Serious Games Need Serious Performance Measures: A Case Study

Susannah J. Whitney; Philip Temby; Ashley Stephens
Defence Science and Technology Organisation
[susannah.whitney, philip.temby, ashley.stephens]@dsto.defence.gov.au

Abstract. Serious games are increasingly being used to support military training. One important instructional feature of these games is their AAR capabilities, providing data such as summary performance statistics. Clearly, if such data is to be of use to military training audiences, it is important that performance statistics are an accurate and valid measure of trainee performance. However, there is very little data on the utility of AAR summary statistics, particularly in the context of how they compare to traditional methods of training evaluation such as Subject Matter Expert (SME) assessments. This paper describes a case study from the authors’ research into the training effectiveness of serious games. In this study, a group of dismounted infantry soldiers undertook training in small unit operations with a serious game. Their performance was measured in two ways: using the game’s AAR data capture tool, and by SME assessment against pre-defined criteria. Examination of the results demonstrated that the SME data showed clear trends across the training sessions, whereas the AAR data showed no clear trends. For the particular task being trained, the SME measures of team processes were more appropriate and the AAR data were of limited use for evaluating team performance. In this paper, we discuss the ways in which the two different sets of data can best be used to assess individual and team performance, depending on the particular task being trained. We also provide recommendations for improvements to AAR data capture capabilities, incorporating measures of team processes and outcomes.

1. INTRODUCTION

Serious games\(^1\) are increasingly being used to support military training. This is due to several factors, including the potential cost savings of supplementing live training with game-based training, and improved training capabilities associated with the ability to customize, record and evaluate training events.

One important instructional feature of these games is their After Action Review (AAR) capability, which can provide data such as summary performance statistics. Clearly, if such data is to be useful to military training audiences, it is important that these performance statistics are accurate and valid measures of trainee performance. However, little research has examined the utility of AAR data, particularly in regard to how it compares to traditional measures of training evaluation such as Subject Matter Expert (SME) assessments.

This paper describes a case study comparing the utility of summary performance data captured by an in-game AAR tool (or ‘AAR data’) with SME assessment of team performance (or ‘SME data’) during game-based training in a military context. The paper begins with a brief overview of previous research in the area, the rationale for the current study, and then presents the methodology and results of the case study. This is followed by a discussion of the results, and recommendations for improvements to performance measures in the area of serious games.

\(^1\) As defined by Morrison, Parr and Ward (2007), a ‘serious game’ is a game used for non-entertainment purposes, such as training.
automated performance measures such as the number of rounds fired per enemy death, and the number killed on each side, to compare the effectiveness of infantry sections of different sizes. However, Barlow et al. (2004) did not comment on whether these AAR data were useful performance measures in the context of their study.

More recently, the utility of automated performance measures was explored by Hussain and Feurzeig (2008). They reviewed the type of information provided by the AAR capabilities of various serious games, and concluded that these tools are currently insufficient for accurate assessments of individual and team performance. However, these authors’ claims have yet to be empirically tested.

In summary, while some researchers consider automated performance measures (and associated AAR data) to be useful, and have used them to measure team performance, other researchers have criticized them for their lack of usefulness. Furthermore, previous research has mainly focused on their utility for experimentation rather than training purposes. Consequently, there is a need for further research into the application of automated performance measures in the area of serious games for training; this became the driver for our case study.

3. CASE STUDY

3.1 Overview

The following case study is based on the authors’ research into the training effectiveness of serious games being conducted at the Australian Defence Science and Technology Organisation (DSTO). The data presented in this paper are from a larger study (conducted in September 2008), which involved comparing training outcomes for small infantry teams undertaking either traditional instruction or game-based training with the serious game Virtual Battlespace 2 (VBS2). In this paper, two sets of data from the game-based training group are discussed, namely performance data from the AAR tool and SME assessment data.

3.2 Participants and method

All participants were members of the Australian Army. A nine-man infantry Blue Force (BLUEFOR) section was formed, comprising an experienced section commander, and eight novice infantry soldiers. The eight novice soldiers had an average of less than four months military experience; they had only recently graduated from basic training. None of the section members had previously played VBS2, but eight of them played similar first-person shooter games on a regular basis.

The section received eight hours of game-based training on section attack procedures, with an emphasis on cognitive and teamwork skills (Figure 1). The section attack is a fundamental task for infantry teams, and relies on good teamwork skills, such as leadership, communication, and situation awareness to achieve the objective of eliminating the enemy. Each member of the section had a different role (e.g., scout for information collection, machine gunner for fire support), even though they were working as a team. The training was conducted by the section commander, under the supervision of an SME, who was a Warrant Officer with 20 years of military experience.

To add realism and competition during the missions, the section played against a two to three-man opposing force (OPFOR) who, in turn, were played by experienced military personnel. A total of six missions were conducted during training. While each mission was slightly different, the primary objectives were to eliminate the OPFOR and to work as a team.

Figure 1: Participants during game-based training

The SME assessed the section’s performance using a 25-item checklist, which contained five categories of behavior: Communication, Leadership, Coordination, Supporting Behavior, and General Points. Example items included: passage of information during assault (communication), group commander awareness of troop positions (leadership), fire and movement effective (coordination), reaction of troops to directions (supporting behavior), and suitability of fire positions (general points). Each item was assessed on a 0-5 rating scale, where 0 = not undertaken/not attempted, 1 = very poor, 2 = poor, 3 = satisfactory, 4 = good, and 5 = very good. The checklist items and rating scale were based on those currently used to assess section attack in the Australian Army. The AAR tool in VBS2 automatically recorded the following data for each mission: rounds fired, enemy killed, enemy wounded, friendly killed, friendly wounded, and hit ratio\(^3\) for BLUEFOR and OPFOR. These data were then transcribed for analysis. The AAR tool also records a 2D and 3D view of in-game events. While these recordings were not formally analyzed, they were used by the section commander when conducting his AAR after each mission.

\(^2\) Entry to the Royal Australian Infantry Corps is restricted to males only, hence the use of gender-specific language.

\(^3\) The hit ratio is a function of the number of enemy and friendly force casualties divided by the total number of rounds fired by the force of interest.
3.3 Results

3.3.1 AAR data

Figure 2 shows the percentage of BLUEFOR (white bars) and OPFOR (black bars) fatalities in each mission. It is clear that BLUEFOR suffered the highest number of casualties in Missions 2 and 6, and the least casualties in Mission 4. The only mission where all OPFOR were eliminated was Mission 3.

![Figure 2: Percentage of fatalities in each mission](image2)

Figure 2: Percentage of fatalities in each mission

Figure 3 shows the number of rounds fired in each mission for BLUEFOR (white bars) and OPFOR (black bars). A logarithmic scale is used as the number of rounds fired in Mission 4 was at least one order of magnitude greater than any of the other missions.

![Figure 3: Number of rounds fired in each mission](image3)

Figure 3: Number of rounds fired in each mission (Y-axis scale is logarithmic)

The hit ratio for BLUEFOR (white bars) and OPFOR (black bars) are shown in Figure 4. Higher hit ratios indicate greater efficiency in the use of fire. It is clear that BLUEFOR’s hit ratio was highest in Mission 6 and lowest in Mission 4 (where the greatest number of rounds was fired). The high hit ratio in Mission 6 could be interpreted as some improvement in efficiency over the missions, but the percentage of BLUEFOR casualties in this mission is identical to the percentage in the Mission 2.

![Figure 4: Hit ratio for each mission](image4)

Figure 4: Hit ratio for each mission

With regard to Figure 4, the high hit ratio in Mission 6 is due to one OPFOR being killed by one participant with a 100% hit ratio (1 round fired for 1 enemy wounded), and the remaining two OPFOR killed by a participant with a 33% hit ratio (6 rounds fired for 2 enemy killed). Averaging these values gives a hit ratio of 67%. However, this does not reflect the fact that other members of the section fired numerous rounds during this mission without killing or wounding any OPFOR. From these data, there are no obvious trends apparent, nor do the data shown in the three figures appear to be correlated in any way; therefore it is difficult to determine if the section’s performance is improving during training.

3.3.2 SME data

The SME’s ratings for each of the five categories, and the average of all items, are shown in Table 1. There are a number of trends apparent. At the beginning of training, there are clear differences in the starting levels of performance; for example, General Points has a mean score of only 0.2, whereas Leadership is 3.4, and the other behaviors are rated in between. This suggests that the section had higher levels of skill (albeit still rudimentary) in some areas than in others. The higher starting level of Leadership is largely due to the presence of the experienced section commander in the team.

The SME data also present a clear picture of the pattern of improvement across training. In all categories, except

---

4 The high number of BLUEFOR casualties across missions is attributed to problems detecting the location of enemy fire, and general issues with situation awareness. In reality, a high number of casualties would cause a mission to be aborted, but in this study the focus of training was on teamwork processes and thus, the missions were allowed to continue.

5 Except in Mission 4, where the large number of rounds fired was associated with a very low hit ratio; this is unsurprising given that rounds fired is used in the calculation of hit ratio.
Leadership, performance initially increased, then tends to plateau towards the end of training. By the end of training, the differences between scores for each behavior category have almost disappeared.

**Table 1: Team performance scores across missions**

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Mission</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>1.5</td>
<td>3.0</td>
<td>3.0</td>
<td>2.8</td>
<td>3.0</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>3.4</td>
<td>3.6</td>
<td>3.4</td>
<td>3.4</td>
<td>3.2</td>
<td>3.6</td>
<td></td>
</tr>
<tr>
<td>Coordination</td>
<td>2.3</td>
<td>2.3</td>
<td>2.6</td>
<td>2.9</td>
<td>2.9</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Supporting behavior</td>
<td>1.3</td>
<td>1.5</td>
<td>2.0</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>General points</td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
<td>2.8</td>
<td>3.0</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>1.7</td>
<td>2.2</td>
<td>2.4</td>
<td>2.8</td>
<td>2.9</td>
<td>3.0</td>
<td></td>
</tr>
</tbody>
</table>

In order to provide a clearer comparison between the AAR and SME data, the hit ratio data for BLUEFOR (white bars, primary y-axis) and OPFOR (black bars, primary y-axis) were plotted on the same figure as the SME ratings (black line, secondary y-axis). As can be seen in Figure 5, there is no clear relationship between the SME and AAR data.

![Figure 5: Comparison of HR data (primary y-axis) and SME rating (secondary y-axis)](image)

**4. DISCUSSION**

The SME and AAR data paint two very different pictures about the section’s performance during training. The SME data indicate that the section’s performance improved across training, initially quite quickly, and then tapering off. In comparison, the AAR data showed that performance was quite variable across the training missions with no clear trends.

The AAR data also did not demonstrate the complexity of what was occurring during the missions. For instance, in missions where there were a large number of rounds fired (e.g. Mission 4), did this occur because section members were surviving longer (i.e. more opportunity to fire rounds), or did it occur because they were firing rounds more quickly, irrespective of their effectiveness? Clearly, looking at any one of the AAR summary statistics does not give a complete indication of section performance. Looking at them in combination does not tell the full story either.

When the two methods of assessment are compared, it becomes clear that each is measuring different things. Firstly, the AAR data represent individual and team performance (as an aggregate of individual performance), whereas the SME data represent team performance. Clearly, team performance is more than the sum of the performance of a few individuals; one or two high-performing individuals do not necessarily result in higher team scores. Furthermore, aggregating individual performance data is not always appropriate because individual team members will have different roles within the team (and hence their performance is not necessarily comparable using such measures), as in this study.

Secondly, the AAR data provide objective measures about the section’s actions, while the SME data are a subjective evaluation of the effectiveness of the section’s actions in achieving their goal. Researchers refer to these two types of measures as outcome measures and process measures respectively (McAlinden, Durlach, Lane, Gordon, & Hart, 2008; Salas et al., 2009).

Researchers have suggested that both outcome and process measures are necessary when measuring team performance (McAlinden et al., 2008; Salas et al., 2009). For example, Salas and his colleagues note that outcome measures can alert trainees and instructors to areas requiring improvement or correction, although they do not provide diagnostic information about the underlying reasons for performance. In addition, as noted by McAlinden et al., outcome measures can in some cases be used to make inferences about team processes. The analogy they use if that if a trainee takes action towards a particular line of effort, it implies they are thinking about accomplishing that line of effort. The implication here is that if performance measures are valid, then we should expect a good correlation between process and outcome measures.

In this case study, however, there was little correlation between the outcome (AAR) and process (SME) measures. The outcome measures were too generic and simplistic to provide useful information about the processes being undertaken by the team. Furthermore, as we have indicated, the outcome measures were an aggregate of team performance, and in doing so average the contributions of each team member; in an infantry section (as with many teams) such averaging of performance is not wise since the roles of each team member are different and therefore their individual performance might be expected to vary. Although the primary objective of the section attack is to eliminate the enemy using the least amount of ammunition, in this study it was of greater interest to train a team in section attack procedures (and requisite behaviors) using a serious game. The fact that the section was not always able to achieve the objective of eliminating the enemy was due mainly to limitations with the game that made...
it difficult to identify enemy positions in the virtual environment. As a consequence, the more useful measures of team performance for our purposes were the process measures. This was due in part to the use of specific and clearly defined team-behaviors. Therefore, it appears that, in this case study, the AAR data did not support any reliable assessments about team performance during training. This finding is consistent with the views of Hussain and Feurzeig (2008), who suggest that current AAR tools in serious games are insufficient for accurate assessments of individual and team performance.

The findings from this case study suggest there is a need for better objective, automated measures of performance within serious games. While subjective measures are considered by researchers to be useful, they are not without their limitations, including the potential for rating bias and the load they place on SMEs to accurately monitor and assess trainee performance, usually in real-time. With this in mind, we now consider how additional automated measures could be applied in future studies with serious games, using examples from our case study.

Given the clearly defined and specific process measures in our study, if certain additional outcome measures had been present in the AAR capability, they may have been able to better complement the SME’s evaluation of the section. For instance, one of the checklist items in the supporting behavior category was fire support effectiveness. Given that the AAR data includes rounds fired, enemy killed, and enemy wounded, the addition of summary information about the range and direction of fire could enable automated assessment of fire support effectiveness. Another item in this category was reaction of troops to directions, which could potentially be assessed automatically if sufficient information was captured regarding the section commander’s communication, and the subsequent movement of player avatars and other entities within the game environment. Other measures that could be automated include objective measures of spacings suitable and bounds appropriate. Automating these measures could provide researchers and instructors with valuable information about trainees’ performance. In addition, they could be used to provide trainees with real-time feedback, for example by displaying a warning message whenever the length of a bound exceeded a predetermined distance or time. By incorporating more objective measures into serious games (that are relevant to the task being trained) then we can increase the level of confidence in evaluations of training, and provide assessors with data regarding trainee performance that can be reliably used as part of the AAR debrief.

It is worth noting that while better automated measures are needed, it is still difficult to capture the context surrounding events within game-based training. This is where the current playback capabilities of AAR tools can be useful. Indeed, the AAR tool was able to provide information about the section’s performance that the SME data could not. Most notably, during one mission, it initially appeared that one member of the section had inadvertently shot his section commander. However, he did not have any friendly killed or friendly wounded recorded against him in the AAR summary statistics. To examine this in more detail, the 2D AAR playback was analyzed and it was ascertained that the fatal shot had been fired by a member of the OPFOR. This level of detail was not apparent to the SME or the study team during the mission. Clearly, even if better objective measures can be adopted into the AAR capability of serious games, the playback feature is still useful in providing the context for in-game events.

5. CONCLUSIONS AND RECOMMENDATIONS

In this case study, automated performance measures generated from an AAR tool provided little useful information about team performance during training with a serious game, due to the generic and simplistic nature of what was measured.

This conclusion is based on a case study of a small infantry unit conducting section attack missions. However, we believe the findings will generalize to other tasks within the military domain, given that leadership, communications, and situation awareness are critical to most military operations.

While most serious games include AAR tools that provide summary performance statistics of individual and team performance, these measures tend to focus on traditional outcome measures of performance (e.g. engagement outcomes), and fail to provide the fine-grain level of detail required to support training objectives. This does not mean that outcome measures are without utility. Rather, it implies that better (‘more serious’) performance measures are needed if automated data capture tools are to be effective for training purposes.

In order to improve the utility of such automated performance measures, we believe they should be expanded to include additional information that can be used to help assessment of individual and team processes. In particular, we make the following recommendations when using serious games:

- Researchers and instructors should clearly define the type of behavior that they are training, and then develop appropriate measures of these behaviors;
- Researchers, instructors, and industry need to work together to implement these measures into serious games;
- Where possible, industry should ensure that any new performance measures can be added to existing serious games with minimal cost and time; and
- Automated measures of performance should not be used as stand-alone indicators of trainee performance; rather they should be used in
conjunction with SME assessment of trainee performance.

5.1 Final Remarks

Automated data capture tools are perceived to be an important component of serious games. To date, there has been little published research on the utility of automated performance data (AAR summary statistics), particularly in comparison to SME assessments. This case study has provided a significant contribution to the literature by highlighting the need to develop better automated performance measures when using serious games. Although this case study is based on a military example, we believe the principles are applicable to other domains (e.g., health, mining, transport) where serious games are being used for training purposes. In closing, we believe that developing more serious performance measures for serious games will help to improve the overall capability of these training tools, as well as improve their cost-effectiveness.

REFERENCES


