Published in final edited form as:

Comput Human Behav. 2014 May 1; 34: 1–11. doi:10.1016/j.chb.2014.01.044.

Quantifying Engagement: Measuring Player Involvement in Human-Avatar Interactions

Anne E. Norris^a, Harry Weger^b, Cory Bullinger^c, and Alyssa Bowers^b

^aCollege of Nursing, University of Central Florida, 12201 Research Parkway, Orlando, FL 32826, USA

^bNicholson School of Communication, University of Central Florida, 12405 Aquarius Agora Dr., Orlando, FL 32816-1344

^cDepartment of English, College of Arts & Humanities, University of Central Florida, P.O. Box 161346, Orlando, FL 32816-1346

Abstract

This research investigated the merits of using an established system for rating behavioral cues of involvement in human dyadic interactions (i.e., face-to-face conversation) to measure involvement in human-avatar interactions. Gameplay audio-video and self-report data from a Feasibility Trial and Free Choice study of an effective peer resistance skill building simulation game (DRAMA-RAMA™) were used to evaluate reliability and validity of the rating system when applied to human-avatar interactions. The Free Choice study used a revised game prototype that was altered to be more engaging. Both studies involved girls enrolled in a public middle school in Central Florida that served a predominately Hispanic (greater than 80%), low-income student population. Audio-video data were coded by two raters, trained in the rating system. Self-report data were generated using measures of perceived realism, predictability and flow administered immediately after game play. Hypotheses for reliability and validity were supported: Reliability values mirrored those found in the human dyadic interaction literature. Validity was supported by factor analysis, significantly higher levels of involvement in Free Choice as compared to Feasibility Trial players, and correlations between involvement dimension sub scores and self-report measures. Results have implications for the science of both skill-training intervention research and game design.

Keywords

engagement; involvement; human-avatar interactions; simulation

© 2014 Elsevier Ltd. All rights reserved.

Address of Correspondence: Anne E. Norris, College of Nursing, University of Central Florida, 12201 Research Parkway, Rm 475, Orlando, FL 32826-3265, 1-407-823-4185 (Office), 1-407-823-5675 (College Fax), anne.norris@ucf.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

¹The need for increased detail and coloration to increase engagement was identified by E2i Creative Studios, at the Institute for Simulation and Training, University of Central Florida.

1. Introduction

More and more computer-based simulation games are being used for teaching and training purposes in healthcare (Nehring & Lashley, 2009), school (Gee, 2005), and military (Garris, Ahlers, & Driskell, 2002; Oswalt, 1993; Smith, 2010) settings. “Digital puppetry,” or the real-time animation of human-controlled virtual characters, or avatars, is a technology that can be used to create a highly interactive simulation game play experience (Norris, Hughes, Hecht, Peragallo, & Nickerson, 2013). Highly interactive games are assumed to be highly engaging (Charoenying, 2010) and thus more effective for teaching and training (Dickey, 2005; Salen & Zimmerman, 2004; Vogel et al., 2006). However, no measurement approach exists to assess the engagement properties of these games.

The purpose of this methodologic paper is to assess the reliability and validity of using Guerrero’s (2005) rating system for assessing involvement in human dyadic interpersonal interactions as a measure of player involvement in the human–avatar interactions occurring in a skill-building simulation game called DRAMA-RAMA™ (Norris et al., 2013). This rating system provides an opportunity to add an objective measure to the battery of current engagement instruments that are reliant on self-report data. Towards that end, audio-video data from an early and subsequent version of the game were analyzed and used to investigate the reliability and validity of this measurement approach for capturing the involvement aspect of engagement.

2. Background

Narrative engagement theory (or NET; Miller-Day & Hecht, 2013), data regarding real-time simulation games, and game theory alternately identify aspects of effective skill-building interventions and features of successful games that can inform measurement of player engagement. NET identifies perceived realism and involvement as key aspects of engagement, and indicative of a game’s ability to teach the player new behavioral skills.

In the context of real-time simulation games, engagement is indicated by the player’s involvement in game play and the player’s perception that the game play experience is real (i.e., the game play “rings true” to his or her own experiences; Lee, Hecht, Miller-Day, & Elek, 2011). Player involvement increases with the perceived behavioral realism of the avatar (Guadagno, Swinth, & Blascovich, 2011; von der Pütten, Krämer, Gratch, & Kang, 2010). The avatar’s capacity to act in a realistic manner contributes to what McMahan (2003) refers to as the social realism of the virtual environment. Social realism, in combination with perceptual realism—that is, how convincing the game environment looks and sounds—helps build and sustain an overall sense of realism for the player.

Game theory identifies two characteristics as critical components of highly engaging games (Sweetser & Wyeth, 2005): (1) the absence of predictability (Koster, 2005), or the game’s ability to generate surprise or mystery (Dickey, 2005; Garris et al., 2002; Salen & Zimmerman, 2004), and (2) the presence of flow, or the game’s ability to create an experience where the player loses track of time (Csikszentmihalyi, 1990). Therefore, games that are perceived as predictable (Koster, 2005) or games that fail to create an optimal flow

experience (Csikszentmihalyi, 1990) are not engaging because they are not fun to play (Sweetser & Wyeth, 2005). Moreover, the literature suggests that engaging games are more apt to facilitate learning (Moneta & Csikszentmihalyi, 1996; Shernoff, Csikszentmihalyi, Schneider, & Shernoff, 2003).

Currently, several measures of player engagement have been developed for various game types including single-player, entertainment games (Brockmeyer et al., 2009; IJsselstein, Poels, & de Kort, 2008; Mayes & Cotton, 2001), educational games (Fu, Su, & Yu, 2009), and virtual simulations (Witmer and Singer, 1998). Variations in existing measures reflect both the wide range of video game genres and goals, and the overall lack of definitional consensus for key theoretical concepts, such as engagement, flow, and immersion (Brockmeyer et al., 2009; Procci & Bowers, 2011). Despite this, one notable similarity is the use of a questionnaire format for gathering player self-report data. Self-report data are undoubtedly essential for gauging players' subjective experiences of game play but may not fully quantify all aspects of player engagement. Objective measures of observable verbal and nonverbal behavior offer additional and complementary information regarding engagement, and they have the potential to provide a more thorough or richer assessment than subjective measures alone.

2.1. Conceptual framework

We propose a conceptual model for measuring player engagement in real-time simulation games that includes predictability, flow, involvement, and perceived realism (see Figure 1). Predictability, perceived realism, and flow are properties of a game that are reflected in subjective experiences of game play. They are not directly observable. In contrast, involvement is directly observable because as defined in the Oxford English Dictionary, it is the "fact or condition of participating in something" (Involvement). Involvement does not lend itself to self-report because it is not a subjective experience. Hence, it is best assessed with coded or rated observations of game play behavior.

2.2.1. Measurement of involvement—Observational rating systems of nonverbal involvement emerged from researchers interested in interpersonal interaction (e.g., Coker & Burgoon, 1987; Guerrero, 1994; 2005). This system for rating behavioral cues of involvement in human dyadic interactions (i.e., face-to-face conversation) is well established. Although Guerrero's (2005) rating system was developed for human dyadic interactions, we believe that the human–avatar interactions facilitated by DRAMA-RAMA™ can be similarly rated for player involvement. The participants in our study are engaging in dyadic interactions with avatars, albeit via a mediated channel, and the interactions in the game are designed to mimic face-to-face interaction. Our goal is to determine how well this system of observational behavior rating can be used to evaluate involvement behaviors in such interactions.

The system is comprised of six scales: Immediacy, Expressiveness, Altercentrism, Interaction Management, Composure, and Positive Affect. Each scale is designed to capture a dimension (or set) of behaviors relevant to involvement and are defined as follows: (1) *Immediacy* dimension behaviors measure the physical proximity between two individuals;

(2) *Expressiveness* dimension behaviors communicate the level of energy, activity, and enthusiasm toward the conversation partner; (3) *Altercentrism* dimension behaviors reflect the degree of focus on the conversation partner during the interaction; (4) *Interaction Management* dimension behaviors support a smooth flow of conversation; (5) *Composure* dimension behaviors reflect an absence of nervous body movements or the presence of confidence; and (6) the *Positive Affect* dimension includes smiling, laughing, and other behaviors that reflect good feelings about the interaction and partner.

The Guerrero (2005) system is uniquely suited for measuring player involvement in a live simulation game involving digital puppetry for three reasons. First, Guerrero's system was designed to determine the degree to which an individual is actively involved in a real-time conversation, and our interest is in using a system to capture player involvement in real-time interactions with animated avatars.

Second, this system permits an observer to rate rather than code specific behaviors. This allows the capture of multiple channels, functions, and types of behavior indicative of involvement. In contrast, coding only focuses on one behavior channel at a time and often necessitates overly simplistic or overly complicated categories to achieve satisfactory reliability. Rating systems also use continuous rating scales (semantic differential or low-to-high behavioral frequency), which require less time to apply than the individual, discreet categories required by coding systems. This time savings is important given our desire to assess the number and variety of involvement behaviors that may potentially occur in a human–avatar interaction.

Third, although the system involves six different dimensions, the use of particular dimensions can vary according to the focus of the research without impacting measurement reliability and validity. This is advantageous in the present study where it is clear *a priori* that Immediacy dimension items relating to physical contact or proximity are not applicable because the nature of the game technology used in our research does not involve or simulate physical contact or physical co-presence.

2.3. Real-time simulation game: DRAMA-RAMA™

DRAMA-RAMA™ uses synchronous human–avatar interactions for the purpose of interpersonal skill training, specifically resistance to peer pressure. The game is controller-less; thus, no input device is necessary. Instead, players communicate verbally and nonverbally in real time with avatars displayed on the game monitor and earn points for resisting avatar pressure to engage in risky behavior. Additional information on the game can be found at mightygirls.ist.ucf.edu.

The avatars are three-dimensional characters designed to resemble typical early adolescent, or middle school age, girls and boys. All in-game avatars are controlled remotely by a single inter-actor, or an adult actor trained in the use of technology and interactive performance (Wirth, 1994). Inter-actor speech, body motions, and facial expressions are transmitted in real time to the avatars via an Internet connection using a combination of Skype™, motion capture, and artificial intelligence (see Figure 2). Players are led to believe that the avatars

are computer controlled. They are not made aware that the avatars are controlled by a human inter-actor.

Although the game follows a scripted story, the game's technology enables the inter-actor to improvise and realistically tailor avatar responses to fit each player's unique verbal and nonverbal responses instead of relying on pre-programmed dialogue. For example, in the opening game scene, the inter-actor introduces players to this game style by pointedly commenting on player clothing or hairstyle such as "That's such a cute shirt; the blue really brings out your eyes!" and "You look so cute when you wear your hair down like that, [player name]!" or by calling attention to player behavior such as "What's going on over there?" if a player is looking away from the screen. Thus, the inter-actor's responses vary depending on what the player says or how the player acts, making the avatar verbal and non-verbal behavior fit (or be more appropriate to) what the player is saying or doing.

Moreover, rather than constrain players with pre-programmed responses from which to select, players instead *speak* their own unique replies throughout the game and inter-actors customize the dialogue accordingly. This capacity for immediate and dynamic adaptation facilitates realistic synchronous communication between player and avatar, which renders the interactions more similar to face-to-face conversations than currently available simulation training applications that utilize pre-scripted dialogue and limit player response options to selecting from a list of possible text responses; cf. the 'Dating Game' component of *It's Your Game* (Markham, Peskin, Shegog, & Tortolero, 2014) or *The Selling Experience* (Experience Builders LLC, 2006). Thus, the lack of such canned replies in DRAMA-RAMA™ effectively personalizes each player's game play experience, which is notable for two reasons. First, it helps increase the avatar's social presence and thus their behavioral realism (Guadagno, Blascovich, Bailenson, & McCall, 2007). Second, this capacity for dialogue adaptivity increases the game's challenge and thus its appeal (Lopes & Bidarra, 2011).

Each game play session contains three scored story scenes, followed by a fourth un-scored debriefing scene that provides feedback to the player via a disembodied voice-over (also provided by the inter-actor using a "narrator voice;" no avatar speaks directly to the player) and an onscreen display of points earned during the game. The communication is one-way in this fourth scene. All avatars from the game are present but none are directly speaking to the player. Thus this final scene lacks the interactivity and player agency of the previous game scenes and renders the debriefing more similar to interactions with computer-controlled characters (also referred to as embodied conversational agents, or ECAs). Compared to avatars, ECAs are unable to provide the meaningful social interaction that positively stimulates and sustains player involvement (Guadagno et al., 2007; Lim & Reeves, 2009). Therefore, involvement scores are expected to be lower for this fourth scene.

2.3.1. Nature of human-avatar interactions in DRAMA-RAMA™—Human-avatar interactions facilitated in the game were driven by storyline and demarcated according to three scored story scenes, each of which contained relatively short, approximately two-to-five minute-long, interactions. (As already noted, interactions did not occur in the fourth, un-scored scene.) Scenes were consistent in the contained plot elements. However scene length

varied due to player style and personality—self-confident, talkative players produced longer interactions, whereas shy, taciturn players' interactions were shorter.

Each interaction consisted of players communicating verbally and nonverbally with the on-screen, inter-actor-controlled avatars and the avatars responding to the players' body language and verbal replies. Inter-actors were also able to tailor the avatars' nonverbal behavior, and although these were limited to facial expressions and head and hand gestures, previous research suggests that their inclusion will increase the avatars' behavioral realism and thus encourage player engagement (Chen et al., 2012). During key points in each scene (i.e., when players were being pressured to agree to risky behavior), players were awarded or deducted points based on their use of verbal and nonverbal resistance strategies. This assessment was visualized in real time during the interaction via an on-screen scoring mechanism.

2.3.2. Game prototypes—Two DRAMA-RAMA™ prototypes were played and contributed data to the analyses described in this paper: an early prototype used in a Feasibility Trial (see Figure 3) and a revised prototype used in a Free Choice study (see Figure 4). Game play was essentially the same for both prototypes (i.e., consisted of three scored scenes and a fourth un-scored scene that provided feedback on gameplay).

Both prototypes had fully colored avatars, but differed in features related to engagement. As can be seen in Figure 3, the environments in the early prototype lacked high-level detail or textures and contained little coloration (they were predominately grey). The early prototype game story also had limited branching: Regardless of how they responded to the offer, all players ended up at a house party with no parental supervision and freely available alcohol. The only story variation was whether the player had a final conversation with an older boy or the female best friend character at the end of the game. In contrast, the revised prototype contained highly detailed, fully colored environments¹ and an alternate story branch that allowed participants to go to either the mall or the party, thus providing new scenes and story lines (see also Figure 4). This branching was designed to maintain engagement over repeated playthroughs.

The two prototypes also differed in their potential for technical interference. The early prototype used a non-directional microphone, and early testing of game equipment indicated that a hissing or static sound occurred at times during game play. Improvements in the revised prototype equipment eliminated these sounds, which should also contribute to increased perceptual realism and thus engagement.

3. Study overview

This study focused on the validity and reliability of Guerrero's (2005) rating system as a measure of involvement in human–avatar interactions. We use theorized relationships in our conceptual model to assess construct validity. In addition, we assessed criterion validity of this measure by comparing: (1) involvement scores for the early and revised game prototypes and (2) involvement scores for the final game scene.

Specific hypotheses to be tested are:

1. Reliability will be supported at the rater- and dimension-scale levels.
2. Construct validity will be supported by correlations with self-report measures of involvement, specifically:
 - a. Positive correlations between involvement, and perceived realism and flow.
 - b. Negative correlations between involvement and predictability of game play.
3. Criterion validity will be supported by:
 - a. Lower involvement dimension scores for an early game prototype as compared to a revised prototype designed to be more engaging.
 - b. Lower involvement dimension scores for final game scene in which the participant does not interact with a specific avatar as compared to earlier game scenes.

4. Methods

4.1. Site, sample, and design

Audio-video and self-report data used in the analyses described here were collected as part of two studies, the Feasibility Trial (early prototype) and the Free Choice study (later prototype), both of which involved girls enrolled in a public middle school in Central Florida that served a predominately Hispanic (greater than 80%), low-income student population. The Feasibility Trial data included here are only for those participants ($n = 18$) who played that game as part of a randomized control trial that occurred in Spring 2011. The Free Choice data included here are for participants ($n = 13$) who played the later prototype in Spring 2012. Neither group of participants was offered incentives for playing the game or for completing study measures.

Participants in the Feasibility Trial study were slightly younger (Mean age = 11.67, $SD = .91$, range = 11–14) than Free Choice study participants (Mean age = 12.26, $SD = .65$, range = 11–14; $t = -2.35$, $df = 38$, $p < .05$). All participants in the Feasibility Trial were Hispanic/Latino by study design. In contrast, Free Choice study was open to all girls at the study site school, resulting in a sample that was 83% Hispanic/Latino, 13% African American, and 4% other. However, the two study samples did not differ significantly with respect to ethnicity (Fisher's exact, $p = .10$).

4.2. Study procedure

The procedures for both studies were similar with three exceptions. First, Feasibility Trial participants were offered a curriculum teaching them peer resistance skill content prior to game play (see Norris et al., 2013), whereas Free Choice study participants were only given a flyer containing information about these skills.

Second, Feasibility Trial participants were assigned to play either the original prototype ($n=18$) or Dancing with the Stars™ ($n=20$; data for these participants not included in study

analyses presented here; see Norris, et al, 2013 for analyses comparing Wii and DRAMA-RAMA game play experiences). In contrast, Free Choice study participants ($n = 27$ in total) could choose to play the revised prototype, the Wii™ Dancing with the Stars™ game, or both. The majority of Free Choice participants played both games (20 or 74%). Few played only one: 15% (4) only played the Wii game; 11% (3) only played the revised prototype. Girls who played both games expressed stronger positive affect regarding DRAMA-RAMA than the Wii ($p < .01$).

Third, unlike the Free Choice study that conducted game sessions in a private room with an adult down the hall in a waiting room, the game sessions of the Feasibility Trial took place in a temporary “cubicle” within a large classroom that visually obscured the player from an adult who remained in the room. Players in both studies were directed to ask this adult for assistance if needed, particularly for technical issues.

Recruitment for both studies began after university institutional review board approval. School and study staff distributed informational packets to girls to take home to parents or legal guardians. The packets contained a letter of support from afterschool program administrators, consent forms, and information about how to contact the principal investigator (PI) or a Spanish-speaking research staff member should they have any questions. Prior to beginning any research activities, verbal assent was obtained from girls who brought back signed parental consent forms and expressed interest in participating.

Game play occurred in a small, private room at the school and lasted approximately 15 minutes. Players sat at a table that held a 42-inch television monitor with a small Web camera affixed to the top to record their verbal and nonverbal behavior. Following the game, participants completed a paper Game Experience Questionnaire in a different room. This questionnaire took about 5 minutes to complete.

4.3. Self-report measures

All participants completed measures of demographic characteristics and a Game Experience Questionnaire containing self-report measures of perceived realism, flow, predictability, and technical interference. Feasibility Trial participants completed detailed demographic items as part of a baseline attitudes, intention, and behavior survey (see Norris et al., 2013). Free Choice study participants completed a short set of demographic items attached to the beginning of the Game Experience Questionnaire. Only demographic items and self-report game experience items relevant to the analyses presented here are described next.

4.3.1. Demographic items—Participants were asked to provide their current age in years and their ethnicity. They could check off as many ethnic groups as applied using a list and also write in an ethnic group. Groups listed were: Hispanic/Latino, African American, and Haitian.

4.3.2. Perceived realism, predictability, and flow—Single-item measures in the Game Experience Questionnaire were used to assess: perceived realism (“Talking with the game kids felt real.”), predictability (“I could tell what was going to happen next in this game.”), and flow (“Time went by quickly when I was playing this game.”). Response

options for these items were: (1) strongly disagree, (2) disagree, (3) agree, and (4) strongly agree. These response options were collapsed into dichotomous agree or disagree categories for data analysis because of discontinuous, skewed, multimodal distributions preventing the use of *t* test, Mann Whitney, or chi-square analyses involving 2×4 tables. Test–retest reliability values were calculated using the Feasibility Trial data only (players in the Feasibility Trial played the game twice) and are as follows: perceived realism ($r = .68$), flow ($r = .54$), and predictability ($r = .38$). The lower value for predictability is expected because predictability should increase (i.e., not be stable) with players' subsequent game play sessions.

4.4. Involvement rating

All audio-video game play data were rated using measurement scales adapted from the Guerrero (2005) rating scheme for studying involvement during human dyadic interactions. We chose items that represent the major sources of behavioral cues correlated with involvement (see Coker & Burgoon, 1987; Guerrero, 2005). These included: (1) vocalics, or vocal qualities such as variation, expressiveness, warmth, etc.; (2) kinesics, or body movement such as posture, relaxation, gestures, etc.; (3) fluency, or the smooth flow of interaction such as talk overs and interruptions; and (4) facial expressions, which include gaze direction and intensity, positive/negative affect, and interest.

4.4.1. Raters and rating process—Two graduate students were trained to recognize and rate nonverbal and verbal involvement behaviors using seven-point interval scales, ranging from “never exhibited” to “constant.” Training included definitions and examples of relevant behavior, reviewing and practicing rating techniques on sample videos, and assessing preliminary reliabilities for rating consistency or areas in need of further training. Rater training continued until satisfactory inter-rater reliability was achieved ($r_1 > .60$).

Involvement ratings were obtained for each player for the four separate game scenes. Scenes ranged from two-to-five minutes in length, depending on the player. To improve accuracy and limit mental fatigue (La France, Heisel, & Beatty, 2007), each of these four scenes was rated twice—once for verbal cues and a second time for nonverbal behavioral cues. When rating the verbal cues, raters placed a sticky note over participants' faces to limit visual distractions. When rating the nonverbal cues, videos were played at double-speed, with all sound muted and any necessary adjustments made to optimize participant visibility. Raters took breaks at least every twenty minutes and limited rating sessions to no more than two hours, as recommended by La France et al. (2007).

Both raters independently rated every player for which adequate audio and visual data were available. Three issues with recording quality or circumstances created missing rating data. First, some recordings contained unintelligible or missing player and/or avatar audio, resulting in no ratable verbal behavior (early prototype, $n = 0$; revised prototype, $n = 4$). Second, some contained poor video quality or did not provide a full view of the player either because of a player's short stature relative to the webcam angle or because the player leaned too far forward on the game station table, situating herself partially or completely out of view of the webcam, resulting in no ratable nonverbal behavior (early prototype, $n = 1$;

revised prototype, $n = 9$). Third, some were missing audio and video for the feedback scene, resulting in no ratable data for the final game scene (early prototype, $n = 2$; revised prototype, $n = 1$). Consequently our sample size for analysis was reduced to 16–18 for the early prototype and 12–13 for the revised prototype.

4.4.2. Preparation of involvement rating data for further analysis—Prior to tests of study hypotheses, rating data from both studies were: (a) examined for signs of systematic patterns consistent with error or missing data; and (b) submitted to factor analysis. The results of each of these preparation steps are discussed next.

Three different systematic patterns were observed in these data. First, little or no incidence of head nods and back channels (i.e., sounds made to encourage or affirm another speaker's statements) were observed, thus preventing these rating items from being used in further analysis.

Second, forward lean and body orientation item ratings were negatively rather than positively correlated, which raises questions about the validity of these items which are conceptualized as being positively related. Further inspection of the video data and physical setup of game play revealed that as participants leaned in towards the avatar during game play (consistent with increased immediacy), they put their hands/arms on the table in front of them onto which the screen displaying the avatars was placed. This resulted in a closed body position, which was defined in the rating system as indicating less immediacy. Similarly, when they leaned back (consistent with decreased immediacy), their arms went to their sides, resulting in an open body position, which is defined as indicating greater immediacy. Hence, these immediacy items were eliminated from further analysis.

Finally, poor audio quality in the final game scene created missing data regarding behaviors correlated with vocalics, such as response latencies, vocal pleasantness, etc. This had three consequences. First, we left the final scene out of our factor analyses and composite measures of involvement so that all variables contributed equally to the composite scores. Second, we did not use ratings from the final, un-scored scene when computing inter-rater reliability. Finally, in our analyses comparing involvement scores for the final scene with involvement scores for earlier scenes, we computed the Positive Affect variable without the two vocalic measures to ensure the means across scenes represented the same variable combinations.

After eliminating the aforementioned problematic items, the remaining rating data were factor analyzed using best practices as recommended by Osborne, Costello, and Kellow (2008). These practices included use of the maximum likelihood method for extracting factors and a varimax method of rotation. The selection of number of factors was based on Eigen values greater than one and interpretation of the scree plot. Items were retained only if they loaded above .50 and did not have a secondary loading above .30.

In all, we found six interpretable factors in this analysis: Expressiveness, Altercentrism, Interaction Management, Positive Affect, Composure, and Vocal Relaxation. The first four of these factors lined up with the dimensions of involvement identified for human dyadic

interactions (Coker & Burgoon, 1987; see Table 2). The Immediacy dimension factor did not emerge in this analysis, but items specific to this dimension were omitted from the analysis because of missing data (head nods, back channeling) or conceptual inconsistency between the meaning of behavioral cues in Guerrero's (2005) system and the meaning of these cues under our game play conditions (forward lean, body orientation).

Two of the factors that emerged (Composure, Vocal Relaxation) were comprised of items that previously loaded together on a single factor in Coker and Burgoon's (1987) work involving ratings of human dyadic interactions. Hence, we examined whether extracting five, rather than six, factors would result in these items loading together, but they continued to load on different factors. Table 1 contains the factor loadings for the best fitting solution. Factor means, standard deviations, and intercorrelations for the six-factor solution are provided in Table 2.

5. Results

5.1. Reliability of rater- and dimension-scale levels

As can be seen from the data in Table 1, item- and scale-level reliability statistics generally supported our hypotheses stating that reliability would be supported at the rater- and dimension-scale levels. The intra-class correlation analysis measure of inter-rater agreement indicated that the raters were relatively consistent at the item level across the involvement dimensions ($r_1 = .65-.95$), with the exception of the Vocal Relaxation items ($r_1 = .42-.49$). Cronbach's alpha values exceeded .70 for all scales except Vocal Relaxation, which had a Cronbach's alpha value of .60, indicating acceptable to strong scale level reliability.

5.2. Construct validity of self-report measures of involvement

The direction and magnitude of the correlations between ratings of involvement and participants' self-reported ratings of perceived realism, predictability, and flow used to evaluate construct validity were computed by summing involvement dimension scores across game scenes and collapsing both these scores and the rating data across prototype (see Table 3). Our sample size limits power for significance testing to values that exceed a moderate effect size ($r = |.32|$). However, effect sizes in this range indicate support for study hypotheses testing construct validity.

Referring to the correlations in Table 3, we find some support for the hypothesized relationships between the involvement dimensions and our self-report measures of perceived realism, predictability, and flow: Expressiveness was correlated with flow (.51); Altercentricism was correlated with both predictability (−.42) and flow (.43); and Positive Affect was correlated with flow (.36). However, none of the dimensions were correlated at the moderate level with predictability, and neither Composure nor Vocal Anxiety was correlated with any of the selfreport measures.

5.3. Criterion validity of involvement scores across game and scene type

Criterion validity hypotheses were investigated using multivariate analysis of covariance and analysis of variance for repeated measures (MANCOVA, ANOVA-RM) to evaluate

differences in involvement scores for the early and revised prototypes as well as for the final as compared to early game scenes. First, involvement scores were summed across the first three game scenes and submitted to MANCOVA using age as a covariate, prototype (early, revised) as the independent variable, and the six involvement dimensions as the dependent variables. Consistent with our hypothesis, there was a significant main effect for prototype (Pillai's Trace = .424, $F [6, 22] = 2.70$, $p < .04$, $\eta^2 = .42$). Univariate results indicated that Expressiveness and Positive Affect were higher, and Composure lower, in the revised as compared to the early prototype with no significant differences observed in Altercentrism, Interaction Management, or Relaxed Vocalics ($p \geq .15$; see also Table 4).

Second, we used ANOVA-RM to compare involvement dimension scores for the final scene with those obtained for earlier game scenes. This analysis was performed separately for each prototype. Unfortunately, as discussed in the Methods (section 4.4.1.), we were somewhat limited in testing this hypothesis due to audio quality problems. Hence, only three dimensions were assessed in this analysis: Altercentrism, Positive Affect, and Composure. Additionally, Positive Affect scores were re-computed for these analyses using only the facial expression items (see also Methods 4.4.2.). Table 5 depicts results for both the early and revised prototype. Consistent with study hypotheses, the Altercentrism and Positive Affect dimension scores for the early prototype were lower in the final scene as compared to the earlier scenes, and scores on all three involvement dimensions (Altercentricism, Positive Affect, Composure) were significantly lower in the final scene for the revised prototype.

6. Discussion

This project sought to determine whether a method for rating involvement in human dyadic interactions could be used to evaluate involvement in human–avatar interactions in a real-time computer simulation game as an aspect of overall player engagement. Our results provide evidence for both the reliability and validity of this rating system for measuring involvement in players' interactions with avatars in DRAMA-RAMA™. Reliability statistics for the Vocal Relaxation dimension were lower than desirable, but this may reflect the difficulty in using observational data to assess subjective states such as anxiety and tension.

Acceptable-to-good reliability occurred across almost all dimensions of involvement with only a moderate amount of rating training time required to achieve moderate-to-high agreement between raters. These reliability results mirror those in the human dyadic interaction literature (see Guerrero 1994; 1996). The ease of rater training combined with the more than adequate level of agreement makes behavior rating methods an attractive option for researchers interested in assessing the level of involvement as an indicator of engagement in a live simulation game.

The dimensionality of involvement ratings in human–avatar interactions observed in the factor analysis conducted during data preparation supports validity because it is consistent with that which has been previously identified for human dyadic interactions (see Coker & Burgoon, 1987; Guerrero, 2005). However, some of the findings with respect to the

relationships between these dimensions and with other measures of engagement appear at odds with what would be expected given past research regarding human dyadic interactions.

Nevertheless, to fully understand our evidence for validity, it is important to discuss the pattern of correlations and mean differences observed in tests of our validity hypotheses within the context of: (1) the challenges inherent in relating scores obtained for observational-based measures to those obtained for subjective experience self-report measures of distinct but conceptually related constructs; (2) the cultural differences between our sample and the samples used to develop and establish the reliability and validity of Guerrero's (2005) rating system; and (3) the impact of aspects of the game (e.g., technology, physical setup, skill training-related story elements, game scoring) on the communication behaviors being rated. Understanding this evidence is also hampered by a lack of prior research findings regarding inter-correlations between Guerrero's involvement dimensions or correlations between these dimensions and other engagement constructs such as those measured in this study.

Both our rating instrument and our self-report items captured various components in our measurement model of engagement, but these components were quite distinct. Our rating instrument assessed social interaction behaviors while our self-report items measured internal experiences and perceptions (i.e., predictability, flow, and perceived realism). Therefore, we would not necessarily anticipate that the all involvement dimensions would be strongly correlated with every self-report measure of game play-related internal experience. For example, Interaction Management, Composure, and Vocal Relaxation had little or no correlation with any of the self-report measures, and perceived realism was not correlated with any of the involvement dimensions. Clearly additional research is needed to better understand how conceptually distinct measures of engagement (involvement, flow, predictability, etc.) interrelate.

Expressiveness, Positive Affect, and Altercentrism each had substantive correlations with flow. This suggests that energetic, enthusiastic, warm, and friendly social interaction behavior that is focused on the interaction partner (i.e., avatar) is related to losing a sense of time during game play. The negative correlation between Altercentrism and predictability suggests that being able to guess what happens next in a game decreases the player's ability to maintain a focus on the avatar, which is consistent with decreased engagement. Thus, these correlations support the validity of these three involvement dimensions. Furthermore, the pattern of findings obtained in comparisons of the early and revised prototypes and of the early and final game scenes provides support for Composure (i.e., feeling at ease in the interaction). Composure scores were greater under conditions hypothesized to be more involving. Thus, we find support for the validity of all but two involvement dimensions, Interaction Management and Relaxed Vocalics, and would argue for additional research to rule out cultural influences, measurement problems, or true differences between interactions involving avatars and human dyadic interactions contributing to this lack of support.

Although almost of our participants self-identified as Hispanics, the majority of participants in past studies of human dyadic interactions self-identified as White, Anglo, or Caucasian. Some research suggests that, compared with American Whites, Hispanics communicate in a

more animated and energetic way (Albert & Nelson, 1993; Johnson, 2006). Rather than high involvement being marked by increasing composure, relaxation, and decreasing interruptions or talk overs, it is possible that Hispanic cultural interaction norms might mark high involvement with increased energy leading to more talk overs, more interruptions, and vocal excitement. This explanation is consistent with the close-to-moderately sized, negative correlation between perceived realism and Composure and the smaller, negative correlation between Interaction Management and flow that we observed. It might also explain the negative correlation between Positive Affect and Interaction Management. For Hispanic participants, a warm and pleasant conversation with a friend might also include excited vocal tones, talk overs, and animated body movement. Clearly, additional research is needed to understand the potential influence of culture on the use of social interaction behavioral cues in general.

It is also possible that aspects of the game (such as story elements and scoring) combined to alter the conceptual meaning of the behavioral cues in the Composure dimension from how they were originally conceived by Guerrero (2005) and colleagues (Coker & Burgoon, 1987). For example, consistent with criterion validity, we found that participants displayed more overall Expressiveness and Positive Affect while playing the revised as compared to the early prototype. However, these participants also displayed less Composure, which is at odds with how increased involvement has been conceptualized in research involving human dyadic interactions (see Coker & Burgoon, 1987; Guerrero, 2005). Yet, we would argue that the behavior cues consistent with less Composure (increased body movement, fidgeting, etc.) are also consistent with increased arousal and engagement in DRAMA-RAMA™, given the game's story elements and scoring. The game's skill-training story elements were used to create opportunities to resist peer pressure to make risky choices, and the scoring made it clear that points could only be earned by resisting this pressure without insulting the characters or losing face. If these story elements and scoring created a realistic simulation of peer pressure and challenge to socially appropriately respond, one might expect an increase in the player's level of arousal as reflected by lower composure scores.

Additionally, the game's physical set-up appeared to have a substantive impact on the meaning of the Immediacy dimension behavioral cues. Specifically, our decision to seat participants at a table in front of a monitor led many to put their hands/arms on the table as they leaned in towards the avatar during game play. Although they were leaning in (indicative of increased immediacy), their hand and/or arm placement on the table constituted a closed body position, which is defined in Guerrero's rating system as indicating less immediacy. Similarly, when they leaned back (consistent with decreased immediacy), their arms went to their sides, resulting in an open body position, which is defined as indicating greater immediacy.

In summary, it would be a mistake to use our findings to identify distinct differences between human-avatar and human dyadic interaction communication or to infer problems with the validity of using Guerrero's system to capture involvement in skill training games that use human-avatar interaction. Instead, our findings highlight the importance of considering how the physical setup for playing such games can impact the researcher's ability to measure human-avatar communication.

6.1. Limitations

There are three limitations to the research presented here. First, the human–avatar interactions, although synchronous, were relatively short. Moreover, the presence of an in-game scoring mechanism necessarily meant that the conversations between the player and avatars were being assessed and quantified in a very visual way, unlike normal human dyadic conversation. Thus, additional research is needed with other applications that foster longer interactions and with applications that do not use in-game scoring of interaction behavior to fully address the differences and similarities between human dyadic interactions and human–avatar interactions. Such research should involve the use of all of Guerrero's (2005) rating items. However, we found Guerrero's method adaptable and useful for the purpose of measuring the involvement component of engagement in our skill training simulation game.

Second, the quality of the video in some cases was less than desirable, contributing to missing data in these analyses. In future projects, a more strategic recording arrangement would improve raters' observations, thus further improving performance of the measurement instrument and its capacity to fully assess involvement. Additionally, we encourage other researchers to perform a number of early field tests of audio- and video-capture capabilities that incorporate planned playing conditions, anticipated variation in sizes of players, and behavioral rating of these audio-video data. Our early field testing incorporated only the planned playing conditions, which undoubtedly impacted the quality and quantity of our audio-video data.

Third, audio quality and having the player sit at a table in front of a large monitor while playing the game contributed to problems with assessing the Immediacy dimension of involvement. Thus, we cannot offer any conclusions as to the absence or presence of this aspect of involvement with respect to human–avatar interactions using results from this study.

6.2. Directions for future research

Despite these limitations, our findings have at least three implications for simulation and training research. First, although we failed to observe back channeling and nods, findings from previous work (Norris et al., 2013) indicate that the desired training effects did occur without eliciting these behavioral cues. This finding suggests that these aspects of involvement behavior may not be critical to ensure training effects for our targeted skillset. The importance of these behavioral cues could vary with the nature of the skill training. For example, the ability to elicit these cues might be important for applications designed to train social skills in children with autism spectrum disorders or to train psychotherapists. For such games, the use of Guerrero's (2005) involvement rating system could be critical during iterative game testing and refinement to ensure that back channeling and nods were elicited.

Second, it is important in future research using Guerrero's (2005) rating system to examine study results within the context of the particular simulation game story and technology being studied. For example, if we had placed the game play monitor on a stand instead of a table, the forward lean and body orientation items might have performed quite differently and been

more consistent with the conceptual definitions in the rating system. Similarly, the lack of participant nods and back channels may be more an artifact of our game and its technology than a phenomenon unique to human–avatar interactions. These behaviors, especially head nods, are heavily mimicked or mirrored during human dyadic interactions. Head nods typically require a certain length of dialogue to be elicited, but our game interactions were kept short to meet the design goal of a 12–15 minute skill-training session. This may have prevented the head nods from emerging during the interaction. Additionally, we used Skype™ as a secure and low-cost approach for transmitting verbal and nonverbal information, but this caused the drop out of any back channel information voiced when a player or avatar was already speaking. Given the nature of back channels (i.e., that they *only* occur when the conversation partner is speaking), it is likely that this was the cause of our failure to observe back channels in this study.

Third, aspects of this rating system may be useful for game design. This system allows repeated assessment of involvement during game play without interrupting play to have the player self-report or relying on recall or a summary measure once play is complete. In addition, close examination of behavioral cues we conducted uncovered a very low level of involvement (and hence engagement) during both prototypes' final debriefing scene. We had assumed that participants would score low on involvement in this scene due to our use of a disembodied voice to deliver verbal feedback to accompany a numerical score and canned animation of game avatars clapping. What we did not anticipate, and had missed in our previous analysis of game effects, was just how disengaged participants became. In particular, the decrease in Altercentrism suggested that participants were least attentive, alert, and oriented toward the game when receiving feedback about their game choices. This finding was consistent with comments made by our raters during the rating process and has prompted us to redesign this final debriefing scene (i.e., use an avatar to embody the voice delivering feedback) to increase the potency of the DRAMA-RAMA™ skill training.

Moving beyond mere assessment of the involvement aspect of engagement in game play, we would argue that Guerrero's rating system provides opportunity for simulation training and game design teams to identify the extent to which simulated conditions map and fail to map to real-life social situations. This mapping process is critical to advancing the science of both skill-training intervention research and effective game design.

Acknowledgments

Funding for this research was provided by the University of Central Florida's College of Nursing, Institute for Simulation & Training, and Office of Research & Commercialization, and by the National Institute of Nursing Research (R15NR012189).

References

- Albert RD, Nelson GL. Hispanic/Anglo-American differences in attributions to paralinguistic behavior. *International Journal of Intercultural Relations*. 1993; 17:19–40.
- Brockmeyer JH, Fox CH, Curtiss KA, McBroom E, Burkhard KM, Pidruzny JN. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*. 2009; 45:624–634.

- Charoenying T. Accountable game design: Structuring the dynamics of student learning interactions. *Journal of Educational Computing Research*. 2010; 43(2):135–163.
- Chen G-D, Lee J-H, Wang C-Y, Chao P-Y, Li L-Y, Lee T-Y. An empathic avatar in a computer-aided learning program to encourage and persuade learners. *Educational Technology & Society*. 2012; 15(2):62–72.
- Coker DA, Burgoon J. The nature of conversational involvement and nonverbal encoding patterns. *Human Communication Research*. 1987; 13(4):463–494.
- Csikszentmihalyi, M. *Flow: The psychology of optimal experience*. New York: Harper Perennial; 1990.
- Dickey MD. Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development*. 2005; 53(2): 67–83.
- Experience Builders LLC. *The Selling Experience*. 2006. Retrieved from <http://www.experiencebuilders.com/demos/TheSellingExperience/default.htm>
- Fu F-L, Su R-C, Yu S-C. eGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*. 2009; 52(1):101–112.
- Garris R, Ahlers R, Driskell JE. Games, motivation, and learning: A research and practice model. *Simulation & Gaming*. 2002; 33(4):441–467.
- Gee JP. Learning by design: Good video games as learning machines. *E-Learning*. 2005; 2(1):5–16.
- Guadagno RE, Swinth KR, Blascovich J. Social evaluations of embodied agents and avatars. *Computers in Human Behavior*. 2011; 27:2380–2385.
- Guadagno RE, Blascovich J, Bailenson JN, McCall C. Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology*. 2007; 10(1):1–22.
- Guerrero, LK. An application of attachment theory to relational messages and nonverbal involvement behaviors in romantic relationships. (Doctoral dissertation). 1994. Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 9432854)
- Guerrero LK. Attachment-style differences in intimacy and involvement: A test of the four-category model. *Communication Monographs*. 1996; 63:269–292.
- Guerrero, LK. *The sourcebook of nonverbal measures: Going beyond words*. Mahway, NJ: LEA; 2005. Observer ratings of nonverbal involvement and immediacy; p. 221–235.
- IJsselsteijn, W.; Poels, K.; de Kort, YAW. *The Game Experience Questionnaire: Development of a self-report measure to assess played experiences of digital games*. Eindhoven: TU Eindhoven; 2008.
- Involvement. (n.d.). Oxford English Dictionary Online. Retrieved from <http://www.oed.com>
- Johnson R. Confounding influences on police detection of suspiciousness. *Journal of Criminal Justice*. 2006; 34:435–442.
- Koster, R. *A theory of fun for game design*. Scottsdale, AZ: Paraglyph Press; 2005.
- La France BH, Heisel AD, Beatty MJ. A test of the cognitive load hypothesis: Investigating the impact of number of nonverbal cues coded and length of coding session on observer accuracy. *Communication Reports*. 2007; 20(1):11–23.
- Lee JK, Hecht ML, Miller-Day M, Elek E. Evaluating mediated perception of narrative health messages: The perception of narrative performance scale. *Communication Methods and Measures*. 2011; 5(2):126–145. [PubMed: 21822459]
- Lim S, Reeves B. Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *Journal of Human-Computer Studies*. 2009; 68:57–68.
- Lopes R, Bidarra R. Adaptivity challenges in games and simulations: A survey. *IEEE Transactions on Computational Intelligence and AI in Games*. 2011; 3(2):85–99.
- Markham, C.; Peskin, M.; Shegog, R.; Tortolero, S. Social support. *It's Your Game: A theory-based pregnancy, HIV, and sexually transmitted disease prevention program*. 2014. Retrieved from <https://sph.uth.edu/iyg/social-support/>
- Mayes DK, Cotton JE. Measuring engagement in video games: A questionnaire. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2001; 45:692–696.

- McMahan, A. Immersion, engagement, and presence: A method for analyzing 3-D video games. In: Wolf, MJP.; Perron, B., editors. *The video game theory reader*. New York, NY: Taylor & Francis; 2003. p. 67-86.
- Moneta GB, Csikszentmihalyi M. The effect of perceived challenges and skills on the quality of subjective experience. *Journal of Personality*. 1996; 64:275–310. [PubMed: 8656320]
- Miller-Day M, Hecht ML. Narrative means to preventative ends: A narrative engagement framework for designing prevention interventions. *Health Communication*. 2013 Advance online publication.
- Nehring WM, Lashley FR. Nursing simulation: A review of the past 40 years. *Simulation & Gaming*. 2009; 40:528–552.
- Norris AE, Hughes C, Hecht ML, Peragallo N, Nickerson D. Randomized trial of a peer resistance skill-building game for Hispanic early adolescent girls. *Nursing Research*. 2013; 62(1):25–35. [PubMed: 23150043]
- Osborne, JW.; Costello, AB.; Kellow, JT. Best practices in explanatory factor analysis. In: Osborne, JW., editor. *Best practices in quantitative methods*. Thousand Oaks, CA: Sage; 2008. p. 86-101.
- Oswalt I. Current applications, trends, and organizations in U.S. military simulation and gaming. *Simulation & Gaming*. 1993; 24(2):153–189.
- Procci K, Bowers C. An examination of flow and immersion in games. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2011; 55:2183–2187.
- Salen, K.; Zimmerman, E. *Rules of play: Game design fundamentals*. Cambridge, MA: MIT Press; 2004.
- Sherhoff DJ, Csikszentmihalyi M, Schneider B, Sherhoff ES. Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*. 2003; 18(2):154–166.
- Smith R. The long history of gaming in military training. *Simulation & Gaming*. 2010; 41:6–19.
- Sweetser P, Wyeth P. Game Flow: A model for evaluating player enjoyment in games. *Computers in Entertainment*. 2005; 3(3):1–24.
- Vogel JJ, Vogel DS, Cannon-Bowers J, Bowers CA, Muse K, Wright M. Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*. 2006; 34(3):229–243.
- von der Pütten, Krämer NC, Gratch J, Kang S. "It doesn't matter what you are!" explaining social effects of agents and avatars. *Computers in Human Behavior*. 2010; 26:1641–1650.
- Wirth, J. *Interactive acting: Acting, improvisation, and interacting for audience participatory theatre*. Fall Creek, OR: Fall Creek Publishing; 1994.
- Witmer BG, Singer MJ. Measuring presence in virtual environments: A presence questionnaire. *Presence*. 1998; 7(3):225–240.

Highlights

- A system for rating involvement in live human-avatar interaction games was tested.
- Moderate-to-high agreement between raters was achieved with moderate training time.
- Factor structure and reliability mirrored those for human dyadic interactions.
- Engaging features of game design associated with higher involvement ratings.
- Involvement rating system is useful to the design of simulation training games.

Figure 1.
Theoretical Model of Engagement

Figure 2.

a. & b. Top, an inter-actor is shown wearing a motion capture hat and using an ErgoDex keypad and iPad to control avatar movement during game play. Head motions are captured and mapped in real time to the avatar she is controlling. The inter-actor's display includes a virtual view of the game (including any character[s] that can be controlled), live video feed of the participant, and an instant message window tracking branching choices. Below, a game player talks with an inter-actor-controlled avatar.

Figure 3.

a. & b. Environments from the early prototype. Note low level of detail and coloration. Early prototype created by SREAL, Institute for Simulation & Training, University of Central Florida.

Figure 4.

a. & b. Environments from the revised prototype. Note increased detail and use of color. Revised prototype created by E2i Creative Studios, Institute for Simulation & Training, University of Central Florida.

Table 1

Primary factor loadings, inter-rater and scale reliability analyses

Factor Name	Variable Behavioral rating item	Primary Loading	Intraclass r between raters	Scale α
Vocalic Expressiveness				.97
	Expressiveness -- Inexpressive	.94	.85	
	High variation -- Low variation	.93	.78	
	Dull -- Full of life	.91	.86	
Altercentrism				.94
	Attentive -- Inattentive	.95	.74	
	Focal -- Nonfocal gaze	.90	.68	
	Distracted -- Focused	.79	.75	
	Steady -- Random gaze	.81	.50	
	Alert -- Not Alert	.65	.76	
Interaction Management				.93
	Speaks-- Does not speak when expected	.86	.85	
	Many -- No interruptions	.95	.93	
	Many-- No talk overs	.88	.93	
	Long-- Short response latencies	.74	.77	
Positive Affect				.89
	Positive-- Negative facial expression	.87	.81	
	Warm-- Cold facial expression	.88	.81	
	Much -- No smiling	.77	.85	
	Friendly -- Unfriendly Voice	.63	.82	
	Cold -- Warm Voice	.52	.70	
Composure				.72
	Large-- Small amount of body movement	.79	.83	
	Large amount-- No fidgeting	.64	.83	
	Frequent adapters -- No adapters	.59	.79	
	Loose -- Rigid	.58	.62	
Vocal Relaxation				.60
	Anxious -- Relaxed	.76	.42	
	Tense -- Calm	.65	.49	
No Clear Factor Loading				--
	Much laughter -- No laughter	.43	.75	
	High--Low Rocking/Twisting	.58	.49	
	Leans toward -- Leans away	.41	.78	
	Closed-- Open body position	.29	.60	
	Fluent -- Nonfluent	.42	.60	
	Many -- Few gestures	.51	.78	
	Tense -- Relaxed Body	.55	.62	
	Detached -- Involved	.58	.79	
	Bored -- Interested Facial Expressions	.61	.76	

Factor Name	Variable Behavioral rating item	Primary Loading	Intraclass r between raters	Scale α
	Choppy – Smooth delivery	.57	.57	

Table 2

Involvement scale means and intercorrelations, all participants, n = 31

	M	1	2	3	4	5	6
(SD)							
1. Expressiveness	3.97 (1.43)		.55**	-.35*	.54**	-.13	.11
2. Altercentrism	5.35 (.85)		--	-.02	.32*	-.18	.22
3. Interaction Management	6.11 (1.63)			--	-.45**	-.08	-.13
4. Positive Affect	4.02 (1.10)				--	.21	.17
5. Composure	2.99 (1.23)					--	-.10
6. Vocal relaxation	5.78 (.56)						--

**
 $p < .01$;*
 $p < .05$

Pearson correlations between involvement dimension scores and participant self-report measures, all participants, n = 31.

Table 3

Self-report measures:	Nonverbal Involvement Ratings					
	Expressiveness	Altercentrism	Interaction Management	Positive Affect	Composure	Vocal Relaxation
Perceived Realism	-.08	.19	.20	-.18	-.29	-.16
Predictability	-.05	-.42**	-.25	.15	.04	.11
Flow	.51**	.43**	-.22	.36*	-.19	-.10
*** $p < .01$;						
* $p < .05$						

Results of MANCOVA testing differences in involvement dimension scores for early and revised game prototypes, using participant age as a covariate

Table 4

Variable	Early Prototype <i>n</i> = 18	Revised Prototype <i>n</i> = 13	<i>F</i>	<i>p</i>	η^2
Expressiveness	2.99 (.63)	4.39 (1.19)	18.78	.00	.41
Altercentrism	4.60 (1.12)	4.86 (.90)	.46	.51	.02
Interaction Management	2.99 (.57)	3.22 (2.71)	.38	.54	.01
Positive Affect	3.21 (.99)	4.14 (1.16)	6.09	.02	.18
Composure	3.13 (.95)	2.41 (.88)	4.57	.04	.15
Relaxed Vocalics	5.86 (.26)	5.79 (.25)	1.24	.28	.04

* Equal variances not assumed

Standard deviations are in parentheses

Table 5

Early and revised prototype scene specific involvement dimension score means and results of repeated measures ANOVA

Game/Variable	Scene 1	Scene 2	Scene 3	Scene 4	F	p	η^2
<i>Early Prototype n = 16^a</i>							
Altercentrism	5.39 ^a	5.16 ^a	5.24 ^a	3.83 ^b	7.49	.00	.31
Positive Affect ^b	7.74 ^a	3.72 ^{ab}	3.41 ^b	3.00 ^c	4.99	.00	.20
Composure	3.65	3.15	3.27	2.62	2.44	.07	.13
<i>Revised Prototype n = 12^a</i>							
Altercentrism	5.25 ^a	5.36 ^a	5.41 ^a	3.42 ^b	21.05	.00	.66
Positive Affect ^b	4.37 ^a	4.13 ^{ab}	4.23 ^{ab}	3.55 ^b	2.84	.05	.18
Composure	2.65 ^a	2.65 ^a	2.83 ^a	1.48 ^b	12.04	.00	.52

Note: Means in rows that do not share a subscript differ at the $p < .05$ level of significance.

^aNo Scene 4 data were available for two participants who played the early prototype and one participant who played in revised prototype

^bPositive Affect is computed without vocalic measures across all scenes