Proceedings of the Twentieth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2024)

Evaluating the Effects of AI Directors for Quest Selection

Kristen K. Yu, Matthew Guzdial, Nathan R. Sturtevant

Computing Science Department, University of Alberta
Alberta Machine Intelligence Institute
{kkyu, guzdial, nathanst}@ualberta.ca

Abstract

Modern commercial games are designed for mass appeal, not for individual players, but there is a unique opportunity in video games to better fit the individual through adapting game elements. In this paper, we focus on AI Directors, systems which can dynamically modify a game, that personalize the player experience to match the player's preference. In the past, some AI Director studies have provided inconclusive results, so their effect on player experience is not clear. We take three AI Directors and directly compare them in a human subject study to test their effectiveness on quest selection. Our results show that a non-random AI Director provides a better player experience than a random AI Director.

1 Introduction

AI Directors (AID) (Booth 2009) - also called player-centric game AI (Charles et al. 2005), experience management (Thue 2010) or drama managers (Yu and Riedl 2013) - have been successfully deployed in commercial games. The most famous example of an AID is in Left 4 Dead (Valve 2008), where the AID maintains a stressful experience for the player by dynamically changing the number of enemies, health packs, and ammunition available in the level (Booth 2009). A natural question arises from this example: "What was the impact of the AID on the player experience?". For the Left 4 Dead AID, we do not know because we do not have comparative data on the player experience with their AID, no AID, or other AIDs. Although we do not conclusively know the impact in Left 4 Dead, we can begin to study the impact of AIDs on player experience in other games.

We may assume there is some positive impact on player experience because AIDs are present in many commercial games (Thompson 2017). Of the studies on AIDs, some have concluded that the AIDs are effective at manipulating the player experience to achieve their desired results (Yannakakis and Hallam 2009; Nygren et al. 2011), while other studies have provided inconclusive results (Thue 2007; Yu et al. 2022). We target these inconclusive AIDs for further evaluation because we want to better understand the strengths and weaknesses of these AIDs. We can consider the impact across two facets: the first being quantitative changes to player behavior, and the second being qualitative evaluations of the players' perception of their experience. The goal of this paper is to design an experiment to evaluate AIDs along these two facets.

One crucial decision for our experiment is choosing the game in which to evaluate the AIDs. Based on previous studies and examples, AIDs are generally designed for a specific game (Thue et al. 2007; Dias and Martinho 2011). This suggests each game requires a unique AID. To address this problem, we choose to implement an AID for a specific system in a game, rather than the complete game. This allows the application of the AID to generalize to any game that uses this system, and we may assume that the findings of the evaluation to hold to other similar games. Specifically, we are interested in quest systems where the player has an option to choose between several quests at a time. We call this problem quest selection, and this design pattern is present in games such as in the Nook Miles+ system in Animal Crossing: New Horizons (Nintendo 2020).

For the experiment we used FarmQuest (Yu, Guzdial, and Sturtevant 2022), a video game test bed for AIDs. We compared PaSSAGE (Thue et al. 2007) and a reinforcement learning based AID (Yu et al. 2022) to a random algorithm. We chose PaSSAGE and the reinforcement learning AID because they had previously inconclusive results. We chose to include a random AID because it is commonly used in commercial games due to its ease of implementation (Yu et al. 2022). We evaluated these AIDs in a human subject study.

Our contributions are two-fold. First, we present an evaluation of previously inconclusive AIDs to further characterize their use. Second, to the best of our knowledge, this is the first attempt at directly comparing two existing academic AIDs on the quest selection problem. Our results demonstrate that there is a quantifiable difference in how players play the game with different AIDs, and there is one measured qualitative difference in how players perceive the differences in playing the game. From these findings, we conclude that a curated AID, either PaSSAGE or the reinforcement learning based AID, performs better on the quest selection problem than random.

2 Background

In the past, there have been successful evaluations of AIDs using human subject studies (Wauck and Fu 2017; Yun et al.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2010) where the authors concluded that their approach was effective at meeting their target goals. However there have been other studies where the authors found their results inconclusive. In this background, we discuss a few specific AIDs which have inconclusive results.

In the study by Thue et al., they proposed PaSSAGE, an AID that changes the story based on the player's previous actions (Thue et al. 2007). Ninety students participated in a human subject study, and they evaluated the hypothesis that PaSSAGE would be more fun and provide more agency to the players than a non-adaptive version of the story. They only collected survey data. They found that female participants rated PaSSAGE higher in fun and agency, with confidence levels of 93% and 86%. They identified these results as inconclusive, as other subgroups did not rate PaSSAGE higher, or had a low statistical confidence interval.

In the study by Dias and Martinho, they implemented a rule-based AID designed to change narrative and other content based on player personality types (Dias and Martinho 2011). Twenty-one males were selected for participation such that the different personality types were roughly equal. Within these groups, they compared an adaptive to a non-adaptive version of the game. They collected telemetry data and survey data for their game, such as difficulty, pace, and immersion. They concluded that their AID showed promise, though all but the immersion metric produced inconclusive results. We did not choose this AID for this experiment because it requires participants to take a Myers-Briggs personality test.

In our previous study, we implemented a reinforcement learning AID that changed which quests to select for a player based on previously accepted quests (Yu et al. 2022). The reinforcement learning AID was compared to a random AID. 208 players played the game, and telemetry data was collected such as time played in game and number of quests presented, accepted, and completed by the player. Survey data was also collected but could not be linked to the AID that the player experienced, so strong conclusions could not be drawn from this data. We concluded that the AID had inconclusive results, because the acceptance rate of quests was not statistically significant, and players spent more time playing the game with the random AID.

3 Experimental Setup

In this section we cover the necessary details to understand our experiment. This includes a formal definition of the quest selection problem, a description of the test bed, and a description of the AID used in the experiment.

3.1 Quest Selection Problem

Quest systems in games take on many forms. One common design pattern is to present the player with a few quests at a time. Skyrim (Bethesda 2011) and Animal Crossing: New Horizons (Nintendo 2020) are two examples of games that have these kind of systems.

We formalize the quest selection problem as follows. In the game, a player p repeatedly sees n quests q at the same time, and there is a maximum number of quests m that p can have at the same time. All q come from the set Q, and q cannot be repeated. p views $\langle q_1, q_2, ..., q_n \rangle$ at the same time. p chooses which of these quests to accept, up to m quests. An AID d selects which q should be shown to p, and selects $\langle q_1, q_2, ..., q_n \rangle$.

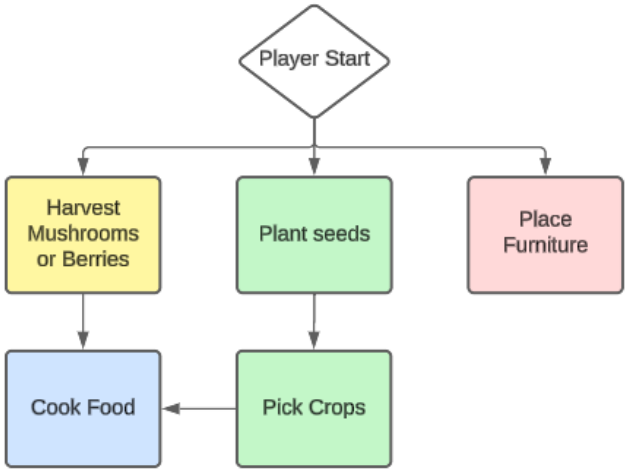

Figure 1: The four main areas of gameplay are placing furniture, planting crops, harvesting, and cooking.

3.2 FarmQuest

FarmQuest is a research test bed video game for AIDs (Yu, Guzdial, and Sturtevant 2022). The game loop consists of planting crops, harvesting mushrooms and berries, placing furniture, and cooking recipes, shown in Figure 1. Figure 2 shows the main level where players access the different types of gameplay. The region labeled "A" is where berries can be harvested, and the region labeled "B" is where mushrooms can be harvested. The region labeled "F" is where seeds can be planted, and the region labeled "D" is where furniture can be placed and recipes can be cooked. There is a shop to buy and sell items in the game, labeled "E", and a quest board, labeled "C", where a player can submit or accept quests. The goal is to earn enough coins to pay off their mortgage. The player can earn coins by selling items in the shop or by completing quests. The player starts with 300 coins, and needs to earn 1000 coins to finish the game.

The FarmQuest AID is intended to be embedded in the quest system, where the AID changes which quests are shown to the player. The quest board, shown in Figure 3, allows the player to interact with the quest system. There are two tabs, a submit quest tab and an accept quest tab, and the quest board starts on the submit tab. In the accept quest tab, the current AID selects three quests to present to the player. We will discuss the AIDs in Section 3.3. The player accepts a quest by clicking on it, and can have a maximum number of three quests at a single time. Every time the accept tab is clicked by the player, new quests are chosen by an AID. Once a player completes a quest, the quest can be submitted in the submit tab. Accessing the submit tab of the quest board does not prompt the AID for new quests.

Each quest is associated with a gameplay type that can be

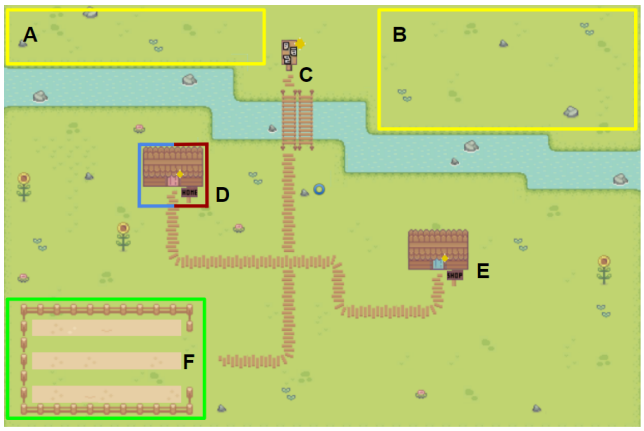

Figure 2: An overview of the main level in FarmQuest

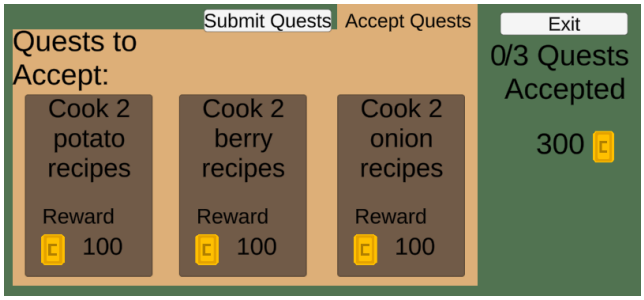

Figure 3: The Questboard that was shown to the players

completed using systems in the game. Each quest gives the same reward of 100 coins, to ensure that the reward isn't influencing the decision of the player to accept a quest. There are six place quests, six plant quests, eight cook quests, and six harvest quests as shown in Table 1. There are at least six of each quest because that is the minimum number of quests that is needed for the quest system to function. This covers the case where the player accepted three of the same quest type, and the AID selects three more of that quest type to present to the player. There are eight cooking quests in order to cover the breadth of gameplay items that can be cooked. For the purposes of this paper, each quest is labeled according to the main gameplay type of the quest, where place furniture quests are labeled "F", plant crop quests are labeled "P", cook recipe quests are labeled "C", and harvest berries and mushroom quests are labeled "H".

3.3 AI Directors

There are three AIDs used in this experiment. The first AID is a random algorithm. A common way to implement this quest system is to randomly choose a quest from a pool of existing quests (Yu et al. 2022). However, this comes at the cost of the player experience. Players may be presented with something they do not want to do, which may reduce their motivation to complete the quest. Players may also be asked to undertake a task they have recently accomplished independent of a quest, which may introduce fatigue by completing a quest with those actions. The random algorithm selects a quest from the pool of total quests with a uniform random probability. In our context, it is slightly more likely to select cook quests because there are more cook quests.

The second AID is PaSSAGE (Thue et al. 2007). PaSSAGE was originally designed to modify quests in a branching narrative to better suit the player preference. We have adapted it to fit quest selection. PaSSAGE models the player's preference for gameplay based on the actions they have previously taken, and assumes that if a player is taking the action they are enjoying that action. Our implementation of PaSSAGE tracks the number of times a player completes one of the four main gameplay types to use as a player model. PaSSAGE uses rollouts to estimate player return, and determines which of the quests are most appropriate to show to the player. The rollouts stop when there is a decision point - in our case, each rollout is a singular step because there is a singular decision point. PaSSAGE then predicts which type of quest is the most suitable for the player. From there, a specific quest is chosen with uniform random probability from the set of possible quests for the predicted type.

The third AID is the reinforcement learning AID, which uses a combinatorial multi-armed bandit (CMAB) (Yu et al. 2022). From this point forward, we refer to this AID as the CMAB AID. This AID was originally deployed on a similar quest system to the one in FarmQuest, so the only modifications were the ones necessary to transfer it to the FarmQuest domain. Instead of a traditional bandit algorithm where each quest would be an arm, the CMAB AID uses a set of quests as an arm. As shown in Figure 3, there are three quests presented to the player at the same time, and this set of presented quests is the arm for the CMAB AID. CMAB generates the set of all possible quest arms based on the quest type, rather than the set of all quests. For example, a place quest, a cook quest, and a plant quest is a valid arm, but "place 2 furniture", "cook 2 onion recipes" and "plant 3 carrots" is not a valid quest arm. The CMAB AID rewards super arms. A super arm is the set of all arms that have at least one quest type in common, and when one arm is rewarded, all of the arms in the super arm are rewarded. The reward is assigned by the player accepting or not accepting quests, where an accepted quest gives a reward of 1 and an unaccepted quest gives a reward of 0. Then, a particular arm in a super arm is selected using UCB, which gives a set of three quest types. The individual quest is then selected using a uniform random probability from the pool of quests of a given type.

Given these three AIDs, we hypothesized that the CMAB AID will perform the best on quest selection. The random algorithm has no curation, which might cause the previously outlined friction in the player experience. PaSSAGE assumes that a player's previous actions are an indicator for which quests they will prefer, but players may take actions for various reasons. CMAB assumes that accepting a quest is an indicator of which quests they will prefer. We believe that accepting a quest is a stronger indicator of what a player will prefer than the player's previous actions.

4 Methodology

Our AID experiment was an AB test. Though we have three AIDs, we chose to not have participants experience all three

Place Quests	Plant Quests	Cook Quests	Harvest Quests
		C1Cook 2 berry recipes	
F1 Place 2 furniture	**P1**Plant 3 carrots	**C2** Cook 2 carrot recipes	**H1** Harvest 3 berries
F2 Place 3 furniture	**P2** Plant 3 green onions	**C3** Cook 2 green onion recipes	**H2** Harvest 4 berries
F3 Place 4 furniture	**P3** Plant 3 lettuce	**C4** Cook 2 lettuce recipes	**H3** Harvest 5 berries
F4 Place 5 furniture	**P4** Plant 3 onions	**C5** Cook 2 mushroom recipes	**H4** Harvest 3 mushrooms
F5 Place 6 furniture	**P5** Plant 3 potatoes	**C6** Cook 2 onion recipes	**H5** Harvest 4 mushrooms
F6 Place 7 furniture	**P6** Plant 3 tomatoes	**C7** Cook 2 potato recipes	**H6** Harvest 5 mushrooms
		C8 Cook 2 tomato recipes	

Table 1: All of the quests in FarmQuest

directors. Instead, each participant was randomly assigned two out of three of the AIDs in order to shorten the study and reduce confusion involved in comparing three AIDs.

We show the flow of the experimental setup in Figure 4. First, we asked each player to sign a consent form. Then, the player filled out a demographic survey and completed a short tutorial to learn how to play the game. Then, they played FarmQuest with their first AID until they paid off their mortgage, and answered a short survey about their experience. Then, they played FarmQuest again but with a different AID, and answered the same short survey again. Finally, each player answered a survey comparing the two experiences. The entire experiment took one hour or less to complete.

The demographic questions were as follows:

- **DQ1** What is your current age?
- **DQ2** What is your gender?
- **DQ3** Do you consider yourself cisgender or transgender?
- **DQ4** Do you consider yourself a gamer?
- **DQ5** How often do you play games in a week?
- **DQ6** What kind of genres of games do you play the most? Select all that apply.

Each demographic question is labeled "DQ" for ease of reference. These questions were all asked because they are considered factors that could affect the player's perception of video games. Age (Whitbourne, Ellenberg, and Akimoto 2013), gender (Desai, Zhao, and Szafron 2017; Phan et al. 2012), and familiarity with video games (Manero et al. 2017) all could have an impact.

The short survey questions were targeted to the participant's most recent playthrough. We based this survey off of our previous study questions (Yu et al. 2022). We asked players to think about their most recent experience, and to answer the questions accordingly. The short survey questions were all Likert questions on a scale from one to five, where one is strongly disagree and five is strongly agree. The short survey questions were as follows:

- **SQ1** I felt like I was able to accept quests that I wanted to do
- **SQ2** I felt like I was **not** able to accept quests that I wanted to do
- **SQ3** I felt like I was able to complete quests that I wanted to do
- **SQ4** I felt like I was **not** able to complete quests that I wanted to do
- **SQ5** I felt like there was a variety of quests for me to complete
- **SQ6** I feel like I enjoyed playing the game
- **SQ7** I feel like I did **not** enjoy playing the game
- **SQ8** I feel like I would recommend playing this game to a friend
- **SQ9**I feel like I would **not** recommend playing this game to a friend

Each short survey question is labeled with "SQ" for ease of reference. We asked SQ1 and SQ2 to learn about accepting quests, SQ3 and SQ4 to learn about completing quests, and SQ5 to learn about variety. We asked SQ6, SQ7, SQ8 and SQ9 to learn about enjoyment, as it can have an effect on the perception of the AID (Yu et al. 2022).

The comparison survey questions compared the experience between the two game sessions. We developed the comparison survey questions out of the short survey questions. We asked players to think back on their first experience and compare it to their second experience. We first asked a series of Likert questions, using the same scale as the short survey questions. The Likert Questions were as follows:

- **CQ1** I preferred my first experience playing the game
- **CQ2** I preferred my second experience playing the game
- **CQ3** I felt like I accepted more quests that I wanted to do in the first experience
- **CQ4** I felt like I accepted more quests that I wanted to do in the second experience
- **CQ5** I felt like I completed more quest that I wanted to do in the first experience
- **CQ6** I felt like I completed more quests that I wanted to do in the second experience
- **CQ7** I felt like I had more fun in the first experience
- **CQ8** I felt like I had more fun in the second experience

We then asked the players two short answer questions:

- **CQ9** What was your favorite activity to do in the game? Did you feel like you got to experience a lot of that activity? Why or why not?
- **CQ10** Did you feel that there was a difference between your first and second experiences? Why or why not?

Figure 4: The flow of the participant through the study

Each comparison survey question is labeled with "CQ" for ease of reference. We asked pairs of questions to focus on different aspects of player perception. We asked CQ1 and CQ2 for preference, CQ3 and CQ4 for quest acceptance, CQ5 and CQ6 for quest completion, and CQ7 and CQ8 for enjoyment. We asked CQ9 and CQ10 so players could describe their experience in their own words. We asked CQ9 to determine if players would identify as a particular player type. We asked CQ10 to determine if players could tell a difference between AIDs.

This study was approved by the Research Ethics Board at the University of Alberta (UofA), ethics ID number (Pro00129326). To advertise for the study, we posted digital calls for participation in classroom forums and Discord at the UofA, and AI and video game studio Slack spaces. The only requirement for participation was that a person be over the age of 18. There was no financial incentive.

5 Results

In this section, we discuss the demographic, qualitative, and quantitative results of the human subject study.

5.1 Demographic Results

First, we present the demographic results in order to gain an understanding of any potential biases from the audience that participated in this study. In total, 108 players started the study. Of these, only thirty-nine players fully completed the study. Most players stopped the study after filling out the demographic questions. Of the thirty-nine complete responses, seven of these included data where the players refreshed the page at some point, which caused the system to assign a different AID. This means that we cannot know which experience they are talking about when they answered the survey questions, so we do not include this data. We are left with thirty-two complete responses that have usable data.

Table 2 shows the self-reported age and gender of players. Twenty-nine players identified as cisgender, two identified as transgender, and one did not answer. Table 3 shows the genres of games played by the players, where each player could select more than one option. For whether a player considered themselves a gamer or not, twenty-eight reported yes and four reported no. Table 4 shows the self-reported number of times the participants played games in a week.

Table 5 shows the number of players and the specific ordering of AIDs each player experienced. In total, twenty-two players experienced PaSSAGE, twenty experienced CMAB, and twenty-two experienced random.

5.2 Quantitative Results

We present the quantitative results from the study. First, we present the number of quests presented, accepted, and completed by each player, shown in Table 7. We wanted to see if the order of sessions had an effect on any of this data, so we ran a Kruskal-Wallis statistical test. None of these tests yielded statistically significant results, so we assume that ordering did not have an effect and combined the data. We wanted to see if there is a difference in any of this data based on AID, so we ran Mann-Whitney U tests. Only the number of presented quests had statistically significant results, shown in bold in Table 7. The CMAB AID had statically significantly fewer presented quests than both PaSSAGE ($p = 0.032$) and random ($p = 0.047$) AIDs.

Reported Age	Players	Reported Gender	Players
18-25	7	Man	23
26-35	16	Woman	6
36-45	8	Non-Binary	2
45 or older	1	Prefer not to say	1

Table 2: The reported age and gender of players

Genre	Players	Genre	Players
Puzzle	17	Platformer	12
Arcade	7	Sports	3
RPG	25	Racing	6
FPS	20	Simulation	16
Action	20	Fighting	2
Adventure	26	Other	5

Table 3: The genres of games that the players played

Second, we present the average playtime. To calculate the time spent in game, we cannot use the timestamp data included in our telemetry due to a problem with the logging. Instead, we use the number of days spent in game as an approximation. In game, each day is thirty seconds, each twilight is ten seconds and each night is twenty seconds, so a complete in-game day is one minute. Thus, we can measure the time spent in game to floor of the nearest minute.

Table 6 shows the average, stdev, and standard error of the

Reported time spent playing games	Players
Less than once a week	1
Once a week	5
Two or three times a week	3
Four or five times a week	10
Six or seven times a week	13

Table 4: The reported amount of time spent playing games in a week

AID Ordering	Number of Players
PaSSAGE then Random	5
PaSSAGE then CMAB	3
CMAB then PaSSAGE	7
CMAB then Random	3
Random then PaSSAGE	7
Random then CMAB	7

Table 5: The number of players for each ordering of AIDs

AID	Average In Game Days	Stdev	SEM
Session 1 Passage	8.75	4.20	1.48
Session 2 Passage	8.35	6.36	1.70
Session 1 CMAB	12.30	4.72	1.49
Session 2 CMAB	8.70	4.90	1.54
Session 1 Random	**11.28**	5.61	1.49
Session 2 Random	**7.25**	6.48	2.29

Table 6: The average number of in games days per AID and session, where bold indicates statistically significant results

mean (SEM) on the number of in game days played by players. The SEM is a measure of how far the sample mean is from the population mean. To see if ordering of session had an effect on the duration that players played the game, we ran a Kruskal-Wallis statistical test. Only the random AID had a statistically significant result ($p = 0.011$) with a 95 $\%$ confidence interval, shown in bold in Table 6. This means that ordering of session does not have an effect in the cases of the PaSSAGE and CMAB AIDs, but does have an effect with a random AID. For the random AID, we wanted to verify if the playtime for session 2 is statistically lower. We ran a Mann-Whitney U test with the alternate hypothesis that session 2 is greater than session 1. This showed a statistically significant result with a 95 $\%$ confidence interval ($p = 0.006$), so session 1 is statistically longer than session 2. This means that players who were assigned the random AID in session 1 played statistically longer than the players who were assigned the random AID in session 2.

5.3 Qualitative Results

We present the qualitative results from the study. First, we present the results from questions SQ1-SQ9. We wanted to determine if the ordering of the AIDs had an effect. We ran a Kruskal-Wallis statistical test on the responses, where a statistically significant result would show that ordering does have an effect. Only SQ6, "I feel like I enjoyed playing the game" had a statistically significant difference for the random AID with a 95$\%$ confidence interval ($p = 0.028$). This means we can treat the SQ6 answers from sessions without a random AID as coming from the same distribution, so we combined the data. Table 8 shows the results from the short survey with SQ6 for random removed.

The median result for session 1 of the random AID is 4.00, and the median result for session 2 of the AID is 2.50. The average result for session 1 of the random AID was 3.61, and the average result for session 2 of the random AID was 2.12. According to the Likert scale, a higher number means that they agree more with the statement "I feel like I enjoyed playing the game".

To determine if the results for session 2 are from a lower distribution, we ran a Mann Whitney-U test on the results from SQ6 with the alternate hypothesis that session 1 is greater than session 2. We ran this test on all three AIDs. The PaSSAGE ($p = 0.765$) and CMAB ($p = 0.578$) AIDs were not statistically significant, but the random AID was statistically significant, indicating less enjoyment ($p = 0.016$).

The results from the comparison questions are shown in supplementary material (Yu, Guzdial, and Sturtevant 2024) because there was no statistical significant with a 95$\%$ confidence interval.

6 Discussion

In this section, we discuss the two facets of the impact of AIDs: quantitative changes in behavior and player's perception of impact.

6.1 Quantitative Changes in Behavior

We hypothesized that the CMAB AID would be the most suitable for this particular game and the quest selection problem. However, our results do not show that the CMAB AID is the clear winner, and instead showed that either the PaSSAGE or CMAB AID is preferable to players. We looked at the number of quests presented, accepted, and completed by the player, the total play time, and the short survey data to support this claim.

The CMAB AID had statistically significantly fewer presented quests than both the Random and PaSSAGE AIDs, while still having a similar number of accepted and completed quests. This shows that players were able to accept and complete similar numbers of quests with all three AIDs, while the CMAB AID needed to present the fewest number of quests to do so. This shows that the CMAB AID was effectively able to reduce time spent searching for quests.

The total playtime shows differences in how long players played, but is not statistically significant, as shown in Table 6. The playtime for CMAB trends higher than random in both sessions, which suggests that the CMAB AID is able to perform well in both session 1 and session 2 for the player. The playtime for PaSSAGE is lower than the random AID during session 1. We compared PaSSAGE session 2 to random session 2, and PaSSAGE trended higher than the random AID. However, this result was also not statistically significant. This data shows there may be a preference for either PaSSAGE or the CMAB AIDs.

In the short survey data, SQ6 had a statistically significant result. Players indicated that they did not enjoy playing random during the second session, shown in Section 5.3. In this experiment, they experienced either the PaSSAGE or CMAB AIDs first, where AIDs curated the quests to the player. We anticipate the difference between a curated and random experience was more apparent when faced with a curated experience first. This data showed a preference in players for either non-random AIDs in terms of enjoyment.

Presented	Average	Std Dev	SEM	Accepted	Average	Std Dev	SEM	Completed	Average	Std Dev	SEM
PaSSAGE	**27.27**	13.85	2.95	PaSSAGE	8.59	3.40	0.72	PaSSAGE	7.09	3.61	0.77
CMAB	**20.68**	8.36	1.92	CMAB	7.89	1.88	0.43	CMAB	6.58	2.01	0.46
Random	**29.71**	22.06	4.81	Random	8.24	2.96	0.65	Random	7.05	2.65	0.58

Table 7: The number of presented, accepted and completed quests by AID, where bold indicates statistically significant results

AID	SQ1	SQ2	SQ3	SQ4	SQ5	SQ6	SQ7	SQ8	SQ9
PaSSAGE	4	2	4	2	4	4	3	3	3
CMAB	4	2	4	2	4	3	3	2	3
Random	4	2	4	2	3	-	2	3	3

Table 8: Comparison of the median value for short survey questions, where 1 is strongly disagree and 5 is strongly agree

6.2 Player's Perception of AIDs

Thus far we have discussed the evidence for how AIDs quantitatively affect the player experience, but the question remains: did players perceive a difference in experience? We conclude that players under different AIDs will play the game differently, but do not perceive a difference. We discuss the short survey data, the comparison survey data, and some free responses to support this conclusion.

For the short survey data, most of the responses showed no difference between AIDs. Additionally, the comparison questions did not show any statistically significant results, as shown in the supplementary material (Yu, Guzdial, and Sturtevant 2024). Players did not notice a difference between their experiences for the majority of the qualitative questions, and their responses reflect that. When directly asked in question CQ10, there were some players who responded "No, I didn't notice any difference", or "I did not feel like there was a huge difference between the two experiences." However, we had measurable quantitative differences. We conclude that players play differently with curated vs non-curated AIDs, but fail to notice a difference in their play style that is caused by the AID.

7 Threats to Validity

One possible threat to validity to our findings is the effect of salience on players selecting quests. Salience is the idea that some objects, when presented together, are more noticeable by a person based on their ordering. As shown in Figure 3, three quests are presented to the player in a row. There could be a salience effect where native English speakers notice the left-most quest more because English speakers read left to right. To determine if this is the case, we analyzed the placement data of each quest according to quest type, shown in the supplementary material (Yu, Guzdial, and Sturtevant 2024). We ran a Chi-Squares contingency test on the number of presented quests in each position compared to the number of accepted quests in each position, and we found that there was no statistically significant results with a confidence interval of 95%. We conclude that salience did not have an effect on players selecting quests.

Another possible threat to validity is the effect of difficulty on selecting quests. It could be that players are selecting quests based on what is easiest. To address this, we analyzed the difficulty within each quest type. For place and harvest quests, we assume that a smaller number of items placed or harvested means an easier quest. For cook quests, we assume that C1 and C5 are easier because the ingredients do not require planting, and assume that all other cook quests are of equal difficulty because they take the same number of actions to complete. For plant quests, we assume that all quests have an equal difficulty because they all require the same number of actions to complete. We ran a Chi-Squares contingency test on the number of presented and accepted quests for each individual quest, to see if there is a bias towards quests that are easier. There is only one category that had statistical significance, which was the cooking type quests for the random AID ($pvalue = 0.008$). We believe this is due to the higher number of acceptances of the quest C7. This quest is assumed to be similar difficulty to the other cook quests except for C1 and C5. Because C7 is the quest with a higher number of acceptances, and not C1 or C5, this does not indicate that there are a higher number of people choosing an easier quest. The other statistical tests do not show any significance. Thus, we conclude that the difficulty does not have an effect on players selecting quests.

8 Conclusion

In this paper we evaluated previously inconclusive AIDs on quest selection. We directly compared tested these AIDs in a human subject study. We collected both quantitative and qualitative data, which showed that a curated AID leads to a longer time played and less quests presented than a random AID, but players fail to perceive a difference.

In the future, we hope to gain a larger sample for the comparison between the PaSSAGE and CMAB AIDs, as the small sample size limited some of the analysis on our data. This could help disambiguate which AID is better for this problem, if either. Additionally, we hope to test other AIDs on the quest selection problem. This will help paint a clearer picture of the individual strengths and weaknesses of existing AIDs, so we can better understand the gaps that need to be addressed in future research.

Acknowledgements

This work was funded by the Canada CIFAR AI Chairs Program, Alberta Machine Intelligence Institute, and the Natu-

ral Sciences and Engineering Research Council of Canada (NSERC).

References

Bethesda. 2011. The Elder Scrolls V: Skyrim. PC, PS3, Xbox 360.

Booth, M. 2009. The AI Systems of Left 4 Dead. In *Game Developers Conference*.

Charles, D.; Mcneill, M.; Mcalister, M.; Black, M.; Moore, A.; Stringer, K.; Kücklich, J.; and Kerr, A. 2005. Player-centred game design: Player modelling and adaptive digital games. *Proceedings of DiGRA 2005 Conference: Changing Views - Worlds in Play*.

Desai, N.; Zhao, R.; and Szafron, D. 2017. Effects of Gender on Perception and Interpretation of Video Game Character Behavior and Emotion. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(4): 333–341.

Dias, R.; and Martinho, C. 2011. Adapting content presentation and control to player personality in videogames. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, ACE '11. New York, NY, USA: Association for Computing Machinery. ISBN 9781450308274.

Manero, B.; Torrente, J.; Fernández-Vara, C.; and Fernández-Manjón, B. 2017. Investigating the Impact of Gaming Habits, Gender, and Age on the Effectiveness of an Educational Video Game: An Exploratory Study. *IEEE Transactions on Learning Technologies*, 10(2): 236–246.

Nintendo. 2020. Animal Crossing: New Horizons. Nintendo Switch.

Nygren, N.; Denzinger, J.; Stephenson, B.; and Aycock, J. 2011. User-preference-based automated level generation for platform games. In *2011 IEEE Conference on Computational Intelligence and Games (CIG'11)*, 55–62.

Phan, M. H.; Jardina, J. R.; Hoyle, S.; and Chaparro, B. S. 2012. Examining the Role of Gender in Video Game Usage, Preference, and Behavior. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1): 1496–1500. _eprint: https://doi.org/10.1177/1071181312561297.

Thompson, T. 2017. In the Directors Chair: The AI of Left 4 Dead.

Thue, D. 2007. *Player-Informed Interactive Storytelling*. Master's thesis, University of Alberta.

Thue, D. 2010. *Generalized Experience Management*. Ph.D. thesis, University of Alberta.

Thue, D.; Bulitko, V.; Spetch, M.; and Wasylishen, E. 2007. Interactive Storytelling: A Player Modelling Approach. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 3(1): 43–48.

Valve. 2008. Left 4 Dead. PC, XBox 360.

Wauck, H.; and Fu, W.-T. 2017. A Data-Driven, Multidimensional Approach to Hint Design in Video Games. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, 137–147. New York, NY, USA: Association for Computing Machinery. ISBN 9781450343480.

Whitbourne, S. K.; Ellenberg, S.; and Akimoto, K. 2013. Reasons for playing casual video games and perceived benefits among adults 18 to 80 years old. *Cyberpsychology, Behavior, and Social Networking*, 16(12): 892–897.

Yannakakis, G. N.; and Hallam, J. 2009. Real-Time Game Adaptation for Optimizing Player Satisfaction. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(2): 121–133. Conference Name: IEEE Transactions on Computational Intelligence and AI in Games.

Yu, H.; and Riedl, M. 2013. Data-Driven Personalized Drama Management. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 9(1): 191–197. Number: 1.

Yu, K. K.; Guzdial, M.; and Sturtevant, N. 2024. Evaluating the Effects of AI Directors for Quest Selection. arXiv:2410.03733.

Yu, K. K.; Guzdial, M.; and Sturtevant, N. R. 2022. PWR: A Demonstration of an AI Director Video Game Test Bed. In *Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*.

Yu, K. K.; Guzdial, M.; Sturtevant, N. R.; Cselinacz, M.; Corfe, C.; Lyall, I. H.; and Smith, C. 2022. Adventures of AI Directors Early in the Development of Nightingale. In *Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*.

Yun, C.; Trevino, P.; Holtkamp, W.; and Deng, Z. 2010. PADS: enhancing gaming experience using profile-based adaptive difficulty system. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, 31–36.