Effects of Gender on Perception and Interpretation of Video Game Character Behavior and Emotion

Neesha Desai, Richard Zhao, and Duane Szafron

***Abstract*—Gender in video games is a popular topic. However, the focus is usually on how gender is portrayed within games. In this paper, we examine the effects of players' gender on the perception of virtual character behavior and emotion based on the results of two user studies involving story-based games. The first study compared players' perception of virtual character behaviors. We analyzed perceived differences both by gender and by gaming experience. In this study we found that female gamers were more appreciative of complex behaviors than male gamers. In the second study, we examined the influence of gender on player' ability to identify the emotion being displayed by a virtual character. We found that most emotions were identified comparably, with the exception of anger. Female players were significantly better at identifying angry characters compared to male players. We also investigated any perception differences between emotions expressed by male and female virtual characters, but we did not identify any statistically significant differences. Overall, the studies suggest that there are differences in how male and female players perceive virtual characters, and if game designers want players to perceive these characters in a certain way, they should consider the gender of targeted players.**

***Index Terms*—character behavior, character emotion, gaming experience, gender, machine learning.**

I. Introduction

VIDEO games are continuing to dominate the entertainment market. With the amount of choice now offered to story-based game players (ranging from what their character looks like to the type of playing style they wish to have), players are becoming much more involved in and much more connected to the story. In order to capitalize on this trend, some researchers and developers have begun to investigate what must be done to create more believable virtual characters, which we also refer to as Non-Player Characters (NPCs) in this paper. The work often results in using a combination of AI techniques, animations, and voiced dialogue. However, there has been no research done on how (or if) gender affects the perception and/or enjoyment of these changes. For the purpose of this discussion, we will use the word "gender" to exclusively denote the biological gender, or biological sex.

While the storylines may have become more complex, there has also been a lot of concern over the influence of video games on attitudes about gender. Because of the potential benefits and drawbacks of game play, it is important to consider how gender may play a role. For example, how are NPCs' behaviors influencing game players' perceptions of gender? How effective are techniques used to create complex character behaviors and do male and female game players differ?

Previous research on gender in gaming showed that there are differences between genders in attitudes toward gaming in general and the types of games [14, 34]. It is not unreasonable to hypothesize that there are also differences in how each gender perceive the behaviors of NPCs in story-based games, such as those in The Elder Scrolls V: Skyrim [2].

We are interested in the effects of gender on the perception of NPCs in story-based games through non-facial means. This paper examines two aspects of perception, the perception of behavior and the perception of emotion. Behavior and emotion are closely related, as research has shown that emotion affects people's behaviors [36]. We focus on the effects a player's gender may have on *perception of NPC behavior and/or emotion in story-based games*.

We analyze the results of two different user studies looking at gender differences and similarities. The first study involved examining the believability of NPC behavior. The second study compared the accuracy of participants in identifying the emotion of NPCs.

II. Related Work

Over the past few decades there has been considerable research about the various intersections of video games, gender and emotion.

A. Gender in Games

It is often assumed that the vast majority of video game players are male. However, it turns out that the split is much closer to even - 52% male and 48% female [10].

Psychologists have examined the role of gender in video games; however, many have focused on the effect of video games on players. Ogletree and Drake [34] reported that in a university population, men were more likely than women to play video games, and to indicate that game playing interfered with sleeping and with class preparation. In terms of their preference for in-game avatars, men were significantly more likely than women to choose a male character, and women were significantly more likely than men to choose a female character.

Previous studies have also shown that men and women prefer different types of gaming experience. In 2006, Hartmann and Klimmt [14] studied the characteristics of video games that females dislike. They found the three main characteristics were the lack of meaningful social interaction, the violent game play and the gender stereotyping of characters. However, since this study was produced, it can be argued that there is a plethora of games that start to address some of these concerns. Many recent games, such as Mass Effect 3 [4] and The Elder Scrolls V: Skyrim [2] include much more complex and deeper story lines, and players are often given a lot of choice as to the type of character they play and thus the nuanced role of the character in the game. While there is still a significant amount of gender stereotyping in games, some game developers are acknowledging the problem and even including player input in design decisions. One example is BioWare's involvement of the Mass Effect fan community when re-designing the female version of Commander Sheppard for Mass Effect 3 [8]. These studies suggest that game designers should treat females and males as separate target audience when designing their games.

B. NPC Behavior

The first of our two studies examines the effect of behaviors of NPCs in story-based games. Traditionally behaviors for each individual character are scripted manually by programmers, which is a time and resource consuming process. Mateas and Stern [28] used a specialized programming language, ABL, together with a drama manager to produce complex behaviors for the two characters of the game Façade. However, the game Façade took years of resources to complete for a few hours of game-time experience. Researchers have looked into ways to streamline this process to provide more believable behaviors with fewer resources. ABL has since been adapted to model character conflict resolution strategies [33]. Prom Week [38, 39] is another game created as a result of research into streamlined game design. Other techniques include finite state machines, hierarchical finite state machines, behavior trees [29], and planning-based approaches [30, 31, 32].

Bakkes et al. [35] used case-based adaptive game AI techniques to control characters in a real-time strategy game. In this case, the goal was to win the competition so opponent modelling techniques were used for exploitation. Believability of characters was not a consideration.

Orkin and Roy [18] devised a data-driven approach to generating believable behaviors using unsupervised learning of behavior and dialogue in an online game that simulates a restaurant. Data-driven approaches have also been used in research to analyze player statistics extracted from in-game traces or character attributes [15]. The complex behaviors used for comparison in our study are based on a data-driven approach [24].

C. Emotion

For our second study, we focus on the emotions that are most widely regarded as being “basic” emotions: those that can be most accurately identified across cultures [9]. Specifically, we examine player identification of happiness, sadness, anger and fear.

Examining whether emotion can be recognized from non-verbal cues can be traced back decades [23], but it has most commonly been tested with human actors. Hall [12] examined previous studies looking at gender effects when identifying emotion through non-verbal (visual and/or auditory) cues. She found that females statistically outperform males. However, she also found that there was little to no effect caused by the gender of the labeled character, which means the effect was on those doing the labeling. Our goal was to determine whether the results of that study, based on human actors, are applicable to NPCs. We also wanted to determine whether the results are influenced by each particular emotion being identified.

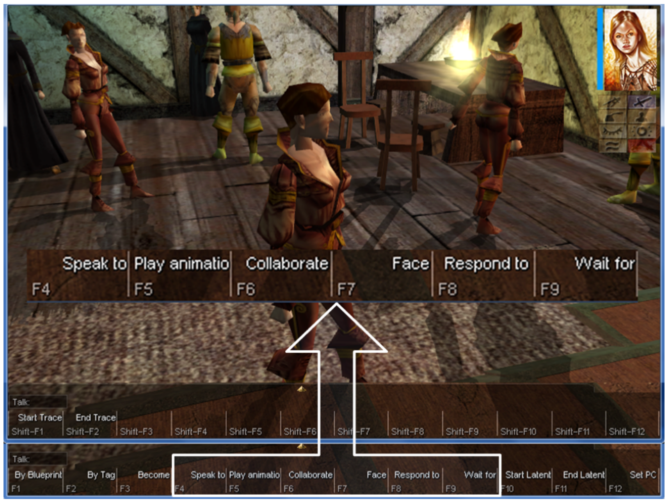

Fig. 1. In training mode, the designer is able to control a character and train the behaviors for this character. Some example actions a designer can choose from are enlarged in the middle, such as to speak to another character.

Animators have long acknowledged that emotion is one of the primary means to produce an illusion of life for NPCs [37]. Emotion has been used by Ochs et al. [36] to help create believable NPCs, where the emotional states of the characters were affected by their pre-defined personality and their role, which in term changed their behavior in social interactions.

III. Perception of Behaviors

Our first study measured whether there were any differences between how males and females perceive NPC behaviors.

We compare three behavior variations, an idle behavior, a simple behavior, and a complex behavior (all described in Section B). The complex behaviors are generated using a learn-by-example data-driven approach [24]. This approach combined Hidden-Markov Models (HMMs) with Behavior Capture. Instead of a programmer writing code to specify how each character should move, speak, and interact with the environment, a game designer controls the character and performs the actions that the designer would like this character to perform. This is done in a training mode, as shown in Fig. 1. As the designer performs different actions, the system captures traces of actions. The system captures these exemplar behaviors and generalizes them to produce behaviors for NPCs when players play the game. This removes the need to write custom programming scripts for each character. The details of the behavior generation process can be found in Zhao and Szafron [24].

Generalization takes a few forms. One trained character can be generalized to a category of characters, such as all tavern patrons of a certain type, as specified by a designer. A trained action can be generalized to a category of actions, such as sitting on any chair in a tavern, instead of the specific chair that was trained on. Most importantly, for each character, by training a Hidden-Markov Model with the exemplar behaviors using the Baum–Welch Algorithm [25], the NPCs can then perform behaviors that are driven by the outputs of the Hidden-Markov Models. This gives controlled stochasticity to the behavior outputs.

In a previous study, the HMM approach to behavior generation has been validated as more believable compared to existing alternatives in commercial games [24]. In this study, we used the HMM behavior as the complex behavior and used it for the purpose of identifying gender differences. In the original study [24], six behavior variations were compared. Four were Behavior Capture-based, with one of them using HMMs, while the other two were a baseline idle behavior and a set of hand-scripted behaviors that mimicked a commercial story-based game: Dragon Age: Origins [26]. Study results showed that the most complex behavior was ranked as most believable. We extended this study to examine the differences in the responses between males and females, along with an additional independent factor: whether the participant is an experienced game player (referred to as a gamer). We defined a gamer as a participant who plays video games at least once per week.

We created identical scenes using the Neverwinter Nights (NWN) game engine [3]. Each scene used an identical setting and an identical set of characters. The only difference was the technique controlling the behaviors of the NPCs in each scene. In this paper, we focus on three of the behavior variations: the baseline idle behavior, the behavior hand-scripted to mimic a commercial game, and the complex behavior generated by an HMM (all described in Section B). We asked the participants to rate each scene according to the believability of the behaviors of NPCs, on a scale of 1 to 4 (very unbelievable, unbelievable, believable, very believable). A neutral option was not included to force a choice. Previous research has shown that when the question was likely to evoke ambivalence, the absence of a neutral position helps shift the mean responses towards one of the extremes [44]. Since participants who do not play games are more likely to be ambivalent, we removed the neutral option. In addition, the participants were asked to complete a questionnaire on their gender and how often they played video games.

A. *Participants*

This study had 79 participants, who were undergraduate students taking a first year psychology class. They were between the ages of 17 and 58 (mean of 20.6), with 51 females and 28 males. Of these, 12 of the females were gamers and 39 were non-gamers, while 12 of the males were gamers and 16 were non-gamers.

The study was conducted in a laboratory on a university campus. Each participant was asked to sit in front of a computer and observe the characters on the screen in a first-person view. The participants were free to navigate around the scenes using the keyboard and mouse but they could not interact with the characters or the environment in any way. Participants were given an hour to observe the scenes and to complete the questionnaire.

B. *Behavior Descriptions*

TABLE I
AVERAGE RATING SCORES OF BEHAVIOR VARIATIONS IN TERMS OF BELIEVABILITY, DIVIDED BY GENDER AND GAME EXPERIENCE

Behavior	Female	Male	Gamer	Non-Gamer	Female Gamer	Female Non-Gamer	Male Gamer	Male Non-Gamer
A: Baseline idle behavior	1.20	1.29	1.33	1.19	1.17	1.21	1.50	1.13
B: Behavior based on commercial game Dragon Age: Origins	2.06	2.32	1.83	2.29	1.42	2.26	2.25	2.38
C: Complex behavior produced using the Behavior Capture plus HMM	3.02	2.89	2.96	2.98	3.00	3.03	2.92	2.88
Difference between B and A	0.86	1.03	0.50	1.10	0.25	1.05	0.75	1.25
Difference between C and B	0.96	0.57	1.13	0.69	1.58	0.77	0.67	0.50

Cutumisu [27] categorizes NPC behaviors as proactive independent, proactive collaborative and latent. A proactive independent behavior is an ambient behavior performed by one character alone such as sitting on a chair. A proactive collaborative behavior is a synchronized ambient collaboration between two characters, such as a conversation between the two. A latent behavior is a behavior triggered by game events, such as NPCs fleeing a building when it catches fire. While latent behaviors can also be independent or collaborative, only latent independents were used in this study.

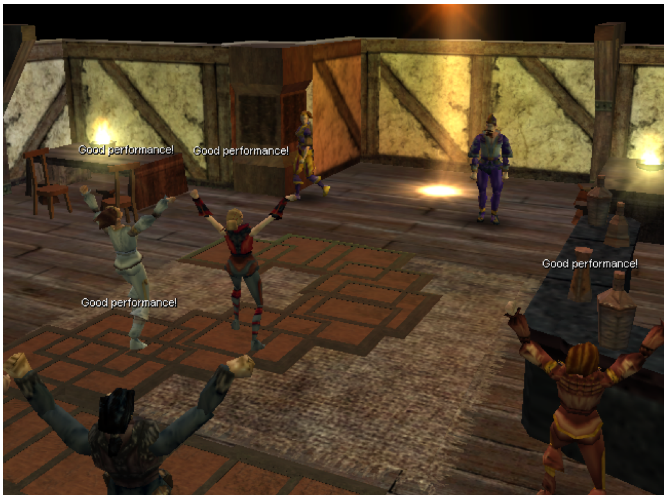

Fig. 2. An example latent behavior at game time: the tavern patrons turn and cheer for the two bards as they finish their performance. The cheer behavior is triggered by the event of the bards leaving their performance spotlight.

We are interested in any differences between the different groups of players in their perception of simpler behaviors and more complex behaviors. Character complexity can take on different meanings, such as behaviors that are socially believable [40, 41]. In this paper we define complex behaviors as behaviors that exhibit independent behaviors, collaborative behaviors, and latent behaviors. A previous study has shown that characters who exhibit this kind of complex behaviors are more believable compared to characters without them [24].

We compare participant responses to these three resulting behavior variations:

A: Baseline idle behavior

B: Behavior hand-scripted based on Dragon Age: Origins

C: Complex behavior produced using HMM

Characters with behavior **A** exhibit only stock idling animations provided by the NWN game engine, such as stretching their arms. They do not move around.

Characters with behavior **B** exhibit behaviors based on the commercial game Dragon Age: Origins. The two most-lively taverns from this game, one in Lothering (bards entertaining) and one in Redcliffe (a server who walks around) were combined together, forming a tavern with tavern patrons, a tavern server, and bards. Tavern patrons engage in conversations but do not respond to bards and do not move around.

Characters with behavior **C** exhibit complex behaviors as provided by an HMM that was generated after a designer trained the NPCs. There are three kinds of characters:

1) Several tavern patrons exhibit independent behaviors (walking around, saying one-liners to themselves, finding a table to stay at, animating their hands, or facing an object); collaborative behaviors (talking to one another on several topics and talking to bartender to order a drink); and latent behaviors (responding to bards). Fig. 2 shows the latent behavior where the tavern patrons turn and cheer for the two bards as they finish their performance. The cheer behavior is triggered by the bards leaving their performance spotlight.

2) One tavern server exhibits independent (walking around) and collaborative behaviors (talking to tavern patron to fill a drink order).

3) Two bards exhibit independent behaviors (performing on stage under spotlight).

C. *Statistical Techniques*

In order to evaluate our data we used two techniques: analysis of variance (ANOVA) and T-tests.

ANOVA is a commonly used hypothesis testing technique for experimental data. It tests the hypothesis that the means among two or more groups are equal. A test result is statistically significant if it is deemed unlikely to have occurred by chance assuming the truth of the null hypothesis. In our application of ANOVA, the null hypothesis is that all behavior and participant groups are simply random samples of the same population. Rejecting the null hypothesis would imply that there are differences between the groups in terms of their perception of behaviors.

Paired T-tests are used to conduct hypothesis testing for the difference between paired means, where the two groups in the dataset are pair-wise dependent. This fits the case in question, since each participant rates every behavior. The behavior scores given by a single participant are therefore related to that participant. Paired T-tests are used to compare the responses to two behavior variations by the participant groups.

T-tests using paired samples are more powerful than independent or unpaired samples since random noise-factors between different participants have been eliminated. However, to compare the responses of two participant groups (as opposed to the responses to two behavior variations by the same participant group), two-sample unequal variance T-test must be used since the two participant groups are independent of each other.

D. Results and Discussions

1) Comparing Behaviors across Participant Groups

Table I presents the results of the study, showing the three behavior variations. We first analyzed whether all groups across gender and gaming experience find that behavior C is more believable than behaviors A and B. We analyzed the ratings using ANOVA. We found that there are statically significant differences between the behavior variations for the participant gender/gaming experience groups at 95% confidence (p-value < 0.05).

Paired T-tests show that on average, participants believe that behavior C is better than all other behavior variations with 95% confidence, for seven of the eight groups: all females, all males, all gamers, all non-gamers, female gamers, female non-gamers and male non-gamers. The only non-significant result is whether male gamers find behavior C more believable than behavior B (with a p-value of 0.110).

2) Female vs Male Participants on Behavior Improvements

We examined whether gender and gaming experience affect perception of NPC behavior. For males, gaming experience does not affect the perception of behavior quality. Values in the last two columns of Table I are similar on a row-by-row basis. However for females, gaming experience increases their ability to discriminate between behavior differences. There is a significant difference between the 1.42 and 2.26 entries in Table I at a 95% confidence level (p < 0.001) using a two-sample unequal variance T-test, implying that female gamers are less impressed than female non-gamers by the level of behaviors in the commercial game Dragon Age: Origins (behavior B). This may indicate that as females gain more gaming experience they will be more appreciative of better behaviors (find them more believable). In other words, female non-gamers may be just as satisfied with existing commercial game behavior quality as male gamers and male non-gamers, but as females become more experienced gamers they are likely to become more discriminating. Devoting resources to producing more complex behaviors could be a key mechanism for attracting more game players and retaining them as they gain experience. This study indicates that female game players especially recognize more complex NPC behaviors.

It is worth noting that games are often framed in a specific cultural context. This study presented a virtual environment of a medieval tavern. Study participants were asked to judge the NPCs in this specific context. However, the scenes were designed to be as neutral as possible with respect to culture and time periods. The behaviors were categorized in terms of their technical complexity which is not restricted to a particular setting. The tavern could be ported to a modern bar setting and little would change in terms of the behaviors exhibited.

IV. Perception of Emotion

The second study we conducted was a gender dependent version of a user study by Desai and Szafron [7], where we examined whether participants could identify the emotion of a character. The study compared two techniques: emotion-specific gaits and what we call emotional incidents.

Emotional incidents refer to how a character interacts with its environment based on its assigned emotion, such as waving at another character when happy or storming by when angry. We tested these techniques by having a mixed-gender group of participants watch a series of 13 scenes during which a female character walked a set path. In five scenes, the character used one of five gaits (four emotional gaits – happy, sad, angry, afraid and one neutral). Four scenes combined emotional incidents with the neutral gait (each scene represented one of the four emotions). And four scenes combined the emotional incidents with the emotion-specific gait. The scenes were created to mimic the setting of a typical story-based game and were presented in a random order.

Gaits are a valuable source of information about a person’s emotional state [43]. Previous research has shown that the amount of arm swing, stride length, heavy-footedness, and walking speed were all differentiating factors in inferring emotion [45]. While our study participants were all attending university in North America, researchers have shown that gait-based perceptions are similar across different cultures [42].

The results from that study indicated that the combination of gaits and emotional incidents resulted in the highest identification statistics, although some emotions could produce similarly strong results using only a single technique. In order to reduce bias, participants were always given five options for labeling a scene: happy, sad, angry, afraid, or none of these. As participants were not aware of the number of techniques being tested, and there were 13 different scenes, it was difficult for participants to know how many times they should use each label. Many participants made use of the “none of these” option, which implies they did not feel constrained to the four emotions presented.

After analyzing the results of the study, we were interested in whether or not a participant’s gender or the character’s gender would influence the results. We created a new study to test this, and to reduce complexity, we chose to only test the scenes with the combination technique, which used both the emotion-specific gait and the emotional incidents.

The test scenes were created using the Unity game engine [20]. The gait animations were based on the results of a study by Roether et al. [19]. In their study, they filmed a group of actors performing emotion specific walks, and then tested the resulting animations to identify the most important features. However, their animations were run using a character that looked like an artist mannequin, with no defined features. Our characters look much more human. We did not animate the face because the goal was to test which emotions can be accurately identified by looking at the character as a whole, not by focusing on the face. Also, in our experiment, the faces had little screen real-estate and were difficult to see, which discouraged participants from trying to differentiate emotion based on facial animations.

We created eight scenes for our participants to identify. In four of the scenes, the participants were trying to identify the emotion of a female character and in the other four scenes the subject was a male character. Each character (male or female) performs each of the four emotions - once per scene. The characters are both dressed in jeans and a t-shirt. The t-shirts are light grey and light tan, with the color assignment being randomly done at the start of the game. All characters (male or female) were using the exact same script and animations within the scenes. So the male and female versions of the same scene were identical except for the character's gender. Each scene used one of the four emotional gaits - happy, sad, angry or afraid. An example screenshot of the four gaits for the female is shown in Fig. 3 and for the male is shown in Fig. 4.

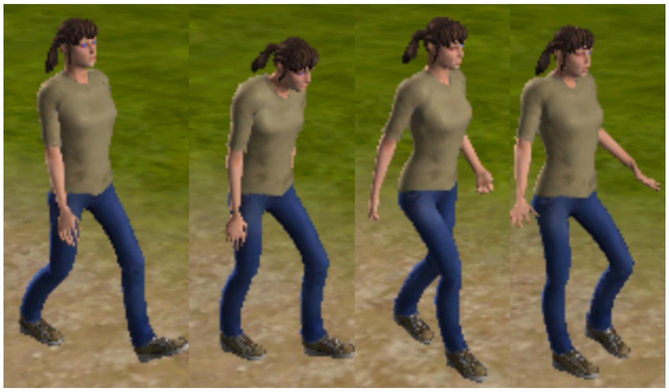

Fig. 3. Female character displaying the four emotion gaits (happy, sad, angry, afraid).

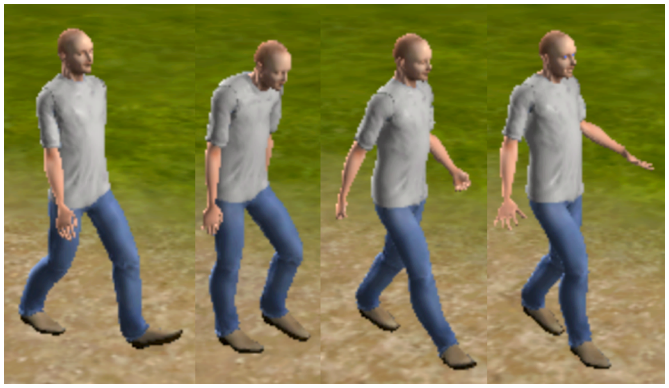

Fig. 4. Male character displaying the four emotion gaits (happy, sad, angry, afraid).

There are two emotional incidents that take place. The first involves the main character walking past a second character, who is sitting on a bench. The seated character waves as the main character approaches. The study focuses on identifying the emotion of the main (walking) character based on how the character responds to the wave of the sitting character (as well as the gait of the main character). The second incident has a small child kicking a soccer ball towards the main character. While the small child (a boy) remains the same between scenes, the gender of the character that is seated on the bench is the opposite of the one who is walking. So when the female character walks the male character sits on the bench and vice versa. The responses to the emotional incidents based on emotion are listed in Table II.

TABLE II
EMOTIONAL INCIDENT RESPONSES

	Character on bench	Kid with ball
Happy	wave, look at character	kick ball to kid
Sad	brief glance, then ignore	pause, look at ball, walk straight
Angry	'glare' at character, speed up to pass	kick ball away
Afraid	slight startle, veer away from bench	startle at ball, then walk around it

Each scene was watched to completion. The length of a scene (ranging from 25 to 55 seconds) depended mostly on the gait speed (e.g. sad characters walked more slowly than angry characters) and on the emotional incident reactions. After watching a scene, participants chose which emotion they thought was being displayed by the walking character. They were given five options: happy, sad, angry, afraid or none of these. They could also choose to re-watch the scene, although this was rarely done (about 4% of the time).

A. Participants

The participants for our study were undergraduate students taking a first year psychology class. They were between the ages of 17 and 31 (mean of 19.7) with 81 females and 81 males. About 45% of the participants reported playing video games at least once a week (24% of females and 67% of males), while 69% of participants reported playing video games at least once a month (51% of females and 86% of males). The study was conducted in a laboratory on a university campus. Each participant sat in front of a computer and observed the characters on the screen in a first-person view. The participants could not interact with the characters or the environment in any way. Participants were given an hour to observe the scenes and to complete the questionnaire.

B. Statistical Techniques

In additional to ANOVA and T-tests, which were both described in the previous section, we used two additional techniques: confusion matrices [21] and bootstrapping [22].

A confusion matrix is used to determine recall and precision for categorical data. Two examples of confusion matrices for our data are shown in Tables IV and V.

Each row in a confusion matrix represents a particular scene, and the row label (happy, sad, angry, afraid) indicates the emotion that we wanted the character to portray. The five columns (happy, sad, angry, afraid, or none) represent how participants classified a scene. RSum is the sum of the row, or how many results for that scene we have. As we had 81 female (and 81 male) participants, the RSum is double this value (81 x 2 = 162) because each participant labeled two scenes for each emotion (one with a male character and one with a female). *Recall* is the ratio of how many participants correctly identified the scene divided by the RSum. The number of correctly identified emotions is on the diagonal, where the row and column labels match (shown in bold). PSum is the sum of a column or how many times participants used that label over all scenes. Finally, *Precision* is the ratio of how often participants used a label correctly divided by the PSum.

The overall precision and recall values are in the bottom right corner. Overall recall is the total number of correctly identified emotions by all participants divided by the total number of participant/scene combinations. Overall precision is the total number of correctly identified emotions by all participants divided by the total number of participant/scene combinations not marked as none.

However, as a confusion matrix only supplies a single value per cell, it is impossible to determine statistical significance. To solve this problem we performed bootstrapping on the data we had collected. Bootstrapping [22] is a method for resampling data that produces 'new' datasets. It creates a new dataset by randomly choosing from the original set with replacement. In our case, we had 162 rows of data, each representing a unique individual. However, as we were interested in comparing male and female participants, we split this original dataset based on gender. This left us with 81 rows for female participants and 81 for males. Using boot-strapping, we created 1000 new datasets of equal size (81 rows per gender). Each row in a 'new' dataset was a row from the original dataset. Because we used sampling with replacement, a new dataset may contain multiple copies of an individual row from the original. Without replacement, each new dataset would be identical to the original. We used this collection of datasets to determine statistical significance.

TABLE III
OVERALL CONFUSION MATRIX FOR FEMALE PARTICIPANTS

	Happy	Sad	Angry	Afraid	None	R Sum	Recall
Happy	**150**	1	1	0	10	162	0.926
Sad	0	**148**	3	6	5	162	0.914
Angry	9	3	**132**	3	15	162	0.815
Afraid	1	0	4	**142**	15	162	0.877
P Sum	160	152	140	151			
Precision	0.938	0.974	0.943	0.940		0.949	0.883

TABLE IV
OVERALL CONFUSION MATRIX FOR MALE PARTICIPANTS

	Happy	Sad	Angry	Afraid	None	R Sum	Recall
Happy	**152**	1	0	1	8	162	0.938
Sad	0	**150**	6	1	5	162	0.926
Angry	9	8	**127**	6	12	162	0.784
Afraid	4	8	2	**117**	31	162	0.722
P Sum	165	167	135	125			
Precision	0.921	0.898	0.941	0.936		0.924	0.843

C. Results and Discussions

We asked two main questions:

Does a participant's gender affect the participant's precision and recall of identifying an NPC's emotion?

Does an NPC's gender affect the participant's precision and recall of identifying the character's emotion?

We ran an ANOVA to determine if the results were influenced by emotion, participant gender, character gender, consistency or participant’s gaming level. The results indicated that emotion, participant gender and gaming level affected the results. There was also an interaction effect between emotion and participant gender and emotion and consistency.

1) Female vs. Male Participants and NPCs

First let's examine the results of female versus male participants. Tables IV and V show the confusion matrices for the female and male participants based on the raw data. Both female and male participants had very high precision values, for all emotions, with none below 89.8%. Since the precision is high in all cases, differences in precision are not interesting.

For recall, although the happy and sad values are above 90% for both female (Table III) and male (Table IV) participants, the results for male participants for the angry and afraid emotions are in the 70's. We used the technique detailed by Hardin and Shumway [13] for comparing confusion matrices using bootstrapping to compare the recall results for female and male participants at a 95% confidence level.

The results show that female participants had significantly higher recall (p-value = 0.003) for afraid than male participants. The recall numbers themselves (from Tables IV and V) are 87.7% for female participants and 72.2% for male participants.

We also compared recall for all participants between male and female characters and found no significant differences for any of the emotions. This matches with what Hall found in her study of previous literature [12]. Although the ANOVA results indicated that there were no interaction effects between character gender, participant gender and emotion, we compared these results and

TABLE V
CONFUSION MATRIX FOR MALE PARTICIPANTS LABELLING FEMALE CHARACTERS

	Happy	Sad	Angry	Afraid	None	R Sum	Recall
Happy	77	0	0	0	4	81	0.951
Sad	0	75	3	1	2	81	0.926
Angry	3	3	68	3	4	81	0.840
Afraid	4	4	2	58	13	81	0.716
P Sum	84	82	73	62			
Precision	0.917	0.815	0.932	0.935		0.925	0.858

TABLE VI
CONFUSION MATRIX FOR MALE PARTICIPANTS LABELLING MALE CHARACTERS

	Happy	Sad	Angry	Afraid	None	R Sum	Recall
Happy	75	1	0	1	4	81	0.926
Sad	0	75	3	0	3	81	0.926
Angry	6	5	59	3	8	81	0.728
Afraid	0	4	0	59	18	81	0.728
P Sum	81	85	62	63			
Precision	0.926	0.882	0.952	0.937		0.924	0.827

TABLE VII
CONFUSION MATRIX FOR FEMALE PARTICIPANTS LABELLING FEMALE CHARACTERS

	Happy	Sad	Angry	Afraid	None	R Sum	Recall
Happy	73	0	1	0	7	81	0.901
Sad	0	73	2	3	3	81	0.901
Angry	4	1	67	2	7	81	0.827
Afraid	0	0	3	71	7	81	0.877
P Sum	81	85	62	63			
Precision	0.948	0.986	0.918	0.934		0.947	0.877

TABLE VIII
CONFUSION MATRIX FOR FEMALE PARTICIPANTS LABELLING MALE CHARACTERS

	Happy	Sad	Angry	Afraid	None	R Sum	Recall
Happy	77	1	0	0	3	81	0.951
Sad	0	75	1	3	2	81	0.926
Angry	5	2	65	1	8	81	0.802
Afraid	1	0	1	71	8	81	0.877
P Sum	81	85	62	63			
Precision	0.928	0.962	0.970	0.947		0.952	0.889

found a couple of significant differences. See Tables V to VIII. First, male participants could identify the angry emotion for female characters (84%) significantly more often (p-value = 0.008) than for male characters (72.8%). And second, female participants could recall the happy emotion for male characters (95.1%) significantly more often (p-value = 0.043) than for female characters (90.1%). However, since ANOVA indicates that these results are NOT significant, a more detailed study is needed to understand whether or not character gender is producing a real effect.

While it can be tempting to try to explain these differences based on animation issues, it is important to know that both characters (female and male) used the same animation files. These animations are based on the animations developed by Roether et al., which were not gender specific (they combined the results of male and female actors).

TABLE IX
DEFINITIONS OF CORRECT AND GENDER CONSISTENT

Character	Happy	Sad	Angry	Afraid
Male	Happy	Happy	None	Afraid
Female	Happy	Sad	None	Afraid
Correct	Both	Female	Neither	Both
Consistent	Yes	No	Yes	Yes

TABLE X
AVERAGE CONSISTENCY BY PARTICIPANT GENDER AND P-VALUE

	Males (%)	Females (%)	P-value
Overall	82.4	84.3	0.255
Happy	92.6	90.2	0.237
Sad	92.7	88.9	0.176
Angry	72.9	74.1	0.393
Afraid	71.5	83.8	0.024

These results indicate that if you want a player to perceive a character's emotion (such as angry), you may need to exaggerate it more for male players than for female players, regardless of the gender of the NPC who is displaying the emotion.

2) Gender Consistency

We also examined whether players of different genders were equally consistent in identifying NPCs of both genders. A participant was considered ‘gender consistent’ if they gave the same label to both the female and male character when they exhibited the same emotion. In this comparison, it did not matter if the participant was correct, just that the same emotion was selected for both the male and female NPCs (correctness was examined in the confusion matrix above using precision and recall). Table IV illustrates of our definition of correct and gender consistent. We analyzed the data for gender consistency and made some interesting observations. First, 44% of participants did not make any recall errors. However, of the 56% of the participants who misidentified an emotion, approximately 81% of them made a gender consistency error.

We started by asking whether females or males were more gender consistent. The result was insignificant (p=0.255), which suggests that the two genders produced similar levels of consistency. After examining the overall results, we looked at the individual emotions to see if there were any differences between the genders. Here there was one significant result. Male participants were significantly (p=0.024) less consistent when it came to identifying afraid characters compared to female participants. The p-values for all of the emotions (as well as overall) are listed in Table X. This result is not entirely surprising, given that male participants had only 72.2% recall in identifying afraid characters compared to the 87.7% achieved by female participants.

3) Other Factors besides Gender

We considered dividing our results into other groups, to see if the differences could be explained by other factors besides gender. However, our participant pool was quite homogenous. Everyone was about the same age (min 17, max 31, mean 19.7, mode 18), same year of study (min 1, max 5, mean 1.97, mode 1), all taking a first year psychology class, and all admitted to the same university. While we did not collect cultural information on the students, based on observation, the percentage of international students was roughly equivalent between the male and female groups. This meant that gamers and non-gamers was the single identifiable factor that could be significant, besides gender. We wondered if, by playing more games, participants were possibly being desensitized towards normal emotional cues displayed by NPCs.

We split all of our participants into two groups - gamers and non-gamers. We defined gamers as those who stated that they play video games a minimum of once a week. We then re-constructed our confusion matrices and bootstrapping analysis. However, unlike the situation where we compared genders, we found no statistically significant differences between gamers and non-gamers. This suggests that the differences in the gender analysis were not the result of a non-intended consequence of gamer versus non-gamer bias, but were the result of gender differences.

V. Conclusion

The paper examines the effects of gender on the perception of video game character emotion and behavior. We describe the results of two studies on the perception of NPCs, one on behavior and one on emotion. The results of this research can be used to improve the quality of game experiences. Overall, the results from the studies suggest that designers of story-based games need to consider the gender of their audience when designing behaviors.

Female gamers have a higher appreciation for complex behaviors of NPCs. This suggests that resources allocated to producing more believable behaviors will enhance the game-play experience of female game players.

While both males and females had similarly high precision in identifying emotions, the significant differences between male and female recall for the afraid emotion and the differences in perception of angry and happy NPCs of different genders indicate that video game designers must be careful in how they try to convey emotion.

Of the 56% of participants who failed to identify a specific emotion, 81% of the errors were gender-consistency errors. This indicates that testing of emotional cues must include both male and female NPCs and both male and female participants to be sure the cues will be read accurately.

We suggest that game designers should proceed with caution and not assume that a single emotion-specific animation or reaction will be sufficient for both female and male players and female and male NPCs.

The results of both studies indicate that if game designers want the players to perceive behaviors in a particular way (high quality, emotion identification), they should consider the gender of the prospective players. Female gamers have a greater appreciation for complex behaviors. There are some emotions that females perceive better than males, and within gender, there are some emotions that are perceived more easily from female characters than male characters (and vice versa).

VI. Acknowledgement

This research was supported by GRAND Network of Centres of Excellence and NSERC. We thank members of the BELIEVE research group at the University of Alberta.

References

[1] C. Basak, W. Boot, M. Voss, and A. Kramer. Can training in a real-time strategy video game attenuate cognitive decline in older adults? Psychology and aging, 23(4):765, 2008.
[2] Bethesda Softworks LLC. The Elder Scrolls V: Skyrim, 2011 [Online]. Available: http://www.elderscrolls.com/skyrim
[3] BioWare. Neverwinter Nights, 2002 [Online]. Available: http://www.bioware.com/en/games/#game-neverwinter-nights
[4] BioWare. Mass Effect 3, 2012 [Online]. Available: http://masseffect.bioware.com/
[5] C. Bonk and V. Dennen. Massive multiplayer online gaming: A research framework for military training and education. Technical report, DTIC Document, 2005.
[6] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovi_c, et al. Predicting protein structures with a multiplayer online game. Nature, 466(7307):756-760, 2010.
[7] N. Desai and D. Szafron. Enhancing the believability of character behaviors using non-verbal cues. In Proceedings of the Eighth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 2012.
[8] M. Dunn. Bioware using facebook to select the look of mass effect 3 main character, 7 2011.
[9] P. Ekman. An argument for basic emotions. Cognition & Emotion, 6(3-4):169-200, 1992.
[10] Entertainment Software Association. Essential facts about the computer and video game industry. Technical report, 2014.
[11] J. Feng, I. Spence, and J. Pratt. Playing an action video game reduces gender dierences in spatial cognition. Psychological Science, 18(10):850-855, 2007.
[12] J. Hall. Gender effects in decoding nonverbal cues. Psychological bulletin, 85(4):845, 1978.
[13] P. Hardin and J. Shumway. Statistical significance and normalized confusion matrices. Photogrammetric engineering and remote sensing, 63(6):735-739, 1997.
[14] T. Hartmann and C. Klimmt. Gender and Computer Games: Exploring Females' Dislikes. Journal of Computer-Mediated Communication, 11(4):910-931, 2006.
[15] T. Mahlmann, A. Drachen, J. Togelius, A. Canossa, and G. N. Yannakakis. Predicting player behavior in tomb raider: Underworld. In Computational Intelligence and Games (CIG), 2010 IEEE Symposium on, pages 178-185. IEEE, 2010.
[16] J. McGonigal. Ted talks - gaming can make a better world, 02 2010.
[17] mtvU. Darfur is dying. 2009 [Online]. Available: http://www.darfurisdying.com/
[18] J. Orkin and D. Roy. Automatic learning and generation of social behavior from collective human gameplay. In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, pages 385-392. International Foundation for Autonomous Agents and Multiagent Systems, 2009.
[19] C. Roether, L. Omlor, A. Christensen, and M. A. Giese. Critical features for the perception of emotion from gait. Journal of Vision, 9(6):1-32, 06 2009. reviewed.
[20] Unity Technologies. Unity. 2014 [Online]. Available: http://unity3d.com/
[21] C. J. van Rijsbergen. Information retrieval, 1979. [Online]. Available: http://www.dcs.gla.ac.uk/Keith/Preface.html
[22] H. Varian. Bootstrap tutorial. Mathematica Journal, 9(4):768-775, 2005.

[23] H. Wallbott and K. Scherer. Cues and channels in emotion recognition. Journal of Personality and Social Psychology; Journal of Personality and Social Psychology, 51(4):690, 1986.

[24] R. Zhao and D. Szafron. Generating believable virtual characters using behavior capture and hidden markov models. Lecture Notes in Computer Science 7168, pp342-353. Springer, 2011.

[25] L.E. Baum, T. Petrie, G. Soules, N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. 41(1), 164–171. 1970.

[26] BioWare. Dragon Age: Origins, 2009 [Online]. Available: http://dragonage.bioware.com/dao/

[27] M. Cutumisu and D. Szafron. An Architecture for Game Behavior AI: Behavior Multi-Queues. Proceedings of the Fifth Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE-09), Stanford, USA, October, 2009, pp20-27.

[28] M. Mateas, and A. Stern. A Behavior Language for Story-based Believable Agents. Intelligent Systems, IEEE, 17(4), pp39-47. 2002.

[29] D. Isla. Handling complexity in the Halo 2 AI. In Proceedings of the GDC 2005. Gamasutra. [Online]. Available: http://www.gamasutra.com/view/feature/130663/gdc_2005_proceeding_handling_.php

[30] J. Orkin. Three States and a Plan: The AI of F.E.A.R. Game Developers Conference (GDC-2006), 2006.

[31] J.P. Kelly, A. Botea, S. Koenig. Offline Planning with Hierarchical Task Networks in Video Games. In Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE-08), 2008.

[32] A. Coman and H. Munoz-Avila. Plan-Based Character Diversity. In Proceedings of the Eighth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-12), 2012.

[33] P. Gomes and A. Jhala. AI Authoring for Virtual Characters in Conflict. In Proceedings on the Ninth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-13), 2013.

[34] S. M. Ogletree and R. Drake. "College Students' Video Game Participation and Perceptions: Gender Differences and Implications" Sex Roles: Volume 56, Issue 7-8, pp 537-542. April 2007.

[35] S. Bakkes, P. Spronck, and J. van den Herik, "Rapid and reliable adaptation of video game AI." IEEE Trans. Comput. Intell. and AI in Games, vol. 1, no. 2, pp. 93–104, 2009.

[36] M. Ochs, N. Sabouret, and V. Corruble. "Simulation of the Dynamics of Nonplayer Characters' Emotions and Social Relations in Games." IEEE Trans. Comput. Intell. and AI in Games. 1 (4): pp 281-297. 2009.

[37] J. Bates. "The role of emotion in believable agents." Commun. ACM, vol. 37, no. 7, pp. 122–125, 1994.

[38] J. McCoy, M. Treanor, B. Samuel, A. A. Reed, M. Mateas, and N. Wardrip-Fruin. "Prom Week: Designing past the game/story dilemma", Proceedings of the 8th International Conference on the Foundations of Digital Games (FDG). pp. 94-101, 2013.

[39] J. McCoy, M. Treanor, B. Samuel, A. A. Reed, M. Mateas, and N. Wardrip-Fruin. Social Story Worlds With Comme il Faut, IEEE Transactions On Computational Intelligence and AI in Games, vol. 6, no. 2, 2014.

[40] M. Eladhari, H. Verhagen, J. McCoy, M. Johansson. Introduction. In: Proceedings of Social Believability in Games Workshop. Proceedings of the 9th International Conference on the Foundations of Digital Games (FDG). 2014.

[41] M. Johansson, H. Verhagen. Social believability in games — the early years. In: Proceedings of Social Believability in Games Workshop. Proceedings of the 9th International Conference on the Foundations of Digital Games (FDG). 2014.

[42] J. M. Montepare, L. A. Zebrowitz. A cross-cultural comparison of impressions created by age-related variations in gait. Journal of Nonverbal Behavior. March 1993, Volume 17, Issue 1, pp 55-68.

[43] C. D. Barclay, J. E. Cutting, and L. T. Kozlowski. Temporal and spatial factors in gait perception that influence gender recognition. Perception and Psychophysics, 23, 145-152, 1978.

[44] S. M. Nowlis, B. E. Kahn, R. Dhar. Coping with Ambivalence: The Effect of Removing a Neutral Option on Consumer Attitude and Preference Judgments. The Journal of Consumer Research, Vol. 29, No. 3, pp. 319-334, 2002.

[45] J. M. Montepare, S. B.Goldstein, and A. Clausen. The identification of emotions from galt information. Joumal of Nonverbal Behavior, I 1, 33-42, 1987.

Neesha Desai received her PhD in Computing Science from the University of Alberta in 2015. Her research focused on creating believable characters in computer games and tools to assist non-programmers to create their own games. She now works in educational technology. She cofounded Alieo Games (a writing game for elementary students) and now helps EdTech companies create fun, engaging games for their games.

Richard Zhao received a BSc in Computer Science from the University of Toronto in 2007, a MSc and a PhD in Computing Science from the University of Alberta, Edmonton, in 2009 and 2015. He is currently an assistant professor at the Pennsylvania State University at Erie, The Behrend College. His research includes the use of artificial intelligence techniques to produce more believable behaviors for virtual characters in story-based games and simulations. His research interests include machine learning, reinforcement learning, planning and scheduling techniques with applications to video games.

Duane Szafron received a PhD in Applied Mathematics from the University of Waterloo in 1978. He is currently a Professor of Computing Science at the University of Alberta. He has been doing research in object-oriented computing since 1980, including language design, language implementation, programming environments and parallel computing. His current research interests are in computer games, especially believable characters in computer games and computer poker. He teaches computing courses to students at all levels, from first year through graduate school.