Werewolves, Cheats, and Cultural Sensitivity

Ruth Aylett, MeiYii Lim,
Heriot-Watt University
Riccarton, Edinburgh EH14 4AS UK
{r.s.aylett}{m.y.lim}@hw.ac.uk

Lynn Hall, SarahTazzyman,
University of Sunderland, UK
Sunderland SR1 3SD, UK
l.hall@sunderland.ac.uk;
snt5@leicester.ac.uk

Birgit Endrass, Elisabeth Andre,
University of Augsburg
8651 Augsburg, Germany
endrass@hcm-lab.de,
andre@informatik.uni-augsburg.de

Chistopher Ritter, Asad Nazir,
Heriot-Watt University, UK
{c.ritter}{a.nazir}@hw.ac.uk

Ana Paiva
INESC-ID, Portugal
a.paiva@inesc-id.pt

GertJan Hofstede
Wageningen University, Netherlands
gertjan.hoftstede@wur.nl

Arvid Kappas
Jacobs University, Germany
a.kappas@jacobs-university.de

ABSTRACT
This paper discusses the design and evaluation of the system MIXER (Moderating Interactions for Cross-Cultural Empathic Relationships), which applies a novel approach to the education of children in cultural sensitivity. MIXER incorporates intelligent affective and interactive characters, including a model of a Theory of Mind mechanism, in a simulated virtual world. We discuss the relevant pedagogical approaches, related work, the underlying mind model used for MIXER agents as well as its innovative interaction interface utilising a tablet computer and a pictorial interaction language. We then consider the evaluation of the system, whether this shows it met its pedagogical objectives, and what can be learned from our results.

Categories and Subject Descriptors
I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence - Intelligent agents

Keywords
Synthetic Characters, Intelligent Virtual Agents, Empathy, Cultural sensitivity, Models of Personality, Emotion and Social/Cultural Behaviour

1. INTRODUCTION
In this paper we discuss a virtual drama approach to the education of primary-school children aged 9-11 in cultural sensitivity. We follow Hofstede's [18] widely-used definition of culture: "Culture is the collective programming of the human mind that distinguishes the members of one human group from those of another. Culture in this sense is a system of collectively held values". As the use of the term *values* suggests, this can result in a strongly positive affective attachment to specific ways of doing things and in an equally strong negative affective response to those whose collectively held values mean they do the same things differently. In an increasingly globalised society, people from quite different cultures mix more than ever before, and cultural misunderstanding is therefore also more widespread, sometimes resulting in serious conflict. Education and training in cultural diversity is an obvious response and now takes place at multiple social levels, from primary schools to large business organisations.

One approach to this type of education/training is informational and descriptive. It is based on the assumption that ignorance is the root of intercultural antipathy. Thus different aspects of an unfamiliar culture are presented in order to make them less unfamiliar. In UK primary education, it might take the form of describing characteristic festivals of local populations considered 'different', say Hannukah, Divali, Eid, with an explanation of why they are significant in their home culture. While an understandable starting point, especially given the importance of specific social rituals in a given culture, this also has some obvious limitations. Firstly, it risks presenting a different culture as exotic or quaint, in the patronising style of orientalism [26]. Secondly, it assumes that merely *knowing* about a different culture is enough to produce sensitivity, rather than explicitly addressing attitudes and behaviour. Thus it ignores the affective issues altogether. Thirdly, this focus on knowledge also fails to take into account models of how cultural sensitivity develops.

The development of cultural sensitivity can be modelled as a process of dynamic change in how individuals handle cultural differences between themselves and others, as in [5]. Bennett's model is formed of six stages along a continuum of intercultural development, of which three are ethnocentric (denial, defences, minimization) and three are ethno-relative (acceptance, adaptation, integration). An alternative but compatible view of this process sees it as developing acceptance of people belonging to a given out-group into ones own moral circle. A moral circle [28] is formed by those who adhere to a common purpose and group identity; those who follow a common set of moral rules and values and who trust each other [19]. Each social setting creates its own moral circle consisting of all people who belong to that setting, its rules only applying to those members. These unwritten rules for behaviour are used to monitor our own and others' performance. All kinds of covert and overt feedback signals and sanctions are used to try to enforce good behaviour. The solidity or permeability of the boundary of one's moral circle is co-determined by one's culture.

The concept of the moral circle and of in- and out-groups underlies the work discussed in this paper. A modified version of

the approach of [5] has been used to develop the framework seen below in Table 1.

In this paper we discuss the MIXER (Moderating Interactions for Cross-Cultural Empathic Relationships) application developed in the eCUTE (Education in Cultural Understanding, Technology Enhanced) project [13], which applies a novel approach to the education of children in cultural sensitivity. MIXER incorporates intelligent affective and interactive characters, equipped with a Theory of Mind mechanism, in a simulated virtual world. We discuss the relevant pedagogical approaches, related work, the underlying mind model used for MIXER agents as well as its innovative interaction interface utilising a tablet computer. We conclude with the evaluation of the system and what can be learned from it.

Table 1. Stages of Development of Cultural Sensitivity

Attitude Stage of Learner	**Emotional goals**	**Cognitive goals**	**Behavioural goals**
Beginner *conscious incompetence* Observation and Acquisition	Be able to recognise own emotions (eg. Fear, anxiety) when dealing with strange behaviours of another group	Start learning the specific practices and values of another group	Be fully present in attending to another's verbal and non-verbal messages
Journeyman *conscious competence* Relating and Experimenting	Be able to observe the behaviour of another group without feeling prejudice	Understand on a basic level differences and similarities between own group and another	Practice skills learned in previous stage; experiment with different forms of behaviour
Expert *unconscious competence* Adapting and Belonging	Be able to share emotions (eg. Sadness, happiness) of members of another group and others' experiences through empathy	Discriminate and select appropriate strategies for the cultural context	Be able to unconsciously participate in another group as a native

2. BACKGROUND

Most intercultural training programs can be categorised along two dimensions. The first, *didactic/experiential*, refers to the process of training, while the second *culture-general/culture-specific,* is applied to its content [6]. Methods which are primarily didactic aim at achieving a cognitive understanding of a given topic through equipping trainees with factual information about another culture, as already discussed above. An experiential approach rests on the assumption that the best way to teach trainees about another culture is to have them experience it. Experiential methods include role plays, in which people act as if they were engaged in a real cross-cultural encounter and simulation games, of which the most popular is BaFa BaFa [27] with its version for children called RaFa RaFa.

Role-play supports the creation of knowledge and meaning from concrete - though imagined - experiences [22]. It uses social interaction and emotional engagement as mechanisms through the invocation of empathy as an individual attempts to understand another by imagining that other's perspective. This may then influence the subjective evaluation of the individual regarding their own social skills and lead to significant behavioural changes.

As role play can be difficult and expensive to organize, requiring skilled facilitation and supporting materials, as well as actors, recent work has often used virtual agents as substitutes for role play actors and training partners e.g. [3, 14, 21, 24, 23, 33]. Some of these systems use real-world cultures as the basis for their virtual agents' cultural model [14, 21, 33], while others use invented synthetic cultures so as to avoid possibly offensive stereotyping of existing cultures [3, 23]. However, none of them are aimed at the primary school age group targeted by MIXER.

Issues of addressing 9-11 year olds include the level of abstraction required to conceptualise culture along with the strength of the Similarity Principle [8] which says that interaction partners that perceive themselves as similar are more likely to feel positive about each other and to display empathic behaviour. This age group have a more limited experience of social life than adults or even adolescents, and are more likely to relate to family or school contexts in which cultural difference is embedded in very specific pieces of behaviour. School is the most general social environment they are likely to have in common and friendship circles their major experience of in- and out-group behaviour. Thus the aliens of ORIENT [3] seem just as inappropriate for this age group as the adult Arab culture modelled in [21].

The experiential mechanism used to train in cultural sensitivity is normally communication frustration or failure. The first stages of Bennett's model [5] refer to *noticing* cultural difference on the grounds that what one does not notice, one cannot adapt to. In the same way, the Beginner Emotional Goals of Table 1 above talk of *recognizing* emotions such as fear and anxiety as a first step towards dealing with them. There is a risk in this approach in that it involves generating negative emotions, and if successful reflection does not take place, this could result in entrenching conflict rather than resolving it. This risk is higher with child groups since they typically have less ability to reflect on their emotions than older participants.

Taking the above points into account, MIXER focuses on schools as its basic social unit and the rules of a specific game as its framework for cultural conflict. Here it draws on the card-game BARNGA [29], which is successfully used for cultural training, making an analogy between game and social rules within the framework of a social game. BARNGA involves pairs of players moving successively to play with new pairs without knowing that slightly different rules for the card game they are all engaged in have been supplied to different groups. Groups are forbidden to communicate through speech, having to use writing and diagrams to resolve conflicts.

BARNGA typically results in a substantial amount of conflict in the play phase, with players often ignoring the 'no speech' rule and shouting at each other, and accusations of cheating. Fundamental to its successful use is the debrief phase that follows the play phase in which reflection takes place and the parallel with cultural norms is internalized.

3. MIXER

MIXER is a narrative-based game containing intelligent virtual agents, designed for children between 9 and 11 years old. The scene is set in a virtual summer camp where two groups of school-children (intelligent virtual agents) come together. A card game like BARNGA was not selected as we felt the visualization of cards would detract from a focus on the virtual characters,

important for empathic engagement. Instead, the virtual agents play *Werewolves*, an enjoyable party game already widely known in many cultures (sometimes under an alternative name: *Mafia*).

In Werewolves, all players except a game-master or narrator are secretly assigned roles: Werewolves, who know each other's identity if there are more than one of them; or Villagers, who know only the number of Werewolves amongst them. The main aim of the game is for the Werewolves and the Villagers each to work to eliminate the other group through strategic game play.

The original game has two phases. In the night phase, all players shut their eyes. The Werewolves then wake up and 'kill' a Villager. During the day phase, the remaining players make accusations about the Werewolves' identity, and vote to eliminate a suspect. The chosen suspect is also 'killed' and then reveals their identity – if not the last Werewolf, then the game continues. Day and night phases alternate until either all the Werewolves have been found or until the Werewolves equal the number of Villagers. A key element of the game is that the player acting as a Werewolf must deceive the other players about their identity.

In MIXER, the game is played by two different teams of six intelligent agents (the Yellows and the Reds, with appropriately coloured T-shirts). They include a game-master, a werewolf and four villagers in each team. The intelligent virtual agents fall into the same age group as the child user, and can be seen in Figures 1 and 2. One of the virtual players, Tom (in the version discussed here) or Lisa (added as one outcome of the results below), is new to the summer camp. The rules are explained to them by the first team, from their own school, and as they play the game, the child user acts as their friend, watching the action and responding to requests for advice on how to react at different stages of the game. Communication is carried out through a tablet application, which uses a pictorial interaction language discussed below.

The child user plays the role of *invisible friend*. They do not directly appear in the virtual world but only interact with the friend character. Earlier work has established the importance of empathic engagement between a child user and a virtual character in achieving pedagogical outcomes [15]. Detaching the child user from actually playing the game themselves is also intended to produce a distancing effect so as to promote reflection [2].

The two teams of IVAs play the game with a significant difference in their rules, which results in a strong episode of conflict as Tom/Lisa switches from one team to another. Because of this different rule, the friend character is "killed" in the very first round of the game where they would not have been using the first set of rules. The child user is then asked for support. The aim is to create reflection on how what appears to be unfair behaviour may actually be due to a different set of rules, hence learning that there is more than one way to run a social situation. This relates to the Beginner stage of the process shown above in Table 1 and lays the basis for accepting people belonging to a given out-group into one's own 'moral circles' [19]. It also encourages thought about the appropriate coping strategy in conflict situations [30].

Figure 1: MIXER characters playing Werewolves

Figure 2: Playing an accusation turn

4. Implementation

The implementation of MIXER drew on existing research software. The intelligent agent 'minds' were developed using the affective agent architecture FAtiMA [11]. The ION middleware [32] was used to connect agent minds to other facilities, including the tablet used for user interaction, and the graphical visualization of the world, implemented in the Unity game engine [31].

Technical innovation lay in two areas: the first was the development of Theory of Mind (ToM) capabilities for the characters, and the second an innovative interaction interface for the child users.

4.1 Virtual Agents ToM

One of the key aspects of MIXER is the application of intelligent affective virtual agents. Such agents with their own goals, motivations and emotions, autonomously selecting their actions, give users a sense that they are 'alive' and hence believable. The utilisation of IVAs in MIXER also has other advantages compared to real-world role-play. Firstly, consistent and repeatable expressive behaviour can be maintained throughout the interaction with users via careful parameterisation and generation of the IVAs actions and expressions. This is important, as the verbal and nonverbal behaviour patterns of the agents will influence children's perception of ethnicity and subsequently their interaction behaviour [20]. Secondly, the MIXER application can be easily multiplied and distributed to allow one-to-one interaction with a whole class of children which is much more difficult with real actors. Lastly, this approach enables reusability. Once a parameterisable IVA has been developed, it can be used in an indefinite number of scenarios with an indefinite number of cultural variables to control the degree of exaggeration in portraying synthetic cultures.

This is achieved using FAtiMA, a cognitive appraisal architecture [11], in which events are assessed in relation to goals and generate an affective outcome as a result. These affective states are then used both in reactive action-tendency rules for immediate responses and in a deliberative planning system driven by the emotions hope and fear [11], generating more complex coping behaviours. FAtiMA also incorporates an autobiographic memory.

However, the Werewolves game poses interesting problems for an autonomous agent. The IVAs playing Villagers must have a rationale for their accusations of other IVAs as the Werewolf. The IVA that is playing the Werewolf most have a more complex strategy for making accusations that will divert attention from themselves and must respond to accusations from others in a believable way. Thus an IVA acting as a Werewolf must be capable of deceitful behaviour, requiring ToM capabilities to predict reactions to possible accusations.

Two different approaches can be taken to incorporating ToM capabilities: one based on reasoning across declarative models of the state of other agents, and a second based on simulating the running of the mind of the other agents in their current situation [17]. In a cognitive appraisal architecture, which can be thought of as a runnable Agent Model, a simulation approach is an economical one since it allows the agent to rerun its own architecture with a paremetrisation for another agent (Figure 3).

The ToM used in MIXER encodes equivalents to Baron-Cohen's [4] Eye-Direction Detection (EDD) and Shared-Attention Mechanism (SAM) capabilities, which work on an agent's incoming perception so it can run the ToM mechanism for the perspective of a different agent [12] – see Figure 4.

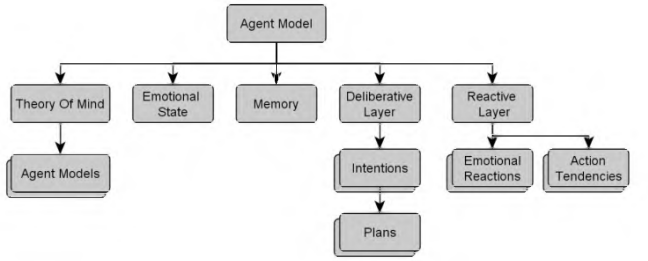

Figure 3: FAtiMA extended with ToM

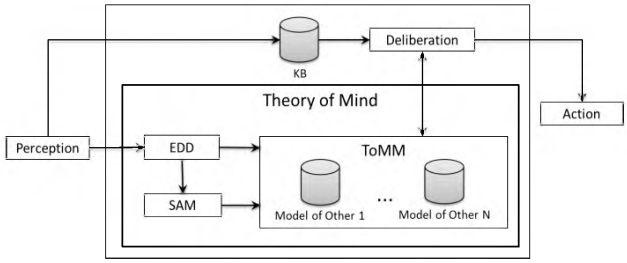

Figure 4: EDD and SAM interfacing with Perception

The MIXER agents' ToM mechanism consists of a collection of Models of Others, each representing the beliefs of a particular known agent. A two-level ToM is modelled which means that a Model of Other will have its own ToM including Models of Others, resulting in a recursive hierarchical tree. While in princi0ple an n-level ToM could be included, two levels is the number required to execute plausible Werewolf strategies.

Thus, a Villager suspects those that were accused by someone it does not suspect; stops suspecting someone who accuses a target it suspects; suspects those who accused a victim that was eliminated in the previous round; thinks that someone who accuses a target suspects that the target is a Werewolf and someone that is accused will suspect the accuser. The Werewolf agent on the other hand will 'lay low' by blending in and avoiding suspicious actions that could denounce him. It will try to deceive the others by trying to make victims believe that he thinks the same way as they do. Using the two-level ToM, a Werewolf will therefore accuse a Villager that is already being accused by another Villager.

Authoring IVAs' behaviour involves establishing their specific goals, actions, inference rules, emotional reactions and action tendencies [11]. Two types of goals are modelled. Active-pursuit goals represent goals that the character actively pursues to achieve a certain state (e.g. accuse someone) while interest goals are goals that the character continuously maintains to avoid threatening situations (e.g. to stay healthy). An active-pursuit goal is activated only if its preconditions are satisfied. Emotion rules are responsible for generating emotions that trigger action tendencies, which specify the actions to be carried out when the character experiences a particular emotional state. Lastly, an inference rule updates the KB with new knowledge.

Goals are constantly checked by the deliberative layer to see if their preconditions have become true. Once a goal becomes active, a new intention is created which represents the agent's commitment to achieve the goal. The agent then creates a plan containing potential action(s) to be performed. Through the specification of ToM preconditions to be tested against a particular Model of Other, the deliberative component is able to trigger goals according to the beliefs of others.

When the deliberative component finds such a precondition it starts by traversing the tree hierarchy of Models of Others and selecting the corresponding Model Of Other to be tested. To model higher-level goals and inference rules – that is, explicit goals and rules to change the mental states of others - local and global effects have to be specified. When an inference rule has an effect on another agent, instead of updating its own KB, the deliberative layer traverse the tree hierarchy in order to update the corresponding Models of Others. As a result, emergent behaviour and story develop as the IVAs play the game, through the changes in their reactions, accusation and interaction sequences providing the users with a rich narrative experience.

4.2 Pictorial Interaction Language (PIL)

MIXER is designed to be used across a variety of countries and cultures and thus an innovative interaction interface was designed for child users. While in theory free text entry gives maximum self-expression, in practice the 9-11 age group includes some very slow keyboarders who may be more focused when typing on the effort involved than the content. On the other hand menus and buttons restrict expressiveness, often to single-item choices.

Touch screens have become widely available through tablets and

Figure 5: Interacting with MIXER via tablet and PIL

smart phones in the recent period, and it was decided that a tablet would provide a good physical interface for interaction, with the advantage that the MIXER screen itself would not become cluttered. This also reinforces the child's role as invisible friend since only they see their communication with Tom – see Figure 5.

The basic interaction concept was inspired by Crawford's Toy Language Deiko [10] designed for interactive storytelling. Deiko defines seven word types, termed: Actor, Stage, Prop, Verb, Event, Attribute, and Quantifier. Four of these - Actor, Stage, Prop, Verb – are thought of as basic. Graphical editing then allows words that are instances of each to be put together into sentences respecting lexicon, grammar and term-relationship constraints. This constrained language is much more expressive than menus and buttons but without the downsides of free typing.

However Deiko is designed for an adult user and works in English only. For the MIXER 9-11 year old target users, the Deiko concept was adapted into a Pictorial Interaction language (PIL) in which each term was represented with a pictorial icon that could be pushed into position on the tablet touch screen.

The vocabulary for the PIL was derived from user-studies. Given the rules of the game, some initial words needed were obvious, such as 'You', 'They', 'Accuse', 'Defend', 'Werewolf' etc. We recruited a total of 70 children (aged 9 to 11) who played the real world Werewolf card game in small groups. The games were recorded and transcribed. In total, we identified 60 frequently used words, such as "I **accused** her because she **looks suspicious**" or "he's being too **quiet**". These words were later grouped, for example as emotions or actions, and structured in a way to match the interaction modes we wanted to use for the game play. As a result of this study, a set of over 60 icons were created.

This set was then used in a further study with 30 children in which their understanding of the icons were tested and they helped to redesign those not well-understood. A further 25 children from a different school were then recruited to test the new set in focus groups and all these icons were successfully identified. Figure 6 shows a small subset of the PIL icons along with their intended meanings.

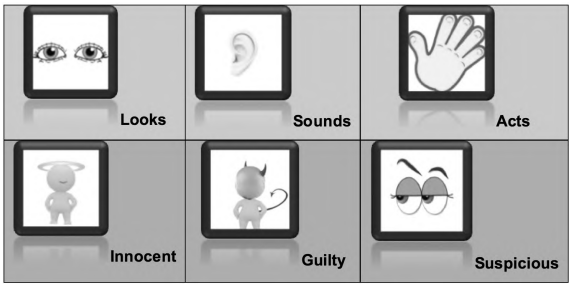

Figure 6: Some PIL icons with their meaning

The Werewolf game itself provides a very strong context for interaction between child user and friend character. One interaction type occurs during the game play and the second when the conflict situation when playing with the second group has taken place. Four different advisory modes were created for the game play cases where user interaction is sought:

DG1: Questioning who is the werewolf
DG2: Reasoning why somebody is the werewolf
DG3: Reacting if another character is being accused
DG4: Reacting if own friend character is being accused

This allowed only those icons required in a particular context to be displayed at the top of the tablet, with same-colour slots along the bottom of the tablet into which icons could be dragged (Figure 7). Depending on what is actually happening in the game, these modes can be used alone or combined to simulate a longer interaction between a child and Tom. For example, after Tom asked who might be the werewolf (DG1), he can ask why the child thinks so (DG2).

Figure 7: PIL interface running on an iPad

. A final test of the PIL, run with 66 children showed that: over three quarters said that the PIL was fun to play with [χ^2 (4, 65) = 41.39, $p < .001$, higher proportion of children rated the PIL as 'fun' than expected given equality across cells]. They found it a good way to play the MIXER game (83%) [χ^2 (4, 65) = 57.23, $p < .001$, higher proportion of children than expected rated PIL as 'a good way to play' with MIXER]. Children found the PIL easy to use (90%), [χ^2 (4, 67) = 93.82, $p < .001$ more children rated PIL as 'easy' to use than expected]; they liked the icons (75%), [χ^2 (4, 64) = 47.88, $p < .001$, more children than expected rated PIL icons as looking 'great'], and found the icons easy to understand (80%) χ^2 (4, 64) = 58.5, $p < .001$, more children than expected rated the PIL icons as 'easy' to understand].

5. EVALUATION

A pilot evaluation using a controlled longitudinal design was carried out to determine whether MIXER meets the desired learning goals. The evaluation involved 91 children aged 9-11 years from four schools in the UK, who completed a pre-test, test and post-test in the classroom over 3-4 weeks. The pre- and post-test workbooks[1], developed using the Transmedia Evaluation approach [16], provide age appropriate engaging activities incorporating three main evaluation measures:

1. **Cultural Intelligence Scale (CQS)** [1]. The behavioural subscale of the CQS was used as a pre and post measure of a child's capability to adapt verbal and nonverbal behaviour in different situations/cultures. This provides data for the question: "Do children who have a more flexible repertoire of behavioural responses in culturally diverse settings recognise more emotion/behaviours in the MIXER application?" This addresses aspects of the behavioural and emotional learning outcomes.
2. **Bryant's Empathy Index** [7]. Factor One from the Bryant Empathy Index was used as a measure of children's empathic behaviour and styles. This provides data for the question: "Are children with higher empathy levels more able to recognise and

[1] All evaluation materials are available at: http://ecute.eu/mixer/evaluation/resources

accept emotions in novel situations?" This in turn addresses the emotional goal of the learning outcomes: "Be able to recognise your emotions when dealing with strange behaviours of another group".

3. **MESSY (Matson Evaluation of Social Skills)** [25]. Factors two and four of the MESSY questionnaire were selected to determine children's capability to adapt to verbal and nonverbal behaviour in different situations/cultures, so as to assess the behavioural goal from the learning outcomes: "Be fully present in attending to others verbal and nonverbal messages".

Children engaged in one of four conditions. A pre-test workbook was completed in the first week. In the subsequent week children interacted with MIXER (or a replacement activity) and in the final week they completed the post-test workbook and engaged in a classroom discussion forum. The four experimental conditions were:

Control Group: MIXER is replaced with an activity that is intentionally unrelated to the learning goals: how to pitch a tent. The replacement activity is of a comparable length to the MIXER software. The purpose of the control condition is to determine whether in the absence of the MIXER software, children's learning and engagement levels remain the same (i.e. no change in learning, attitudes and behaviour without the MIXER software intervention according to the learning goals).

Double Interaction: children interact twice with the software, firstly only interacting with the yellow team and then on the second interaction with both the yellow and red teams. The Double Interaction condition examines whether greater exposure to MIXER reinforces and increases learning, and whether greater familiarity with the characters impacts on engagement levels.

Single Interaction: children interact with MIXER during week two. The aim of the single interaction condition is to assess the impact of MIXER on children's learning and engagement with cultural sensitivity.

Passive Condition: During the test week, children watch a video of the MIXER interaction. This condition is designed to test whether a lack of interaction (i.e. passively watching a video of MIXER) decreases learning outcomes and engagement levels (i.e. Is MIXER more effective if children believe they have some control/involvement over the characters and narrative?) Control through interaction is provided for double and single interaction conditions via the PIL on an iPad as discussed above.

5.1 Results

Children had a mean age of 10.48 (SD: .67). Just over half of the sample were girls (n:49, 53.8%), boys (n:42, 46.2%). Numbers across the four experimental conditions were: Single 15(16.5%); Double 17(18.7%); Control 27(29.7%); Passive 32 (35.2%).

Repeated Analysis of Variance (ANOVA) was carried out on the CQS, Bryant's Empathy Index, and MESSY scales to determine whether there were any significant differences between the four experimental conditions (between-subjects factor) across time (T1 pre-test & T2 post-test) (repeated measures factor). Results provide the main effect of whether the condition is significant (irrespective of time), the main effect of whether time is significant (irrespective of condition), and the combined effect of condition and time together (interaction effect).

The results for the three instruments can be seen in Table 2, where ***Df*** = degrees of freedom, ***F*** = variance of the group means / mean of the within group variances, **η2** = effect size, and ***p*** = how significant the findings were.

Measure	*Df*	*F*	η2	*p*
CQS				
(A) Exp. Condition	3	7.47	.26	<.001
(B) Time	1	7.27	.10	.009
A * B (Interaction)	3	.33	.02	.81
Error (within groups)	64			
MESSY				
(A) Exp. Condition	3	.52	.03	.67
(B) Time	1	8.39	.14	.005
A * B (Interaction)	3	3.62	.17	.02
Error (within groups)	53			
Bryant's Empathy				
(A) Exp. Condition	3	2.44	.11	.07
(B) Time	1	2.18	.03	.15
A * B (Interaction)	3	3.74	.16	.016
Error (within groups)	57			

Table 2: Experimental Condition x Time Factorial Analysis of Variance for CQS, MESSY and Bryant's Empathy

Cultural Intelligence Scale CQS: Higher scores indicate that children adapt and modify their verbal and non-verbal behaviour when they encounter new people and situations. Children's CQS baseline scores (N: 86) ranged from 5 to 24, (*M*:10.67, SD:5.31), and post-test scores ranged from 5 to 25, (*M*:9.51, SD:4.87).

A Repeated ANOVA was carried out on CQS scores at baseline and post-test (repeated factor time) for each of the four experimental conditions (between subjects factor). No interaction effect between time and condition was found, suggesting that children's CQS scores taken together, did not differ between pre- and post-test, and Mixer conditions. Therefore, no support was generated for the hypothesis that the MIXER software would have the most impact on CQS scores in the double and single MIXER conditions.

Bryant's Empathy Index: Scores (N:62) at baseline ranged from 0 to 6 (possible range from 0 to 9) (*M* =2.55, SD:1.46), and at post-test from 0 to 8 (*M*:3.06, SD:1.90). This instrument uses an inverted scale, demonstrating that children were already showing relatively high levels of empathy at pre-test (T1), leaving little room for improvement in scores (i.e. a possible ceiling effect).

Gender & Empathy: A repeated measures ANOVA with time as the repeated factor and gender as the between measures factor was carried out on pre- and post-test Bryant Empathy scores. A significant interaction between time and gender was revealed, $F(1,58) = 4.95$, $p = .03$, ($\eta p2 = .08$). Boy's empathy scores were higher at post-test ($M = 3.18$) compared to pre-test ($M = 2.14$), whilst girl's Bryant empathy scores did not change between pre-test ($M = 2.68$) and post-test ($M = 2.63$). This result suggests that boys became **less** empathic post-test (T2) compared to girls whose scores remained the same. This could mean that boys engaged with the game, whilst the girls did not, but that boys were not engaging with the game in the hypothesized direction.

Bryant's Empathy over time: A repeated measures ANOVA (controlling for gender) with time as the repeated factor and experimental condition as the between measures factor was carried out on pre-test (T1) and post-test (T2) Bryant Empathy scores. A significant interaction effect between time and condition was found (See Table 2). Children's mean empathy scores in the passive mixer condition increased between pre-test

(T1) (M = 2.44) and post-test (T2) (M = 3.69), whilst children's mean empathy scores for the single mixer condition decreased between pre-test (T1) (M = 3.25) and post-test (T2) (M = 1.63). This finding is indicative that there was a **positive** empathic change between pre- and post-test for children who were in the single MIXER interaction condition (i.e. children were more empathic after they had interacted with MIXER in the single condition). The passive MIXER condition appears to have had the lowest impact on empathy scores (i.e. less empathic between pre- and post-test), where no interaction with the MIXER game actually took place.

MESSY: Total scores could range from 15 to 75, with higher scores indicating higher positive social interaction abilities with others (T1: M:58.79, SD:7.25; T2: M:56.31, SD:10.26). Independent t-tests for gender revealed no significant differences in MESSY scores at both pre-test (T1) and post-test (T2).

A repeated measures ANOVA revealed a small significant interaction effect between time and experimental condition (See Table 2). Mean values show us that children's MESSY scores in the double mixer condition (T1: M = 59.9; T2: M = 59.6), control condition (T1: M = 58.24; T2: M = 56.43), and passive condition (T1: M = 56.6; T2: M = 55.87) did not change between pre-test (T1) and post-test (T2). This tells us that the MIXER software in the double interaction did not have desired effect of increasing children's social interaction abilities. However, MESSY scores in the single-mixer condition appear to **decrease** at post-test (T1: M = 60.73; T2: M = 51.09). This is the opposite result to the hypothesized effect of the MIXER software increasing children's social skills.

To summarise, no support was generated for the hypothesis that children's CQS scores would improve following the MIXER software interaction. Some initial support was found for the hypothesis that children's empathic abilities would increase post-mixer interaction. Children's were more empathic at T2 for the single Mixer condition only. Boys were less empathic after interacting with the Mixer software, but this was not the case for girls, whose empathy scores remained unchanged between pre- and post-test. No support was generated for the hypothesis that children's social interaction skills (as measured by the MESSY) would increase following interaction with the Mixer software. In fact, children in the single interaction condition had lower MESSY social interaction scores at post-test compared to all other Mixer conditions.

6. CONCLUSIONS

It is very clear from the pilot results discussed above that MIXER in the form evaluated does not meet its pedagogical objectives and indeed in some ways could be considered to work against them. Given the importance of those objectives and the way we have based MIXER on a strong theoretical foundation, this is a significant outcome that it is very important to explain. Further experiments are required in order to understand the results: here we discuss possible directions.

Empathy is fundamental to an invisible friend approach: it is the way in which the child user is engaged in the scenario and cares – or not – about the outcome. The participants scored at pre-test as highly empathic. At the final conflict in MIXER, the following dialogue occurs:

Tom: "Hey that's not fair. What is going on here?"

Red team: "Calm down! You are dead because no one agrees with you!"

Tom (to user): "They killed me even though I wasn't accused of being the werewolf. And there was no majority vote against me. They are cheating! Maybe they just don't want me to play. What do you think?"

The focus groups, held after the post-test, show that children *did* believe that the red team were cheating. Their comments indicated that rather than MIXER reinforcing the idea that different groups have different ways of doing things, it seemed to have reinforced the view that the in-group is correct and the out-group incorrect. The participants' empathy with Tom may be involved in this result. Changing Tom's utterances is an option here.

The difference in empathy results between boys and girls is easily addressed: we have already introduced Lisa as a female friend character since the similarity principle suggests that girls may find it harder to relate to a male protagonist. However we note that boys' empathy actually dropped, and thus the result may be the girls' empathy drops too.

Though the yellow group play with the usual rules of Werewolves, children did not have prior experience of the game. However the first set of rules is reinforced by greater exposure within MIXER and the variant rules are only perceived once, when the conflict occurs. An obvious experiment is to reverse the conditions so that children play with the red group rules first, and then with the yellow group rules. This would also allow us to determine whether the variant rule chosen is considered inherently unreasonable by children in the given culture.

One should also note that while MIXER applies the basic concept of BARNGA - a negative rule-clash-based conflict - the post-test was carried out *before* group discussion of the outcome. The debrief session is considered a vital part of BARNGA [29], and it may be that this is where the desired learning actually occurs. Running the post-test after the focus groups is also being considered.

As argued above, there is a risk in invoking negative feelings through a conflict situation. While this approach is used successfully with adults, one must consider the possibility that it does not work well with this age group for developmental reasons. This would mean that the learning objectives MIXER is designed to meet were not appropriate for the age group. Certainly the drop in boys' empathy suggests the negative feelings are not resolved. It has been reported [9] that BARNGA works differently with children than it does for adults: "Whereas adult players make compromises to keep the game going, which allows the game to serve as a tool for intercultural communication, the children were looking for the "right" rules, which they normally get from adults." Further, there is evidence [9] of a developmental process for children in relation to rules:

1. Simple individual regularities.
2. Ego-centric imitation of elders.
3. Cooperation.
4. Concern for the rules themselves.

Thus testing participants with an additional instrument to try to establish their position on this sequence could tell us whether the theory on which MIXER is based is problematic for the age group.

We therefore suggest that as well as contributing novel technology in MIXER, the outcome of its initial evaluation is a valuable contribution to overcoming the real difficulties in developing novel approaches to education in cultural sensitivity.

7. ACKNOWLEDGMENTS

This work was partially supported by European Community (EC) and is was funded by project ECUTE (ICT-5-4.2257666). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, which is not responsible for any use that might be made of data appearing therein.

8. REFERENCES

[1] Ang, S., Van Dyne, L., Koh, C., Ng, K. Y., Templer, K. J., Tay, C., & Chandrasekar, N. A. (2007). Cultural Intelligence: Its Measurement and Effects on Cultural Judgment and Decision Making, Cultural Adaptation and Task Performance. Management and Org. Review, 3(3), 335–371.

[2] Aylett, R. S., Louchart, S., Dias, J., Paiva, A., & Vala, M. (2005). FearNot!–an experiment in emergent narrative. In *Intelligent Virtual Agents* (pp. 305-316). Springer

[3] Aylett, R., Vannini, N., Andre, E., Paiva, A., Enz, S., Hall, L.: (2009) But that was in another country: agents and intercultural empathy. In: Proc. 8th Int. Conf. Autonomous Agents and Multiagent Systems, IFAAMAS (2009) 329–336

[4] Baron-Cohen, S.: Mindblindness: An Essay on Autism and Theory of Mind. MIT (1995)

[5] Bennett, M.J.: A developmental approach to training for intercultural sensitivity.International journal of intercultural relations 10(2) (1986) 179–196

[6] Brislin, R.W., Yoshida, T.: Improving intercultural interactions: Modules for cross-cultural training programs. SAGE Publications, Incorporated (1993)

[7] Bryant, B. B. K. (1982). An index of empathy for children and adolescents. Child Development, 53(2), 413–425.

[8] Byrne, D.E.: The attraction paradigm. Personality and psychopathology, Vol 11. Academic Press (1971)

[9] Corbeil, P. (1999) Learning from the Children; Practical and Theoretical Reflections on Playing and Learning. Simulation gaming, 30:163

[10] Crawford, C. (2008). Deikto: an application of the weak sapir-whorf hypothesis. Proc. Hypertext 2008 workshop Creating '08. ACM

[11] Dias J. and Paiva A. (2005) Feeling and reasoning: a computational model for emotional agents. In Proceedings of 12th Portuguese Conference on Artificial Intelligence, EPIA 2005, pages 127–140. Springer, 2005

[12] Dias, J; Aylett,R; Paiva,A. & Reis, H. (2013) The Great Deceivers: Virtual Agents and Believable Lies, Proc CogSci2013, pp2189-2194

[13] eCute: http://ecute.eu Accessed 2/10/13

[14] Endrass, B., Rehm, M., Andre, E.: Planning small talk behavior with cultural influences for multiagent systems. 25(2) (2011) 158–174.

[15] Hall, L; Woods, S; Aylett, R. & Paiva, A. (2006) Using Theory of Mind methods to investigate empathic engagement with synthetic characters." International Journal of Humanoid Robotics 3.03 (2006): 351-370.

[16] Hall, L; Jones, S.J; Aylett,R; Hall, M; Tazzyman,S; Paiva, A. & Humphries, L. (2013) Serious game evaluation as a meta-game, Interactive Technology and Smart Education, Vol. 10 -2, pp.130 - 146

[17] Harbers, K. v. d., Maaike; Bosch, & Meyer, J.-J. (2009). Modeling agents with a theory of mind. In Proc. 2009 ieee/wic/acm Int. Joint Conf. Web Intelligence and Intelligent Agent Technology - v. 02 pp. 217–224.

[18] Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and organizations. McGraw- Hill London (1991)

[19] Hofstede, G.J.: The moral circle in intercultural competence: Trust across cultures. (2009) 85–99

[20] Iacobelli, F., Cassell, J.: Ethnic identity and engagement in embodied conversational agents. In: Intelligent Virtual Agents (IVA), Springer (2007) 56–63

[21] Johnson, W., Valente. Vintage., A.: Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. (IAAI 2008), AAAI (2008) 100–120

[22] Kanfer, F.H.: Self-management strategies and tactics. (1979) 195–224

[23] Mascarenhas, S., Dias, J., Afonso, N., Enz, S., Paiva, A.: Using rituals to express cultural differences in synthetic characters. In: Proceedings of AAMAS 2009. (2009)

[24] Marsella, S., Johnson, B., LaBore, C.: Interactive pedagogical drama. 4th Int. Conf. on Autonomous Agents (AAMAS), Bologna, Italy, ACM Press (2002) 301–308

[25] Matson, J. L., Neal, D., Fodstad, J. C., Hess, J. A, Mahan, S., & Rivet, T. T. (2010). Reliability and validity of the Matson Evaluation of Social Skills with Youngsters. Behavior modification, 34(6), 539–58.

[26] Said, E.W. (1994) Orientalism

[27] Shirts, R.G.: Beyond ethnocentrism: Promoting cross-cultural understanding with bafa bafa. Intercultural Sourcebook: Cross-Cultural Training Method 1 (1995) 93–100

[28] Singer, P.: The expanding circle. Clarendon Press Oxford (1981)

[29] Thiagarajan, R. (2011). Barnga. Faber & Faber.

[30] Ting-Toomey, S.: Intercultural conflict competence as a facet of intercultural competence development: Multiple conceptual approaches. (2009) 100–120

[31] UNITY: http://www.unity3d.com/

[32] Vala, M., Raimundo, G., Sequeira, P., Cuba, P., Prada, R., Martinho, C., , Paiva, A.: ION framework - a simulation environment for world with virtual agents. Proc IVA 2009 LNCS Vol 5773. pp 418–424, Springer

[33] Wu, P., Miller, C.: Interactive phrasebook conveying culture through etiquette. In Blanchard, E., Johnson, W., Ogan, A., Allard, D., eds.: 3rd Int. Workshop on Culturally-Aware Tutoring Systems (CATS2010), Pittsburgh, (2010) pp47–55