Toward Avatar Models to Enhance Performance and Engagement in Educational Games

Dominic Kao and D. Fox Harrell
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, USA
{dkao,fox.harrell@mit.edu}

***Abstract*—This paper presents work toward better understanding the roles that avatars can play in supporting learning in educational games. Specifically, the paper presents results of empirical studies on the impact of avatar type on learner/player performance and engagement. These results constitute work establishing baseline understandings to inform our longer term goal of developing models that use dynamic avatars to best support learners in educational games. Our aim is motivated by a convergence of research in the social sciences establishing that identity plays an important role in learning. Of note, aspects of social identity (e.g., race, ethnicity, and gender) have been shown to impact student performance [1] via triggering stereotypes [2]. Recently, performance and engagement studies in our educational game for Science, Technology, Engineering and Mathematics (STEM) learning suggest these same phenomena can be activated through virtual avatars [3], [4]. Here, we present results of a comparative study between avatars in the likeness of players and avatars as geometric shapes. In our STEM learning game, results show that players that had selected and used a shape avatar had significantly higher performance than players that had customized and used a likeness avatar. Players using the shape avatar also had significantly higher self-reported engagement, despite having lower self-reported affect towards the avatar.**

I. Introduction

Player models can be powerful predictors of human behavior that can augment user engagement [5], [6]. However, inputs to player models have predominantly been gameplay data, physiological signals, and player profile information [7]. We argue that the social science literature suggests that a more interdisciplinary approach can provide more robust player models. This paper aims to lay the groundwork for a new type of player model in educational games that take into account the “real” (sociocultural) and virtual identities of learners. However, in order to build models that take these phenomena into account, baseline understandings and best practices need to be discovered. Here, we contribute to that end through an empirical study on how avatar types can affect player performance and engagement.

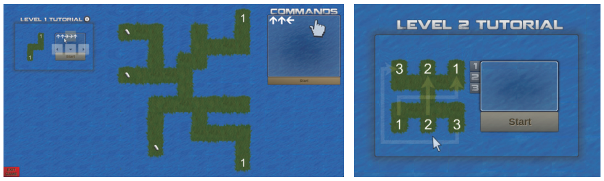
Stereotype threat, the theory that the mere *idea* of conforming to a stereotype can hinder one’s performance, is well-studied in the social sciences [8]. Studies have shown that children as young as five to seven years old are affected. In one experiment, the five to seven year old girls were given a picture to color prior to doing an age-appropriate math test. The girls were assigned either a picture of a landscape, or a picture of a girl their age holding a female doll. The girls that colored the picture with a doll performed significantly worse on the test, i.e., the image invoked gender stereotypes [9]. These effects are not limited to a particular domain, such as about girls being worse at math, nor are their limited to a particular social group. Our recent results have suggested that stereotype threat can be activated by avatars in a Science, Technology, Engineering and Mathematics (STEM) learning game, resulting in lower self-reported engagement [3], [4]. Stereotype threat may be especially harmful in educational games, whereby its effect can translate into dropping out and missed learning opportunities. Here, in light of recent evidence, we suggest that “real” and virtual identity are important elements to consider in modeling players in educational games.

This paper makes several contributions to computational intelligence in games. First, we perform a comparative study between avatars in the likeness of players (anthropomorphic) and avatars as geometric shapes (abstract). Anthropomorphism has emerged as an important distinguishing characteristic in avatars [10], [11], but research comparing avatars along this dimension is severely limited. In our STEM learning game, results show that players that had selected and used a shape avatar had significantly higher performance than players that had customized and used a likeness avatar. Players using the shape avatar also had significantly higher self-reported engagement, despite having lower self-reported affect towards the avatar.

Second, we performed natural language processing and sentiment analysis on players’ linguistic descriptions of their avatars. This was done to better understand how players perceive and identify with the two avatar types. We used a well-known linguistic analysis tool called Linguistic Inquiry Word Count (LIWC) to rank the top ten dimensions on which the avatar types differ. Players using avatars in the likeness of themselves often use 1st person singular (e.g., I, me, my) and present tense, whereas players using avatars as geometric shapes often use articles (e.g., a, an, the), impersonal pronouns (e.g., it, it’s, those) and past tense.

Third, we use supervised learning to predict the number of levels that players complete in our STEM learning game. We did this to build and test an exploratory player model using attributes of virtual and social identities. We used a number of learning algorithms to predict how many levels players would complete. We achieve moderate success, im-

978-1-4799-8622-4/15/$31.00 ©2015 IEEE



(a) *Mazzy*'s first level. (b) Animated tutorial.

Fig. 1: In *Mazzy*, players write "code" to navigate a maze.

proving up to 11.9% over baseline accuracy using decision tree learning. Ranking features using a single-attribute evaluator demonstrates that modeling avatar type is crucial; prediction using avatar type alone adds 7% to baseline accuracy.

The rest of this paper is structured as follows: Section II provides background information on the application domain of our STEM learning game. Section III covers related work on virtual identities, social identities, and task performance. Section IV describes our methodology for this study. Section V presents the results and analyses of our study. Section VI discusses the possible implications of our findings for educational games and player modeling. We conclude with a summary in Section VII, and discuss about potential future work in Section VIII.

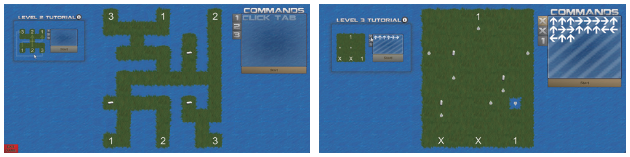
II. The Game

The game we created is called *Mazzy*. *Mazzy* is a STEM learning game, designed to be a fun game, but also to foster computational thinking. *Mazzy* has been used as an experimental testbed for evaluating the impacts of avatar type on performance and engagement [3], [4]. Players use the keyboard to write procedures to guide a character to reach the end of a maze (see Figure 1).

Mazzy's design is grounded in an influential pedagogical approach called "constructionism," in which building objects is central to the process of learning [12]. Constructionism originates in the principle that learning a new concept or idea is easier if it can be assimilated into existing models [13]. In *Mazzy*, the character is "body syntonic" [14]; this means players can identify with it and its motion in space. Players are learning computing by creating programs via a real concrete object that can be manipulated.

Mazzy uses symbols to represent code instead of natural language. This has the advantage of being very simple, since the notion of misspelling a command or forgetting a closing bracket (known typically as "syntax" errors) does not exist in *Mazzy*. This also makes code easily learnable since the symbols are meant to represent their purpose. When players run a program, each symbol is highlighted as it is processed (similar to "debugging"). This stems from the philosophy that building systems is an iterative process, and that things almost never work on the first try.

Three levels have been implemented in the current version of *Mazzy* (see Figure 2). Adding new levels is not technically difficult, although developing levels that are both fun and effective for learning requires skillful game design. Each of the levels features animated tutorials to guide the player. Bonus items challenge players to solve levels in a more complex manner. Levels become harder; for instance level two requires the player to program multiple characters in parallel, and level three requires the player to program boolean logic into the level map. The game is challenging; on average players finish about 1.5 levels.



(a) Level 2. (b) Level 3.

Fig. 2: Levels in *Mazzy*.

III. Virtual and "Real" Identities

We shall investigate if virtual identities can impact players' behaviors (performance on educational activities and learning), attitudes (affect toward the learning experience), and identities as STEM practitioners. This focus necessitates understanding the relationships between sociocultural and virtual identities.

A. Blended Identities

Harrell describes digital self-representations as selective projections of some aspects of a real player (e.g., preferences, control, appearance, personality, understanding of social categories, etc.) onto the actual implemented, virtual, representation [15]. As such, Harrell's notion of a "blended identity" is an approach based on looking at structural mappings from one domain to another that is central to the understanding of virtual identities in this project [16]. This concept builds upon James Gee's notion of the "projective identity", which can be described as "manifesting the ways that real player values are reconciled with values understood as being associated with avatars." [17], [18]. Relating in-game behavior to real-world identities, such as demographic segments [19], [20] has demonstrated useful insight into understanding how to match interaction mechanisms in digital media systems such as games to users in order to provide the most appropriate supports. Such supports can have strong impacts on user behaviors, such as has been shown by research on the "proteus effect", a phenomenon in which users conform to expected behaviors and attitudes associated with an avatar's appearance [21]. Here, our focus is on matching avatar uses with supports for computer science learning by diverse players.

B. Stereotype Threat

This work motivates some of our efforts investigating the impacts of avatars on players. We are motivated by the fact that users' representations may act as triggers prompting more positive or negative outcomes depending on the social group of the user as visually represented by the avatar's appearance. In other words, we are building toward addressing the impact of avatars on stereotype threat. Stereotype threat [2] can lead to a number of harmful consequences, ranging from decreased performance (e.g., women performing worse in math when their female identity is made salient) [22] to altered professional aspirations (e.g., stereotype threat undermines sense of belonging and reduces women's desire to pursue math in the future) [23].

Techniques such as deemphasizing threatened identities [24] and endorsing an incremental view of intelligence [25] have been seen to reduce, and in some cases eliminate, stereotype threat. Our preliminary work suggests that stereotype threat persists in virtual environments. This is consistent with the observation of stereotype-related phenomena identified by other researchers, e.g., Yee demonstrated in [26] that often players' behaviors conform to stereotypes associated with their avatar's gender. In this work, our systematic study of the impacts of virtual identities on learners will enable us to develop systems that could help in inoculating users against stereotype threat in STEM learning.

C. Avatar Impacts on Engagement and Performance

To the best of our knowledge, there has not been extensive work on the impacts of avatars on player engagement and performance. Linebarger et. al compared four avatar types on task performance in a virtual environment and concluded that "simpler, less computationally expensive avatar representations are quite adequate" [27]. More recently, Domínguez et. al explored the impact of avatar color on performance in a virtual scavenger hunt, although their results are so far "not conclusive" [28]. Previous studies using *Mazzy* as an experimental testbed suggest that "face photo" avatars can prompt more negative emotional dispositions towards the game [4]. However, task performance was not different across avatar.

IV. Methodology

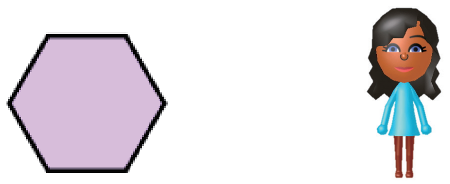
The study we performed consisted of two experiments (N=508) inside of our educational game *Mazzy*. The study compares the impact of selected shape avatars and customized likeness avatars on player engagement and performance.

A. Avatar Conditions

The two avatar conditions we tested were:

1) Likeness: Avatar in the likeness of the player.
2) Shape: Avatar as a geometric shape.

The likeness condition consisted of a Mii avatar. A Mii is a character developed by Nintendo, chosen since Miis were designed with the intention of looking similar to users (Mii is a blend of "Wii" and "me"). Players were asked to use a publicly available Mii customization system prior to the task [29]. Furthermore, players were told to create an avatar that looked like themselves and that this avatar would be used in the subsequent game. The shape condition was a geometric shape; players picked out of eight possible geometric shapes [30]. These players were also told the shape that they picked would be their avatar in the ensuing game. See Figure 3 for examples of these.



(a) A sample Shape avatar. (b) A sample Likeness avatar.

Fig. 3: Sample avatars.

B. Task

The experimental task was to play *Mazzy*[1]. There are three levels in this version of *Mazzy*. In the first level, players can click in the command box, after which they can use the arrow keys on the keyboard to input arrow commands. When participants click on the "start" button, the character begins to move according to the programmed arrows. The arrows are highlighted as each command is processed by the character. The character continues to move until either a) the character moves into a water tile, b) the character reaches the end of the maze, or c) the character has no more commands to process and has not reached the end. In case b), the player can advance to the next level. In any other case, the character disappears and the player should try again.

The second level is a direct extension of the first; players can now program three characters, all starting in different spots in the maze, and all having different goal locations. These start and corresponding end locations are marked with the same number. Beside the command box are three smaller buttons, clicking each of these brings up the code for each of the three characters. Clicking inside the command box allows players to modify the code for a single character. The same rules as in the first level apply to each of the three characters. In the second (and third) level, commands are highlighted for the character whose code is currently on-screen. The player may switch between each character's code view during execution.

The third level is similar to the second level in that there are three characters. However, all of them are already preprogrammed. The code can be viewed for all three characters, but their code cannot be modified. Two of the characters start at locations marked with an "x" (no corresponding end location), and one character starts at a location marked "1" (with a corresponding end location). The player can click *the map itself* to toggle some tiles to either be water or grass. There are some combinations of toggles that allow only the character starting at "1" to reach the end location; doing so passes the level and the game. Toggling of these particular tiles can be done during code execution.

Bonus items are scattered in each level, which the players can optionally pick up. There are a total of nine bonus items, three in each level. In all levels, there is an animated tutorial in the top left of the screen demonstrating the mouse clicks

[1] http://groups.csail.mit.edu/icelab/mazzy/

and keyboard presses required to solve a simpler version of the current level; these levels have the same mechanics, only the mazes are reduced in complexity. Mousing over a help icon next to this animation provides a textual description of the goal for each level.

C. Quantitative and Qualitative Measures

The performance measures we recorded were:

- **Levels completed:** The number of levels completed.
- **Level attempts:** The number of attempts in each level.
- **Level bonus items:** Bonus items collected in each level.

The engagement measures we recorded were:

- **Enjoyment:** Enjoyment rating in each level.
- **Difficulty:** Difficulty rating in each level.

All subjective data was collected using a 5-point Likert scale. These engagement measures (enjoyment and difficulty) were the only engagement data collected in the first experiment. In the second experiment, at the end of the study, players were also asked to rate how they felt overall with respect to the game, their progress, and their avatar, in addition to describing their avatar in text and completing a demographics survey.

D. Participants

508 participants (250 in the first experiment, 258 in the second experiment) were recruited through Mechanical Turk. 38% of the participants were female. 77% of participants were white, 9% black or African American, 5% Chinese, the remaining participants were divided amongst eleven other group categories. Participants were between the ages of 18-68 (M = 31.6) and were reimbursed $2 to participate.

E. Design

A between-subjects design was used: avatar type was the between-subjects factor. Participants were randomly assigned to conditions (i.e., random assignment of avatar type).

F. Experiment Protocol

Prior to starting the task, players were told they could exit the game *at any time*. Then, for each condition players loaded the game in their web browser. After each level that players completed, players were presented with a screen showing the number of "stars" they had earned (corresponding to the number of bonus items they had collected); at this point in the procedure, players could either continue or replay the level. If they chose to replay the level, they were brought back to the previous level (with their previous code still intact). If they continued, they were then asked to report engagement (enjoyment and difficulty). When participants were done playing, they returned to the instructions, which prompted them with additional questions including the demographic survey.

G. Analysis

Our analysis consists of independent-samples t-tests, and results are reported as significant when $p<0.05$ (two-tailed). Furthermore, we perform linguistic analysis and supervised learning as described below.

1) Natural Language Processing: The text we want to analyze are players' linguistic descriptions of their avatars, typically 2-3 sentences long. In order to interpret these, we leverage a text analysis system called Linguistic Inquiry Word Count (LIWC). LIWC is a popular tool in psychology. LIWC was developed over the last couple decades by human judges that categorize common words [31], [32]. LIWC matches text to 82 language dimensions; these range from affective processes (i.e., positive emotion, negative emotion, anxiety, etc.) to part-of-speech (i.e., articles, past tense, present tense, etc.) to thematic categories (i.e., achievement, money, death, etc.). Pennebaker et. al performed one of the earliest text analyses, using sources such as daily diaries and journal abstracts; they found that linguistic style is a "meaningful way of exploring personality" [33]. In our case, we leverage LIWC to analyze players' descriptions of their game avatars; we are interested in exploring how players perceive themselves in relation to their avatars. We use LIWC to calculate scores for each player individually, then present the averages for each condition in the results. Given the large number of language dimensions analyzed by LIWC, we present only results from ten of the dimensions that have the highest difference in score between avatar types.

2) Prediction Algorithms: Here, we are looking to test the effectiveness of a player model that incorporates social and virtual identity in predicting when players will quit our game. In order to make these predictions, we must select some subset of machine learning algorithms to train and test on. We use the WEKA machine learning workbench (version 3.7.12). WEKA was developed at the University of Waikato [34], and contains a collection of machine learning algorithms for data mining tasks. This version of WEKA has by default over 50 different classification algorithms. Furthermore, WEKA's package manger gives access to an additional set of classification algorithms; this makes the total number of classification algorithms available close to 100. Given the large number of choices, we use a similar approach to Mahlmann et. al in that we consider at least one algorithm from each of the families of algorithms [35]. Similarly, we pay especially close attention to algorithms found on the list of top ten data mining algorithms: SVMs, decision trees, belief networks, etc. [36]. The specific attributes used in the algorithms is as follows:

- **Avatar Type:** The avatar type (Likeness, or Shape).
- **Avatar Shape:** The avatar sub-type. For likeness avatars, these were coded as "Mii"; for shape avatars, these were coded as "Triangle", "Square", "Pentagon", etc.
- **Level One Enjoyment:** Player reported enjoyment in level 1.
- **Level One Difficulty:** Player reported difficulty in level 1.
- **Level One Stars:** Number of bonus items in level 1.
- **Level One Attempts:** Number of attempts in level 1.
- **Level One Successful Attempts:** Number of succ. attempts in level 1.
- **Player Age:** The player's age.
- **Player Gender:** The player's gender.
- **Player Race:** The player's race.

We used a simple single-attribute evaluator called 1R to rank these attributes by importance. 1R generates a one-level decision tree that splits on a single attribute (i.e., all predictions for that tree depend only on that specific attribute). 1R has been shown to perform well vis-à-vis more complex algorithms [37]. We use the 1R evaluator on each attribute individually; we then rank those attributes by their prediction scores, giving us a rough approximation of each attribute's merit.

V. Results

A. Experiment 1

Players reported higher engagement in the shape condition. Players in the shape condition (M=3.26, SD=0.96) reported significantly higher enjoyment than participants in the likeness condition (M=2.89, SD=1.11), t(205)=2.35, p=0.02, r=0.16. No other significant differences were found. See Table I.

TABLE I: Results from the first experiment.

Attribute	*L*-Mean	*L*-SD	*S*-Mean	*S*-SD	t-test
Levels Completed	1.90	1.14	1.96	1.16	0.43
Average Enjoyment	2.89	1.11	3.26	0.96	**2.35***
Average Difficulty	2.34	0.95	2.32	0.86	0.26
Total Bonus Items	3.10	3.28	3.18	3.38	0.19
Total Attempts	21.87	21.60	18.65	15.21	1.40

*<.05, **<.01, L = Likeness, S = Shape, SD = Standard Deviation

B. Experiment 2

Players had higher performance and engagement in the shape condition. Players had lower affect towards the shape avatar. Players in the shape condition (M=1.65, SD=1.07) completed significantly more levels than participants in the likeness condition (M=1.08, SD=1.01), t(256)=4.42, p=0.0001, r=0.27. As a result, players in the shape condition (M=16.14, SD=12.86) had more total attempts than participants in the likeness condition (M=12.38, SD=10.99), t(255)=2.52, p=0.01, r=0.16. Players in the shape condition (M=3.45, SD=1.02) rated the game higher than participants in the likeness condition (M=3.07, SD=1.10), t(253)=2.89, p=0.004, r=0.18. Players in the shape condition (M=3.34, SD=1.03) also rated their progress higher than participants in the likeness condition (M=3.06, SD=1.08), t(252)=2.09, p=0.038, r=0.13. Players in the likeness condition (M=3.61, SD=0.94) rated their avatar higher than participants in the shape condition (M=3.06, SD=0.87), t(254)=4.84, p=0.0001, r=0.29. Overall trends remain consistent across both experiments. See Table II.

TABLE II: Results from the second experiment.

Attribute	*L*-Mean	*L*-SD	*S*-Mean	*S*-SD	t-test
Levels Completed	1.08	1.01	1.65	1.07	**4.42****
Average Enjoyment	2.86	0.88	3.05	0.95	1.44
Average Difficulty	2.15	0.82	2.23	0.88	0.62
Total Bonus Items	1.99	2.73	2.69	3.08	1.93
Total Attempts	12.38	10.99	16.14	12.86	**2.52***
Avatar Rating	3.61	0.94	3.06	0.87	**4.84****
Progress Rating	3.06	1.08	3.34	1.03	**2.09***
Game Rating	3.07	1.10	3.45	1.02	**2.89****

*<.05, **<.01, L = Likeness, S = Shape, SD = Standard Deviation

C. Text Analysis

Table III contains a summary of text analysis results on players' descriptions of their avatars. Figures 4 and 5 are word clouds of players' avatar descriptions. Common english words, as well as the words "avatar" and "game" have been removed from these clouds to highlight differences.

TABLE III: Top ten dimensions (ranked by difference) from natural language processing using LIWC.

Attribute	Example	*L*-Mean	*S*-Mean
Biological Processes	Eat, hands, pain	5.87	0.83
Impersonal Pronouns	It, it's, those	4.12	9.05
Articles	A, an, the	8.44	13.25
Present Tense	Is, does, hear	9.15	5.02
3rd Person Singular	She, her, him	4.09	0.16
Past Tense	Went, ran, had	5.32	9.06
Social Processes	Talk, they, friend	7.23	3.95
Space	Down, in, thin	5.32	8.49
Feel	Feels, touch	3.11	0.45
1st Person Singular	I, me, my	8.55	5.96

L = Likeness, S = Shape

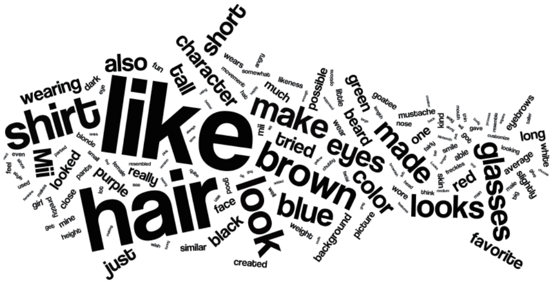


Fig. 4: Words used to describe likeness avatars. Larger corresponds to higher recurrence.

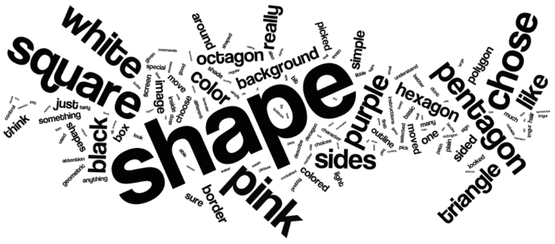


Fig. 5: Words used to describe shape avatars. Larger corresponds to higher recurrence.

D. Level Prediction

To determine the usefulness of modeling aspects of social and virtual identities, we built a player model using only statistics from the first level. We then ran a number of machine learning algorithms to determine if we could predict the final level completed. This involved removing those participants that did not complete the first level (there were 73 such

TABLE IV: Number of players that stopped playing after each level.

Level	1	2	3
Number of Players	65	74	46

participants). This left us with 185 data points, spread out across the remaining levels. See Table IV for this distribution.

In order to have a useful player model, it should outperform at least the baseline accuracy. The baseline accuracy is calculated by finding the final level with the highest number of players, then dividing that by the total number of players. Here, the baseline accuracy is 74/185 (40%). To perform the actual prediction, we picked algorithms from each family as described earlier. We use 10-fold stratified cross-validation in all cases. Parameters in each algorithm are either left at default or tuned manually lightly. See Table V for results of these prediction algorithms. Results show that decision tree learning performs 11.9% above baseline, and that non-linear classification using support vector machines performs 11.3% above baseline. Many of the algorithms, such as multinomial logistic regression and k-nearest neighbours classification, performed only marginally better than baseline.

Next, we used attribute selection using the 1R algorithm to rank the individual attributes by score. See Table VI for these ranked scores. Baseline accuracy was 40%. Thus, knowledge of avatar type alone gives us an improvement over baseline by 7%. Attributes such as bonus items collected, reported engagement (enjoyment and difficulty), and gender were the least effective individual predictors.

TABLE V: Prediction accuracy of various machine learning algorithms. Higher means that the algorithm performed better.

Algorithm	Accuracy
C4.5	51.9%
LibSVM	51.3%
Random Forest	47.6%
Bayes Network	47.0%
Multilayer Perceptron	45.4%
k-Nearest Neighbors	42.7%
Logistic Regression	42.7%
Baseline	40.0%

TABLE VI: The 1R attribute evaluation scores for each feature.

Attribute	1R Score
Avatar Type	47.03
Avatar Shape	44.32
Level One Attempts	43.24
Player Age	43.24
Player Race	41.08
Level One Succ. Attempts	41.08
Level One Difficulty	39.46
Player Gender	36.76
Level One Enjoyment	36.76
Level One Stars	34.60

VI. Discussion

We now discuss some of the broader implications of this work, both for educational games and player modeling.

A. Implications for Educational Games

The results suggest that avatar type has a significant impact on user performance and engagement in our STEM learning game. This has important implications; level completion in an educational game can be seen as evidence that learning has occurred (so long as the task is novel). Therefore, understanding virtual identities' impacts may be crucial in better understanding how they affect learners in educational games.

We might ask why specifically there was a large, measurable difference in performance and engagement between these two avatar types. Depending on the point of view one takes, this can be explained by a number of phenomena. Bowman et. al suggest that avatars more like "objects" cause players to focus more on in-game mechanics and challenges ("pleasures of control") [38]. There is evidence of this in the text analysis. Players detailing their shape avatars are more likely to use impersonal pronouns (e.g., it, it's, those) and articles (e.g., a, an, the), and less likely to use first person singular (e.g., I, me, my). Failure in the game (which is almost guaranteed, the mean number of attempts in the first level was 8.4), may be especially thwarting when the character failing is *you*. This would suggest that, for instance, failing as an abstract shape, but succeeding as a likeness to yourself, would be an effective adaptive avatar representation for learning.

Players using likeness avatars often made personal comparisons, e.g., "My avatar has a likeness to myself [...] she is chubby like me.", "[...] I had black hair and a gray shirt and my red glasses", and one player commented "[...] the avatar's success is my own", seeming to support the above. But some players felt they were unable to adequately represent themselves in the Mii; one participant said "it was difficult to make the avatar look like me" and "there weren't enough colors to customize the shirts." This means that despite the large number of options for hairstyles (72), eyes (48), mouths (24), etc. some players still found the avatar creator to be limiting. Even though this is the case, we found the avatar creator to be more than sufficient for most players. Figure 6 suggests that stereotype threat may have been an additional contributing factor for some players; there were greater disparities between the two avatar types in African American players, i.e., the likeness avatar may have acted as a stereotype threat trigger, as consistent with previous work [4].

Because the actual customization of the likeness avatar was part of the condition, perhaps players simply did not enjoy that aspect of the game. Or it is possible they did enjoy it, but were unsatisfied with the avatar's role in the game. If this was the case, it affected not only their performance, but also significantly affected their disposition towards the game in a negative manner, *despite* the fact that player avatar ratings are strongly in favor of the likeness avatar. It is clear that more work needs to be done in distinguishing the specific psychological effects at play here. However, the results suggest that there

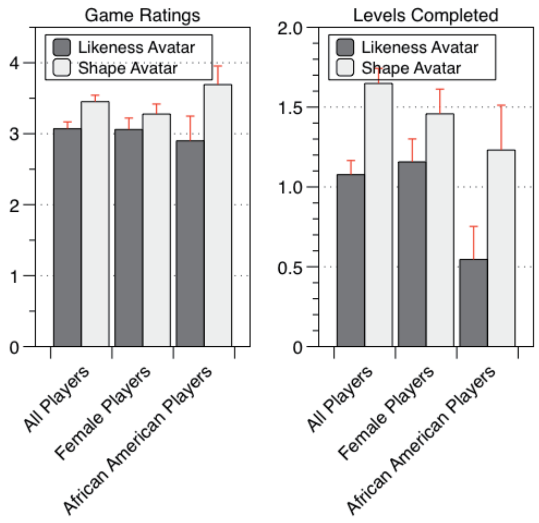


Fig. 6: Experiment 2 game ratings and level completion averages between avatar types across social categories. Here, the focus is on two social groups underrepresented in STEM.

are differences between avatars customized in the *likeness of players* and avatars selected as *geometric shapes*. Were we to make a recommendation to educational game makers based on these results alone, we would be hard-pressed to make a definite statement; however, if faced between simpler, abstract avatars and more complex, customizable avatars, we would be in support of simpler avatars.

The studies we have done here to collect data on comparing avatar types has a broader aim. Virtual identities are now ubiquitous; a systematic understanding of their impacts on user performance, engagement, and learning is crucial. These results help provide the basis of a follow-up project to develop personalization algorithms for adaptive learning systems that dynamically adapt the virtual identities of students to support performance, engagement, and learning within a broader learning ecology.

B. Implications for Player Modeling

Player modeling has tended to focus on capturing 1) gameplay data (i.e., behavioral data), 2) objective data as bodily feedback (ie. physiological responses), and 3) the game context (i.e. actual game events) [7]. One of the challenges in player modeling has been simply a lack of rich data [39]. The contribution in this work is a comparative study of two avatar types in a STEM learning game. We propose that aspects of both players' sociocultural and virtual identities have substantial enough of an impact to improve upon existing player models. For instance, research from the social sciences has made clear the effect of identity salience, in the domains of "math, verbal, analytical, and IQ performance, golf putting, reaction time performance, ..." [8]. These findings may well analogize to performance in learning systems via virtual identities.

Natural language processing revealed considerable differences in the way players describe likeness avatars versus geometric shape avatars. Players using shape avatars tended to use lots of articles, impersonal pronouns, and describe their avatar in the *past tense* (e.g., "It was a pentagon. It wasn't notably memorable, but it was my choice. I like options."). Whereas players using likeness avatars spoke in the 1st and 3rd person *present tense* (e.g., "He is short. He wears a red shirt. He has dark-skin and spiky hair as well."). This suggests that the differences in relationship between the player and the avatar is significant. Banks et. al define one axis of the player-avatar relationship (PAR) as being "identification" (from "my avatar is a digital form" to "my avatar is *me* in digital form") [40]. Harrell et. al proposes one avatar dimension as being between "character external to self" to "mirror of real self" [41]. This player identification with avatar (or lack thereof) may well play a role in player behavior; more work is needed to fully characterize the consequences of these constructs.

Predicting level completion in our STEM learning game was challenging. This is likely the result of limited data, a small number of attributes to infer behavior from, and inherent noise in a very brief exposure. Nonetheless, using decision tree learning we were able to predict level completion 11.9% over baseline. Single-attribute evaluation revealed that avatar type was the individual attribute with most merit. This is evidence that integrating avatar type as an attribute, particularly when avatars can vary in a considerable manner, can prove to be a beneficial behavioral predictor. Modeling even more fine-grained aspects, such as color [27], [28], type [3], [4], degree of anthropomorphism, similarity to self, similarity to other, customizability, in addition to social identity, may produce even more robust models of player behavior.

C. Limitations

We chose to run this experiment in a setting that we had created (*Mazzy*). This was to ensure that we could control exactly the type, size, position, etc. of the avatars in the experiment. This also gave us the ability to ensure the levels were suitably difficult for our sample population (early pilot studies aimed to balance the game such that on average participants could finish about half the levels). However, as with other avatar studies of this type, these results take place in a single game, and across two avatar types. Should we change the setting (e.g., an interactive narrative game) or the art style (e.g., blocky minecraft avatar), we may find that the results are dependent on these factors. Therefore, while it is valuable to conduct these studies and to disseminate results, it is also vital to perform replication studies. This is an interesting, if challenging, area to conduct research in.

VII. Conclusion

In this paper, we have shown results that suggest players using likeness avatars exhibit different behavior as compared to players using shape avatars. In particular, players selecting and using geometric shape avatars had significantly higher performance and higher task affect than players customizing and using likeness avatars. However, players rated the likeness avatar significantly higher. We also performed natural language processing and sentiment analysis; players provided text descriptions of their avatars, and we found significant heterogeneity across linguistic dimensions. In particular, likeness avatars scored higher on those dimensions that relate to player-avatar identification. Lastly, we used supervised learning algorithms to predict the number of levels that players complete. We were able to achieve modest gains above baseline using decision tree learning. Single-attribute evaluation using the 1R algorithm revealed that avatar type alone could achieve prediction 7% above baseline.

We have reported on results of a comparison between avatars customized in the likeness of the player and avatars selected as a geometric shape in a STEM learning game. These avatars are blended identities (a selective projection onto a virtual avatar) that can impact performance and engagement. We suggest more investigation into the psychosociological mechanisms by which avatars impact player performance. Finally, we propose a more robust model that incorporates the virtual and social.

VIII. Future Work

Mazzy is currently being developed to be a longer game (12 levels, new mechanics, more progression, etc.) and to track additional behavioral metrics (mouse biometrics, keyboard biometrics, etc.). These types of studies are being done in the interest of creating a new class of personalization algorithm in adaptive learning systems that will take into account the social identities of learners. We envision virtual representations that are dynamic and may adapt over time, perhaps appearing abstract in some conditions and reflecting users' social identities in others, not only in terms of appearance, but also in terms of behavior, visual style, reflecting user's interests, and other features strongly associated with their cultures.

Acknowledgment

We thank the anonymous reviewers for their feedback. This work is supported by National Science Foundation (NSF) Award IIS-1064495 and a Natural Sciences and Engineering Research Council of Canada (NSERC) fellowship.

References

[1] M. Shih, T. Pittinsky, and A. Trahan, "Domain-specific effects of stereotypes on performance," *Self and Identity*, no. March, 2006.

[2] C. Steele and J. Aronson, "Stereotype Threat and the Intellectual Test Performance of African Americans," *Journal of personality and social psychology*, 1995.

[3] D. Kao and D. F. Harrell, "Exploring construction, play, use of virtual identities in STEM learning," *Jean Piaget Society Annual Conference*, 2015.

[4] ——, "Toward Evaluating the Impacts of Virtual Identities on STEM Learning," *Foundations of Digital Games*, 2015.

[5] G. N. Yannakakis and M. Maragoudakis, "Player Modeling Impact on Player's Entertainment in Computer Games," *Lecture notes in CS*, 2005.

[6] J. Togelius, R. De Nardi, and S. M. Lucas, "Towards automatic personalised content creation for racing games," in *Proceedings of IEEE Conference on Computational Intelligence in Games*, 2007.

[7] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. André, "Player Modeling," *Dagstuhl Follow-Ups*, vol. 6, p. 59, 2013.

[8] C. M. Steele, "Whistling Vivaldi and other clues to how stereotypes affect us," in *Whistling Vivaldi*, 2010.

[9] N. Ambady, M. Shih, A. Kim, and T. L. Pittinsky, "Stereotype susceptibility in children: effects of identity activation on quantitative performance." *Psychological Science*, 2001.

[10] J. F. Morie and G. Verhulsdonck, "Body/Persona/Action! Emerging Non-anthropomorphic Communication and Interaction in Virtual Worlds," *Advances on Computer Entertainment Technology*, p. 365, 2008.

[11] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of Experimental Social Psychology*, vol. 52, pp. 113–117, 2014.

[12] S. Papert and I. Harel, "Situating Constructionism," 1991.

[13] J. Piaget, *Piaget and His School*, B. Inhelder, H. H. Chipman, and C. Zwingmann, Eds. Springer Berlin Heidelberg, 1976.

[14] S. Papert, "Mindstorms," 1993.

[15] D. F. Harrell, *Phantasmal Media: An Approach to Imagination, Computation, and Expression*. The MIT Press, 2013.

[16] ——, "Toward a Theory of Critical Computing: The Case of Social Identity Representation in Digital Media Applications," *CTheory*, 2010.

[17] J. P. Gee, *What Video Games Have to Teach Us About Learning and Literacy*. Palgrave Macmillan, 2007.

[18] D. F. Harrell, D. Kao, C. Lim, J. Lipshin, and A. Sutherland, "The Chimeria Platform: User Empowerment through Expressing Social Group Membership Phenomena," *Digital Humanities*, 2014.

[19] C.-U. Lim and D. F. Harrell, "Modeling Player Preferences in Avatar Customization Using Social Network Data," in *CIG*, 2013.

[20] C. Lim and D. Harrell, "Developing Social Identity Models of Players from Game Telemetry Data," *AIIDE*, 2014.

[21] N. Yee and J. Bailenson, "The Proteus Effect: The Effect of Transformed Self-Representation on Behavior," *Human Comm. Research*, Jul. 2007.

[22] M. Shih, T. L. Pittinsky, and N. Ambady, "Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance," 1999.

[23] C. Good, A. Rattan, and C. S. Dweck, "Why do women opt out? Sense of belonging and women's representation in mathematics." *Journal of personality and social psychology*, Apr. 2012.

[24] L. J. Stricker and W. C. Ward, "Stereotype Threat, inquiring About Test Takers Ethnicity and Gender, and Standardized Test Performance," *Journal of Applied Social Psychology*, vol. 34, no. 4, pp. 665–693, 2004.

[25] J. Aronson, C. B. Fried, and C. Good, "Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence," *Journal of Experimental Social Psychology*, 2002.

[26] N. Yee, N. Ducheneaut, M. Yao, and L. Nelson, "Do Men Heal More When in Drag?" *CHI 2011*, pp. 1–4, 2011.

[27] J. M. Linebarger and G. D. Kessler, "The effect of avatar connectedness on task performance," *Lehigh Univ TR*, 2002.

[28] I. X. Dominguez and D. L. Roberts, "Asymmetric Virtual Environments : Exploring the Effects of Avatar Colors on Performance," *Experimental Artificial Intelligence in Games: Papers from the AIIDE Workshop*, 2015.

[29] Mii Search - Create Nintendo Mii Characters, "Mii Creator," 2010. [Online]. Available: http://www.miisearch.com/mii-creator.html

[30] Math is Fun - Maths Resources, "Shapes," 2014. [Online]. Available: http://www.mathsisfun.com/shape.html

[31] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language. use: our words, our selves." *Annual review of psychology*, vol. 54, pp. 547–577, 2003.

[32] J. Pennebaker and C. Chung, "The Development and Psychometric Properties of LIWC2007," *Austin, TX, LIWC. . . .*, pp. 1–22, 2007.

[33] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of pers. and social psych.*, 1999.

[34] M. Hall *et al.*, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, 2009.

[35] T. Mahlmann, A. Drachen, J. Togelius, A. Canossa, and G. N. Yannakakis, "Predicting player behavior in Tomb Raider: Underworld," *Computational Intelligence in Games*, 2010.

[36] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," 2008.

[37] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, 1993.

[38] N. D. Bowman, R. Rogers, and B. I. Sherrick, "In Control or In Their Shoes? Video games, characters, and enjoyable or meaningful experiences," *Broadcast Education Association Research Symposium "Media and Social Life: The Self, Relationships, and Society."*, 2013.

[39] S. Lucas, M. Mateas, M. Preuss, P. Spronck, and J. Togelius, "Artificial and Computational Intelligence in Games," 2013.

[40] J. Banks and N. D. Bowman, "Close intimate playthings? Understanding player-avatar relationships as a function of attachment, agency, and intimacy," *Selected Papers of Internet Research*, pp. 1–4, 2013.

[41] D. F. Harrell and S. V. Harrell, "Imagination, Computation, and Self-Expression: Situated Character and Avatar Mediated Identity," *Leonardo electronic almanac*, vol. 17, no. 2, pp. 74–91, Jan. 2012.