. Qualitative analyses can provide a means to assess the effects of changing the shape of decision selection areas decision scoring and offers a valuable intermediary step in analysis. Finding ways in which qualitative analyses can be expedited or paired with automated reports may prove useful in both a research context and in practice.

We explore the problem of finding appropriate ways of scoring decision responses through the assessment of a short-term video-game intervention, which is being evaluated as part of a larger study. The aim of this paper is to demonstrate the benefits of qualitative and visually descriptive analyses of decision accuracy combined with automated quantitative methods when assessing sports decision making through video-based testing protocols.

2. Method

2.1 Participants

Forty participants were recruited from the 2014 National Inline Hockey Championships. Two participants were excluded due to incomplete testing. Data were collected on 26 males and 12 female participants spread across three age groups, including youth (12 – 16 years; N=16), junior (17 – 21 years; N=11) and senior (> 22 years; N=11) divisions. For the purpose of preliminary analyses of the data, and as inline hockey is a non-contact sport with equivalent rules played across age and gender competitions, data were collapsed across groups.

Participants completed a short-term video-based decision-making task both before (Pre-test) and after (Post-test) undertaking 30 minutes of video-game activity. The experimental group played a domain-specific sport based video game (NHL14, EA Sports, Burnaby, Canada) and the control group played a first person shooter game (Battlefield 3, Electronic Arts, Redwood City, USA). The control game was chosen to provide a similar level of action and arousal in participants, without associated sport-specific content. The project was granted ethical approval by the Human Research Ethics Committee of the university. Participation was voluntary, with each participant, or both participant and parent or legal guardian (if under 18 years of age) providing informed consent prior to testing.

2.2 Task

The video-based decision-making task was created using Inquisit 4.0.6.0 (Millisecond Software, Seattle, USA) experimental design software. The task included five warm-up video clips of inline hockey game play, followed by 16 trial clips of standard game situations. Prior to data collection, examples of possible selections (pass, shoot, keep possession, move puck to space) were provided visually in Inquisit and paired with written instructions.

Testing clips were sourced from open access game footage from World Championship and North American Championship games and all included aerial views of the rink. Four coaches (two International and two state level) consulted with the lead researcher and individually viewed each clip and then rated available play options, which included passing, shooting, keeping possession or moving the puck to space on the rink. These options were rated 3 (best), 2 (second best), to 1 (third best) option. As per Lorains et al. (2013), where coaches differed on ratings, the selection of the rating was based on the majority answer.

Participants were required to watch a short period of game play and were asked to indicate what they would do as quickly as possible by clicking the mouse on the region of the screen once the clip froze at a
decision point. Having made this decision, the participant was then asked to rate their confidence in their choice on scale of 1 (least confident) to 5 (most confident). The presentation of trial clips was randomized.

2.3 Data analysis

Data collected through Inquisit were in the form of x and y coordinates of the mouse-click to be compared with the expert coaches choice of plays (region of screen identified and rated as first, second or third choice plays), confidence in their decision, and decision latency (in milliseconds) from the time at which the video clip was frozen to the time at which their decision was made.

Data were exported into both SPSS 22.0 (IBM, Armonk, USA) and R Studio 0.98.1103 (R Studio Boston, USA) to complete complementary statistical analyses. Both datasets were first cleaned to remove ‘timed-out’ responses (>3000ms) for consistency with previous sports decision-making testing protocols (e.g. Lorains et al., 2013a,b). Statistical analyses were run on both raw and cleaned latency data. Averages per participant were calculated and then the distribution of age categories, and pre-test measurements were assessed to ensure no statistical differences were found between the experimental and control group. Repeated measures t-tests were used to assess differences in decision latency, whilst decision accuracy and decision confidence were assessed using a Wilcoxon Signed Ranks test.

Further graphical analysis in R involved plotting the decision location (coordinates) and confidence level against the final freeze-frame of each trial video clip. In addition to this, the outlines of the expert coaches choices were highlighted to provide a visual representation of the regions considered when automated accuracy scoring was used. Each outlined coach-scored area could be reconfigured interactively to include as many correct visually-assessed selections as possible. For example, as will be discussed with respect to Figure 6, a number of selections fall just outside of the scoring box, but by visual assessment appear to represent correct decisions. Each edge of the scoring boxes for each video could be interactively extended to include the “correct” responses, and this manipulated scoring area was then used to update all data and statistics, which included the additional selections (data represented as manipulated scoring location).

3. Results

3.1 Latency Data

Figure 1 shows latency data for pre test and post test conditions for all participants (top left) and for each individual participant (bottom left).

![Latency data](image)

Figure 1. Box plots of pre- and post-test latency data for raw and cleaned data (latency<3000ms).
It can be seen from this left panel that the majority of responses occurred within 3000 ms. Graphs in the right panel show the pattern of results to be similar for pre-test versus post-test conditions, and across participants when data is “cleaned” by removing latencies of greater than 3000 ms.

Both experimental and control groups responded faster overall in the post-test condition (experimental group $M$(Pre-test) = 3303 ± 1266 ms, $M$(Post-test) = 2540 ± 1254 ms; control group $M$(Pre-test) = 3110 ± 1122 ms, $M$(Post-test) = 2524 ± 820 ms) when the raw data were assessed. When the cleaned data was assessed, response latencies were shorter in the post-test condition of the experimental group only (experimental group $M$(Pre-test) = 1951 ± 325 ms, $M$(Post-test) = 1731 ± 384 ms; control group $M$(Pre-test) = 1887 ± 451 ms, $M$(Post-test) = 1807 ± 383 ms, see Table 1), suggesting that much of the improvement in the control group was through reduction in the number of extremely long responses.

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>Cleaned</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experimental</td>
<td>Control</td>
</tr>
<tr>
<td>$M$</td>
<td>763.48</td>
<td>585.94</td>
</tr>
<tr>
<td>$SD$</td>
<td>994.26</td>
<td>810.42</td>
</tr>
<tr>
<td>$95%$ CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Lower$</td>
<td>310.89</td>
<td>169.26</td>
</tr>
<tr>
<td>$Upper$</td>
<td>1216.06</td>
<td>1002.62</td>
</tr>
<tr>
<td>$t$</td>
<td>3.519</td>
<td>2.981</td>
</tr>
<tr>
<td>$df$</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>$p$</td>
<td>0.002**</td>
<td>0.009**</td>
</tr>
</tbody>
</table>

**Significant at the 0.01 level

3.2 Confidence Data

Figure 2 shows the decision latencies as a function of confidence for all data and cleaned data respectively. Participants were least confident of their fastest decisions (latencies<1300 ms) suggesting that these decisions may have been self-identified errors in responding. The pattern of response latencies for confidence scores from 2 (low confidence) to 5 (most confident) showed that participants tended to be more confident of their faster decisions. Although decision latency decreased in the post-test condition, decision confidence did not change between pre and post-tests for either the experimental group ($Mdn$(Pre-test) = 4.3, $Mdn$(Post-test) = 4.3, $Z$ = 1.12, $p$ = 0.264) or the control group ($Mdn$(Pre-test) = 4.3, $Mdn$(Post-test) = 4.4, $Z$ = 1.05, $p$ = 0.293).

Figure 2. Decision latency as a function of confidence for all data, and for cleaned data (latency<3000ms).
3.2 Analysis of decision accuracy

Visual representation of changes in the location of the decision for each participant between pre-test and post-test are presented in Figures 3 and 4 for experimental and control groups, respectively. Only decisions occurring within 3000 ms (cleaned data) were included in this analysis. There is no particularly obvious pattern of results distinguishing the two groups, but it should be noted that the grey lines indicate that many participants in both conditions changed their selection between pre-test and post-test.

![Figure 3](image1.png)

Figure 3. Locations of decisions for each participant (experimental group) for each trail (Clip 1 to Clip 16). Lines connect the decision of each participant in the pre-test (red o) and post-test (blue x) condition.

![Figure 4](image2.png)

Figure 4. Locations of decisions for each participant (control group) for each trail (Clip 1 to Clip 16). Lines connect the decision of each participant in the pre-test (red o) and post-test (blue x) condition.
Decision accuracy was scored using up to three selection boxes identifying appropriate decision options. These options were scored as 3 for the best option (blue box), 2 for the second-best option (red box) and 1 for the third-best option (green box) according to options provided by domain experts. Figure 5 shows an example of a scoring scenario where all decisions fall within an identified scoring box. Note that in the post-test condition, one decision has changed to the third-best option, resulting in a lower accuracy score. Furthermore, while all scored decisions fell within scoring boxes, the accuracy score is not 100% for the pre-test condition because decisions that took longer than 3000 ms were scored as 0. The data are presented without a background image for ease of seeing the responses.

![Figure 5](image1.png)

Figure 5. Example of all player selections, and scoring locations signified by coloured scoring boxes (blue = 3 points, red = 2 points, green = 1 point) for pre-test (left) and post-test (right).

In contrast to the scoring scenario of Figure 5, the top two panels of Figure 6 show a scoring scenario where a number of decisions fall just outside the blue scoring box, but visual inspection of the data suggests these decisions represent an appropriate selection of the correct option. The bottom two panels of Figure 6 show an extended scoring box that has been adjusted post hoc to include these data points as correct.

![Figure 6](image2.png)

Figure 6. Example of all player selections, and scoring locations (original = top and manipulated = bottom) signified by coloured scoring boxes (blue = 3 points, red = 2 points, green = 1 point) for pre-test (left) and post-test (right) displayed on the last frame of video for pre-test (left) and post-test (right).
Decision accuracy was assessed using the original strict scoring boxes as well as the scoring boxes manipulated post-hoc to include as many data points as possible that were deemed correct on visual inspection. Decision accuracy was found to be significantly higher in the post-test of the experimental group, while no significant differences were found for the control group after 30 minutes of video-game play. This finding was represented in both original scoring and manipulated scoring datasets (see Table 2). It should be noted that while there is a medium effect size for the experimental group using both scoring methods, there was a small to medium effect size for control group even though it did not reach statistical significance. It may be that the small difference in the number of participants per group (N = 21 experimental, N = 17 control) did not provide adequate power to the control group to register significance.

Table 2. Decision accuracy for both original and manipulated scoring locations.

<table>
<thead>
<tr>
<th></th>
<th>Pre Mdn</th>
<th>Pre Range</th>
<th>Post Mdn</th>
<th>Post Range</th>
<th>Z</th>
<th>r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original scoring locations</td>
<td>Experimental</td>
<td>43%</td>
<td>90%</td>
<td>47%</td>
<td>70%</td>
<td>2.35</td>
<td>0.38</td>
</tr>
<tr>
<td>Controlled scoring locations</td>
<td>Control</td>
<td>37%</td>
<td>60%</td>
<td>37%</td>
<td>70%</td>
<td>1.05</td>
<td>0.17</td>
</tr>
<tr>
<td>Experimental scoring locations</td>
<td>Experimental</td>
<td>43%</td>
<td>70%</td>
<td>63%</td>
<td>67%</td>
<td>2.75</td>
<td>0.45</td>
</tr>
<tr>
<td>Control</td>
<td>40%</td>
<td>60%</td>
<td>53%</td>
<td>60%</td>
<td>1.92</td>
<td>0.31</td>
<td>0.055</td>
</tr>
</tbody>
</table>

*Significant at the 0.05 level; **Significant at the 0.01 level

4. Discussion

4.1 Efficacy of domain-specific video game intervention

The aim of the study was to assess the efficacy of short-term exposure to domain-specific video-game play in improving decisions. The specific focus of this paper was to explore a number of methodological considerations regarding the use of quantitative, qualitative and visually descriptive analyses of decision accuracy, used to assess sports decision-making performance in video-based protocols.

Decreased decision latency and increased decision accuracy in the experimental group provide preliminary support for the idea that video-game play can improve components of decision making. These results align with previous decision-making research, which has demonstrated improvements in decision-making accuracy after a video-based decision-training intervention (Lorains et al., 2013). Though participation in video-game play is not specifically a form of training, it is possible that some tactical knowledge is gained through the process of playing domain specific games because of related content embedded in the games. Previous video-game studies have demonstrated links in improved cognitive abilities associated with gaming in general (e.g., Green & Bavelier, 2012; Oei & Patterson, 2013; but for an alternative perspective see Boot, Blakely, & Simons, 2011). Thus, there was some level of expectation that small changes would occur in both the experimental and control groups (domain-specific game play versus non-specific action-based game play). As previous research findings suggest team sports expertise may be associated more with sport-specific knowledge rather than general cognitive ability (Memmert, Simons, & Grimme, 2009), the larger effect size elicited in the experimental group may have been linked to the sport-specific content of the game-play intervention.

4.2 Methodological issues

Measurement of decision-making quality raised a number of methodological issues. Though quantitative analysis was sufficient to assess decision latency, analysis of accuracy required greater depth of interpretation. Furthermore, while interpretation of decisions through visual inspection appears to be intuitively easy, the quantitative classification and scoring of decisions presents challenges even when expert scoring criteria are available and agreed upon.

As can be seen from the example provided in Figure 5, participants were able to identify appropriate options that fell within the designated scoring location. However, Figure 6 demonstrates how a number of selections fell just short of the designated scoring box but still appeared to correspond to the correct choice of action represented by the box. This highlights how a deliberate and valid selection may not be correctly classified if automated processes are the only analysis method implemented and suggests that the use of automated procedures for assessing decision accuracy is likely to be sub-optimal without the use of visual validation.
The location, extent and shape of the “correct” region will depend on a number of factors, which require thoughtful consideration from domain experts. Yet visual inspection alone may lead to excessive subjectivity in scoring in addition to increased evaluation time. A trade-off is required between analysis tools that provide data quickly but may simplify the nature of the result, versus more analytically complex procedures which exceed useful feedback periods for results to be returned in the field. Thus, a pairing of the analysis methods may be warranted, whereby the protocol or tool is able to make use of data-driven optimisation of scoring, and to provide different analysis options for those interested in practical outcomes versus those interested in theoretical interpretation. For example, the decision to trim latency data at 3000 ms may be driven by specific theoretical interpretations of decision-making processes, even though there were few differences in the outcomes of statistical analyses based on all the data versus the “cleaned” data. However it may also be the case that, while faster and slower methods of decision making employed by participants are equally accurate, decision latency would affect the ability to transfer training from the video-based task to the real world, and thus theoretical considerations may be important to future practically-focused contexts.

4.3 Future directions

A number of future directions regarding video-based decision-making analysis methods are proposed. The first is the development of automated scoring areas based on complex and irregular shapes, rather than rectangles or squares. The second is to implement gradient mapping of the screen coordinates associated with decision selections of domain experts. For example, scoring could be weighted by the distance from the identified decision hot-spot. Thirdly, follow-up studies should also be conducted to further establish whether decision-making expertise can be gained from longer-term exposure to domain-specific video-game play.

5. Conclusions

Results indicate that domain specific video-game play can potentially provide a mechanism for learning domain specific information. Qualitative analyses based on visual inspection of decision-making data using interactive analysis tools was found to be a rapid way of optimising automated scoring of decision accuracy. The combination of qualitative and quantitative methods of assessing video-based decision making resulted in an efficient and effective evaluation tool, which could be fine-tuned to the specific context of the evaluation and would allow for rapid communication and dissemination of practice-based information back to the field.

Acknowledgements

Thankyou to both Skate Australia and Skate Australia Victoria Inline committees and players for their support.

References


