FP7- 257666-eCute

(March, 2012)

Education in Cultural Understanding, Technology Enhanced

Collaborative Project (ICT-2009.4.2)
Technology-enhanced learning

Start date of project: **01/03/2006** Duration: **36 months**

Deliverable 7.2: Assessment and Evaluation Embedding Strategy and Implementation Approach

Due date of deliverable: 30.03.12 Actual submission date: 30.03.12

AUTHORS: Marc Hall, John Hodgson, Colette Hume, Lynne Hall.

CHECKERS: Nick Degens & Eva Krumhuber.

STATUS: [FINAL]

Project co-funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

PROJECT COORDINATOR

Name: Ruth Aylett
Address: School of Mathematics and Computer Science, Heriot-Watt University Edinburgh, EH14 4AS
Phone Number: +44 131 4514189 Fax Number: +44 131 451 3327
E-mail: ruth@macs.hw.ac.uk

TABLE OF CONTENTS

1 Purpose of Document:

The purpose of this document is to describe the embedding strategy of the evaluation activities of the eCUTE project.

To be read in conjunction with:

D7.1 – Evaluation (a web report available at: www.ecute.eu/aaei)

2 Executive Overview:

The purpose of this document is to describe the current state of the eCUTE evaluation strategy with specific attention given to how and to what extent the evaluation will be embedded into the proposed interaction. Section 3.2 establishes the context of the evaluation effort and disambiguates some necessary terminology and concepts. Section 3.3 outlines the battery of instruments we propose for use with both MIXER and TRAVELLER, describing them in general terms. Section 3.4 and 3.5 describes the instruments to be used with MIXER and TRAVELLER respectively, how we will embed them and details about why they have been chosen for that specific interaction. The Appendix section is used mostly to feature examples of the instruments themselves (where they are freely available). Much of the technical detail has been moved to the Technical Annex, here you will find discussions of theoretical background and considerations, and some detailed definitions of relevant ideas that while not critical to understanding this document, are relevant.

3 Introduction

The evaluation efforts in the eCUTE project need to assess and measure several factors. We organise those factors into four themes (detailed in 4.1). eCUTE aims to perform all evaluation activities in such a way that there is no apparent burden on the user, or to put in terms of user experience, so that the user is unaware of the evaluation activities. This is achieved by a process of 'embedding' the evaluation into the *primary interaction* or by presenting the instrument in a more appealing manner. A more detailed discussion of what we mean when we talk about embedding and the primary interaction can be found in the technical annex IV and in the following section respectively. Embedding evaluation is achieved by embedding instruments; we see an evaluation strategy as a battery of such instruments and an instrument as any activity designed to elicit a response to a specific query. In order to make this document easy to navigate and refer to we have split the presentation of the evaluation instruments into three broad sections, the first describing the instruments in their abstract form, the second and third describing them in the context of a MIXER and TRAVELLER respectively.

3.1 Usage and Terminology

In order to discuss our evaluation strategy effectively it is necessary to define some key concepts, if only for the sake of effective communication. Firstly, refer to technical note IV for an exposition of what we mean when we talk about embedding evaluation. We make a distinction between *scenario* and *interaction*. The interaction refers to the user's interaction with the artefact (in this case a virtual world), not to be confused with a more technical idea of the interaction as something limited to the interaction devices, although the two uses of the word overlap. The scenario, which is delivered within the interaction, contains such elements as plot, character development, story and (implicitly) the learning goals.

A concept that is central to our view of evaluation is the 'Magic Circle'our usage is similar to Salen & Zimmerman, (2004) in that processes found in the ordinary and temporary worlds are both taken in to account This represents a conceptual world of which the user should feel they are a part of. The use of the word magic denotes its nebulous and almost hypothetical nature, in which the original concept is open to debate. It occurs due to the mental efforts of the user rather than being a construct presented to the user by the designers. This is why we use the idea of a Magic Circle as opposed to a 'virtual world'. A virtual world comprises of elements presented to the user, while a Magic Circle is much more a fiction created by the user based on stimuli presented to the user by the designers.

Often, throughout this document, the phrases 'primary interaction' and 'primary scenario' are used. This is because often the process of embedding evaluation instruments into the Magic Circle augments or extends the content of the Magic Circle. For example, an evaluation might introduce a new character. This character would be part of the interaction, but not part of the primary interaction. The 'primary interaction' and the 'primary scenario' are used to refer to the interaction and scenario as they stand without embedded evaluation.

From an evaluation perspective the terms 'participant' and 'user' are almost synonyms. All users of the software are participants in the evaluation study and in most cases all participants are also users in some sense (with the exception of individuals who complete evaluation instruments without being exposed to the artefact.) We do however use these two words

selectively given the context. When we are talking about the individual in the context of the evaluation study then s/he is referred to as a participant, when in the context of using the artefact, s/he is referred to as a user.
Finding an effective structure for this document had to take into account that there are some instruments used for both the MIXER and TRAVELLER interaction and some that are specific to each interaction. It's also possible that instruments used for both are used in a slightly different way. Further, some instruments are used to fulfil more than one data gathering requirement, i.e. to get data on multiple aspects of the interaction. To cope with this we define the instruments abstractly; then describe a set of 'alias' instruments for each of the two interactions. An alias is a version of an instrument that has been adapted for implementation in the context of a specific interaction. For example, a demographic questionnaire is an abstract instrument, while a version designed to look like an in-role application form for some fictitious job feature in the scenario is an alias of the demographic instrument and is only applicable to that specific interaction. This structure means that should a reader want to know about how the instruments are used for a specific interaction, they need only refer to one section. The downside to this structure is it incurs some repetition, e.g. where the same instrument is described in two situations. We have tried to balance repetition with structural clarity.

3.2 Considerations

The evaluations of both MIXER and TRAVELER will initially be conducted in the United Kingdom, followed by a comparison of another partner country, Germany and possibly Portugal.
The evaluation requirements fall into four themes, each representing a required area of investigation, understanding or data as indicated by the project team. Some instruments are more suited to specific themes, some cross multiple themes. Some instruments can be used for both the adult and child interactions, many are specific to an age group. The themes are (in no special order):

- **Social** - Contains cultural and intercultural competencies, in-group-out-group dynamics and the learning outcomes or 'impacts'. (More information can be found in deliverables D2.1 Preliminary Cultural Learning Interdisciplinary Network and D3.1 Behavioural Interface)
- **Agents** - All aspects of agent believability and effectiveness, both in terms of presentation and mind architecture.
- **Engagement** - The level of engagement experienced by the user towards the interaction and scenario.
- **Comprehension** - The level of the user's understanding of the events and progression of the scenario and interaction.

3.3 Instruments

The following section provides a quick reference descriptor of the instruments proposed as part of the eCUTE evaluation strategy. Examples of the instruments themselves can be found in the Appendices. This section doesn't deal with embedding, for information about how these will be embedded; refer to section 3.4 and 3.5.

3.3.1 INCA for Adults

Overview

The Intercultural Competence Assessment (INCA) was designed as a means to gain information about an individual's background experience with regard to (inter)cultural interaction. It is comprised of three parts, questionnaires, scenarios and role-play (www.incaproject.org). We intend to make specific use of the questionnaire components.

Method

Participants are asked questions about how they act in specific situations, each with a cultural component. In some cases the questions are worded in such a way as to refer to past experience and theoretical experience. For example, "When meeting new people I…" invites the participant to call upon past experience, but also to imagine that they were in that situation at present, and report how they *would* act.

The first section of the questionnaire features questions that elicit a three-point likert response (not applicable, maybe, fully applicable) while the second section is mostly free text responses. The main embedding of the questionnaire is the way in which it will be presented to the participant.

Application

The INCA will be used as a means to establish a baseline measure of cultural competence, with it being applied before and after the interaction with the application.

3.3.2 Social Connectedness

Overview

The Social Connectedness instrument was validated by (Yoon, Jung, Lee, & Felix-Mora, 2012) and is comprised of two component instruments: Social Connectedness in Mainstream Society (SCMS) and Social Connectedness in Ethnic Community (SCEC).

Method

The instrument is a set of validated questions that ask the participant about their perception of their relation to other cultures based upon items such as behaviour, knowledge and empathy.

Application

The instrument has obvious application to the social theme of our evaluation, but more specifically the nature of the wording of the questions lend it to use with adults rather than children. The questions must be adapted slightly so that they are relevant and meaningful to the participants (e.g. featuring cultures that the participant is aware of).

3.3.3 Empathy Index for Children

Overview

The empathy index for children and adolescents, developed by Bryant, (1982) and initially discussed in the Milestone 3 report produced by Jacobs University (MS3), looks at empathic traits in people. It examines the relationship of aggressiveness and attitudes towards the general population. Depending on the age group the scales can be a simple yes or no answer, or up to a nine-point scale, which identifies between 'like me' or 'not like me'.

Method
As the child participants will derive from a mixed ability group of 9 to 11 year olds we will apply the simple scale of yes or no, this will take in to account the varying literacy levels of the participants.
Application
This will establish a baseline measure of the participants' interpersonal competencies.

3.3.4 Remote Inquisitor

Overview
The Remote Inquisitor (RI) is a character (either virtual or human) with whom the user can communicate, in parallel to the primary interaction.
Method
While taking part in an interaction the user has access to some communication device with which they can chat with an 'online friend' character. This character is a member of the interaction world and is tasked with eliciting responses from the user while presenting itself to the user as an online friend.
Application
The RI can be used to elicit either Social or Comprehension responses. The use of the RI as a means to measure engagement is problematic as, firstly, the interaction with the Inquisitor punctuates the primary interaction and, secondly, the RI character could be seen as part of the wider interaction of which the engagement levels are being measured.

3.3.5 Character Evaluation Questionnaire

Overview
The Character Evaluation Questionnaire (CEQ) uses direct questions to ascertain user perception of characters in the scenario, mainly with regards to believability. An initial character study has been carried and reported in D2.1, in which characters acted out simple greeting gestures. The CEQ aims to evaluate the characters believability within the context of the applications being developed. This evaluation within context will allow for the agent architecture to be taken in to account and the characters behaviour.
Method
After interacting with the application, the user completes the questionnaire.
Application
Although the CEQ targets 'characters' rather than agents, it is the primary insight into user's perceptions of the autonomous agents. The characters are also a factor of engagement so the CEQ can also be used to provide limited-scope engagement measurements i.e. the component of engagement impacted upon by character behaviour.

3.3.6 Motivation Analysis

Overview
An instrument designed to find a user's perception of the motivational state of characters in a scenario

Method
A small scenario, or scene from a scenario, is presented to the user. This can be text or some other media, like video or audio. The scenario is designed such that a user, on the basis of the characters internal motivational state, can explain its proceedings. The results are analysed by categorising the users' impression of the characters' motivations; for example, whether or not the explanation ascribes an internal belief state.
Application
Given that an overarching aim for autonomous agent design is that users perceive them as having a mind of their own, an analysis of the users' ascription of mental states can be used to determine the extent to which users attribute a mind to the agents.

3.3.7 Real-time engagement state

Overview
This instrument takes snapshots of the user's state of engagement during an interaction in such a way as to minimize distraction from the primary interaction. The key strength of this method is that avoids retrospective bias caused by recall during post-hoc self-reporting (Killgore, 1998).
Method
The user is given a touchscreen interaction device. They can touch the device at any time during the interaction. The screen of the device features either a small number of buttons, each representing a level of engagement or discrete engagement state, or an x-y space representing two orthogonal variables (e.g. 'gloating' versus 'pity'. Touching the device creates a time stamped data point holding the value of whatever the user pressed.
Application
This instrument is intended to generate a real-time chart of the users' level of engagement.

3.3.8 Flashcard Quiz

Overview
Flashcard Quiz provides users with a way to answer a set of questions in a way that is engaging. It is based on an interaction paradigm that many users will be familiar with, and is intuitive if even they aren't. It is a paradigm used very effectively in commercial group/party quiz games.
Method
This instrument can be applied to the whole group at once, or as a single-player addition to the primary interaction. Participants are asked a series of questions and respond by presenting what they feel is the appropriate card.
Application
The instrument is highly suited to factual content and is most applicable to the comprehension evaluation theme.

3.3.9 Word Association

Overview

The Word Association instrument is based on Directed Activities Related to Text (DARTs) developed by Lunzer & Gardner, (1979). In this text not only relates to the written word, but also diagrammatic representations or pictures. The activity is designed to be fun by allowing the participant to think aloud and exchange ideas without the fear of being wrong DfESC (2004 a, 2004 b)

Method

Users are asked to pick from a set of words, those words that are most strongly associated with some topic, concept or character.

Application

Depending on the specification of the words available to the participant their choice of words can be used to identify the factors that are most salient to a user's outlook, e.g. the use of emotionally charged words as apposed to emotionally neutral words, or the tendency towards use of words associated with familial relations.

3.3.10 Card Sort: Character Grouping

Overview

Card sort methods are used widely in elicitation practices (Wood & Wood, 2008), generally involving the arrangement, grouping or ordering of concepts. The important component of card sorting tasks, with respect to this project, is the process by which the participant categorises concepts.

Method

Users sort characters into groups, in such a way as to minimise conflict, taking into account the traits of the characters.

Application

If the traits are specified based on the pedagogical learning outcomes, for example cultural dynamics, the way the participant groups the characters would differ depending on their attainment of the learning outcomes.

3.3.11 Card Sort: Concept Grouping

Overview

Much of the cultural theory employed for MIXER and TRAVELLER makes use of the concept of the Moral Circle; which, in a fundamental sense, are mental sets of people. This type of categorical analysis is what the Conceptual Grouping instrument seeks to investigate.

Method

The instrument is split into three phases: a proximity phase, a grouping phase and a descriptive phase. In the first two phases the user works with a set of concepts presented as 'cards' (either virtual or physical). In the proximity phase the user is asked to arrange the cards around themselves such that they place more important concepts closer to themselves and less important ones further away. In the second phase the users are asked to arrange the cards into groups. The scheme by which they group the cards is their choice. In the final, descriptive stage, the users are asked to explain why they grouped the concepts as they did.

Application
With the correct choice of concepts concept grouping can be used to gain insight into the users mental organization of concepts with respect to culture. This instrument also deals with how the user is thinking at the time, rather than based on long-term measures like in many cultural sensitivity instruments, so could offer a fine temporal resolution.

3.3.12 Attainment levels

Overview
In order to measure the level of comprehension attained by child users interacting with the MIXER scenario it is useful, from an analytical perspective, to have a baseline measure of the users' comprehension skills. This baseline measure already exists, at least in the UK, in the form of a teacher assessment as prescribed Attainment Levels. The users' schools will have already assessed the pupils' comprehension skills and encoded the results in an attainment level for each pupil.
Without a correction for the users' comprehension skills measuring the level of comprehension of the interaction scenario could give misleading results; for example, a very uniform real level of comprehension would appear as a wide distribution of comprehension if the baseline level of comprehension also has a wide distribution, such a result wouldn't distinguish between a mixed result or a mixed set of base ability.
Method
The Attainment levels will be requested from the teacher.
Application
As mentioned, attainment levels are available as a base-line measure for childrens' comprehension levels in the MIXER scenario.

3.3.13 Progression

Overview
Progression is presented here as a discrete instrument but is really an activity metric. It exploits the fact that the use of and progression through TRAVELLER require comprehension of its content, thus the extent to which the user progresses through TRAVELLER and the nature of any obstacles to that progression are a basis of analysis of comprehension.
Method
The initial measurement for this instrument is the point in the scenario the user reaches during the interaction. The next step is to establish why the user got as far as they did. There are two broad categories of obstacles: 'the user didn't understand what they had to do' and 'the user knew what to do but not how to achieve it.' Further, every stage in the interaction maps to a set of knowledge or understanding requirements for getting to that stage. The progression of a user can therefor be mapped back to a particular knowledge outcome.
Application
This instrument is highly suited to TRAVELLER, but less so to MIXER. At this stage it is less clear what role comprehension plays in the progression through the MIXER

interaction.

3.3.14 Physiological Measures

Overview

The Physiological measures are provided by, and represent a distinct body of work employing specialized hardware and expertise form, Jacobs. The highly technical nature of their implementation means that their application is separated logistically and independent from the rest of the evaluation methods and instruments A full description of theses methods can be found in MS3 report, however these measures still need to be shown as part of the evaluation strategy and so is integrated into this document as a discrete instrument.

Method

The primary physiological measures are of arousal via skin conductance and heart rate and emotional valence by muscle activation (e.g. corrugator supercillii muscle in frowning, zygomaticus major muscle in smiling.)

Application

The psychological variable being measured by the physiological instruments is a high-level, and involuntary, reaction to emotional relevance. Emotionally salient events trigger a measurable 'spike' that the equipment can measure and record. This variable can be used as a measure of how others fit into the user's unconscious value judgments, which in turn maps onto in-group-out-group dynamics. Those individuals, events or propositions that elicit a strong emotional response are more likely to represent socially and culturally relevant phenomena and as such comprise the 'in-group'.

Physiological measures can also be used to measure general levels of engagement to a situation by taking arousal as an antilog of engagement.

3.3.15 PQ – Presence Questionnaire

Overview

Does the user understand their role within the interaction?

The users acceptance and adoption of a role is related to the concept of social identity theory in which an individuals perception of self or self concept is partly derived from membership of a relevant social group and setting (Turner & Oakes, 1986). In order to interact and engage in the storyline of a virtual environment the user needs an understanding of their position or role (Oatley, 1995) within the depicted social group, giving a foothold for the user in the environment in which the story is set.

The Presence Questionnaire (PQ), was designed by (Witmer & Singer, 1998) to measure the users presence in virtual environments and the degree to which contributing factors influence the intensity of the experience of presence. These factors are shown in the appendix.

Method

The PQ questionnaire relies exclusively on self-report information. A seven-point scale format that is based on the semantic differential principle (Babbitt, 1989). The PQ contains 32-element questionnaire (in appendix), which is administered post use.

Application

The PQ is suitable for use in applications in which the user moves around a virtual environment. The TRAVELLER application will be set in a bar various other geographical locations, it is therefore essential that the user maintains a sense of presence in the virtual environments in order to follow the storyline of the application as it develops. Studies involving the PQ show that immersion and felt involvement are necessary for experiencing presence and that the levels reported of both are relevant to the level of presence experienced. Immersion (Jennett et al., 2008) and involvement (O'Brien & Toms, 2010) are both common features in the assessment of user engagement, therefore the assessment of presence is an indicator of user engagement.

3.3.16 Games Engagement Questionnaire

Overview

The Games Engagement Questionnaire (GEQ) is a 19 question survey. The questions are intended to measure the subjective experience of deep engagement. Brockmyer et al., (2009) defined deep engagement as involving immersion, presence, flow and absorption and developed the GEQ to evaluate the impact of playing video games. The studies conducted in development of the questionnaire focused on violent video games, however the range of questions included in the GEQ relate to levels of engagement in the game and not specifically the amounts of violent content.

Method

Participants complete a three-part questionnaire on media habits and related issues (the GEQ is one of the three parts). Participants provide information relating to their age when they started playing video games, how much they play each week and three favourite games.

3.3.17 Engagement Sampling Questionnaire

Overview

The Engagement Sampling Questionnaire (ESQ) has been developed for the purpose of discovering the users desire to continue use of interactive narrative based application. It pays specific attention to which elements of an interactive storytelling application make the user want to continue the experience, and to which degree he or she wants to continue (Schoenau-Fog, 2011).

Method

The ESQ relies on self-report and during runtime surveys.

Application

During use of an application the user is asked if they want to continue (by a pop-up in the application) and asked to state why, at that point, they want to continue. Depending on the timing of the pop-up the user will give different responses. If the user is enjoying using the application or enjoying the storyline and wants to know what will happen next then the answers given will indicate the reasons. This method would be applicable to both MIXER and TRAVELLER, however, a major issue with the ESQ is if a user chooses not to continue, then they would not reach the end of the

application and would not experience the system in full.

3.4 Embedded Evaluation: MIXER

The previous section detailed the suit of proposed instruments in their abstract form. A subset of the instruments described will be used with each of the interactions. This section presents the instruments that are relevant to the MIXER interaction, how they meet the data gathering requirements for their specific theme and how they will be embedded. For clarity, this section is organized by evaluation theme.

3.4.1 Social

MIXER Remote Inquisitor

Overview

The Remote Inquisitor (RI) is in a sense a role that must be enacted, either by some technological solution or by a human, during the primary interaction. The actions of the role are specified by a procedure specifically designed for the interaction and data gathering requirements; the RI needs to know what to inquire about and when.

Embedding

In MIXER the IR represents a remote friend of the user with whom the user keeps in contact while taking part in the scenario. The remote friend is curious to know what the user is doing at all times and routinely asks for descriptions of and explanations for the events in the scenario.

Procedure

During the interaction the user has access to a communication device through which they can chat to the fictitious remote character. Interaction with the RI and MIXER scenario happens in parallel. The remote agent, whether human or software, abides by a protocol designed to elicit evaluation relevant responses.

A signalling protocol is to be used so that the RI can be made aware of events in the main interaction and can pause the main interaction, if necessary, to make sure required responses are elicited before the world-state progresses e.g. to allow the user to complete the current RI conversation before progressing.

Validity

The key strength of this instrument is that it is flexible and generic. The nature of the data it gathers depends on the nature of the protocol it implements.

The RI could in principal ask the user about anything, but since the RI is in effect a character in the interaction world (albeit part of the ‘extended world’ created by the evaluation), it would be incongruent for the RI to inquire about things that the user would consider beyond the scope of the RI’s knowledge. As a result the RI cannot take the position of an all-knowing helper / elicitor, rather the RI takes the role of a largely ignorant remote agent who is interested in the user’s pursuits.

Card Sort: Concept Grouping

Overview

Users are asked to sort concepts into groups of their choosing.

Embedding

Although it would be possible to construe the Card Sort task into a task that is integral to the primary interaction, to do so would create an unwanted bias to the results. The task is intentionally unfocused, in the sense that the user isn't asked to perform the sorting for any pragmatic reason, so as to give insight into how the user thinks by default rather than when steered towards a specific goal. For this reason this type of Card Sort activity could be best used as a baseline pre-test so that it doesn't interrupt the primary interaction.

Procedure

Users are given a set of cards each representing a concept, and one representing 'me'. They are asked to arrange the concepts around 'me' such that more important concepts are closer to the 'me' card and less important concepts are further away. This arrangement is then captured.

The users are then asked to organise the same cards into any number of sets. They choose how many sets to create and what the sets represent. The users are asked to name each of the sets by labelling them. This arrangement is captured using a camera.

Analysis

The insight generated by this instrument is strongly related to which concepts the cards are chosen to represent. By using concepts relevant to the primary scenario, each with varying cultural significance (based on categorisations like heroes, rituals, values etc. (Hofstede, 1994), as well as culturally neutral concepts), the users' arrangements will engender their internal model of the relationships of those concepts. This could be evaluated from the perspective formation and perception of 'Moral Circles'.

Cluster analysis can be used to derive the variables most relevant to the users' groupings.

Validity

As mentioned, this instrument is based on the use of 'Moral Circle' sets (Mc Breen, Di Tosto, Dignum, & Hofstede, 2011). The aim is to gain insight into the user's internal understanding or implementation of the type of thinking implied by Moral Circle theory. As an example, a user could group culturally related ideas, implying that cultural dynamics are a salient variable in their mental representations of others.

Card Sort: Character Grouping

Overview

Participants use their knowledge of group relational dynamics to figure out how an arbitrary set of characters would behave in different group structures. The characters are presented along with traits that, while representing the character's personality traits, also in some cases indirectly reference cultural traits (e.g. A character likes 'saving money' indirectly references uncertainty avoidance.) The same set of traits reference personality level traits such that characters could share implicit cultural propensities while having apparently different explicit traits, and vice versa (e.g. two

characters might both like a particular specific activity, e.g. bowling, while having cultural traits that are at odds, like being individualist versus collectivist.) A participant that doesn't pick up on the cultural component to the traits will assume the two would be highly compatible, while someone aware of the implications of their cultural differences would be more likely to consider them a bad match.

Embedding

This is presented as an in-role task in which the characters must be arranged in order to provide the most positive experience for those characters, for example, allocating the characters to a set of dormitories.

Procedure

The users are given a set of character cards. Each has a picture of the character and a set of traits. The participant must then group these cards in such a way as to minimize conflict and maximize friendship. The participants are asked to describe why they grouped the characters as they did.

Analysis

Factor analysis can be used to derive which factors are most salient to the participants grouping of the characters. The traits must be designed in such a way that they map to culture and personality level parameters, for example it must be established that a specific trait represents a certain cultural trait and personality trait, and how these relate to the other character's traits. The instrument can also be used with participants who haven't taken part in the primary interaction as a means to measure the impact of the scenario on the users impression of the group dynamics of the characters.

Validity

The instrument is designed to find those factors, whether cultural-centric or personality-centric, which impact upon the participants' assessment of group dynamics. As cultural traits aren't directly referenced users must have some tactic knowledge of them in order to detect and apply them.

3.4.2 Agents

MIXER Character Evaluation Questionnaire

Overview

The Character Evaluation Questionnaire (CEQ) measures directly user perception of the effectiveness and realism of the characters in order to evaluate the scenario and agent architecture.

Embedding

There are problems associated with embedding this type of evaluation component, which are discussed in Technical Note II. Generally the questions have to be reworded so as to avoid them breaching the Magic Circle.

Procedure

The user is asked to suppose that some of the characters are actually robot imposters. They then rate the likelihood, on a five-point scale, that each of the characters is in fact robot imposter.

Analysis

As the data is inherently quantitative standard statistical analysis can be used.

Validity
The measure of the users belief that the character could be a robot imposter is analogous to their impression that the character's behaviour is in some intuitive sense un-human or incongruent with the proper behaviour for normal human interaction, thus is an indirect measure of the 'realism' of the characters. The questionnaire would incorporate those questions used within the initial character evaluation study as shown in D2.1, looking at the characters ability to have a mind or basic emotions.

Empathy Index
Overview
A series of validated questions measure an individual's level of empathy towards others.
Embedding
This instrument will be modified to match the visual style of the interaction.
Procedure
The users will be given the instrument at the beginning of the interaction.
Validity
The user's empathic response to the agents is to be measured but must be corrected for each user's base empathy level. This instrument will be used to get the base-line measure.

Motivational Analysis
Overview
Asking users to explain the behaviour of individuals in terms of their motivation can be used to examine their perception of the character's internal mental characteristics.
Embedding
As the instrument is scenario based, it can be constructed from themes introduced in the primary scenario while featuring specific scenarios that are new, or spin-off scenarios.
Procedure
1. A discrete scenario is presented to the user, either a passage of text or a video.
2. The user is asked to describe what they believe to be the motivation of the characters in the scenario.

Analysis
The data gathered is of a qualitative form but can be coded by categorizing responses as either characterizing an internal mental state (therefore having required a Theory of Mind Method of analysis by the participant) or not. For example, the motivation of a character could be explained in terms of them accusing a player within the game in terms of a physical attribution, such as "hearing a noise" or describing a players non-verbal channels of communication and ascribing a mental state of deception to those actions. Both of which imply attribution of an internal behavioural model capable of the perception of someone being guilty, but the later implies a higher level of Theory of Mind abstraction. More precisely a user could only bestow the ability to 'be deceptive' on an agent that has at least some form of mind.
Validity

The Belief Desire Intention model has been used successfully to categorize Theory of Mind (ToM) responses (Bosse, Memon, & Treur, 2011). The engagement of ToM in the user corresponds intrinsically to attribution of the sorts of internal mental states that the agents are designed to engender.

3.4.3 Engagement

Real-time Engagement State

Overview

The participant uses a touchscreen application to record their state of engagement at any point during the interaction.

Embedding

This instrument is presented in parallel to the primary interaction, as a separate discrete task.

Procedure

A touchscreen device is present with the user at all times. An application featuring an x-y grid runs on the device and is available throughout the interaction. At any time the user touches a part of the x-y space to encode their engagement state.

Analysis

The motivation for such a real-time approach to engagement measuring is discussed in detail in Section V of the Technical Annex. The data captured by the instrument represents a direct measure of the users engagement state over time.

Validity

This instrument represents a trade-off between data reliability and bias, in that in order for the data captured to be accurate the instrument must be done in parallel with the primary interaction, but by doing so the instrument itself is part (albeit in parallel) of the interaction and as such represents a possible bias to the data it is capturing. A simple example is that the instrument might be a distraction from the primary interaction. However, this measurement paradox is in a fundamental sense true of all measurement activities.

3.4.4 Comprehension

Attainment Levels

Overview

Attainment levels are derived from student academic records by their school and kept on record, they will provide a baseline measurement of the users ability to interpret the scenario introduced to them.

Embedding

Acquisition of attainment levels will happen without participants' knowledge and so requires no embedding.

Procedure

Specific consent will be sort to gain access to participants' attainment levels, both from the school and from users' legal guardians.

Analysis

Attainment levels can be used as a baseline measure to adjust the results of comprehension evaluation to correct for varying levels of comprehension skills among

participants.

Validity

Attainment levels take into account a wide range of competencies, so represents a broad measure of comprehension skills, including literary, social, personal etc. A good level of comprehension of the activities in the primary interaction requires comprehension skills covering a range of areas, as well as general literacy abilities. Overall attainment level seems to be a reliable measure of a user's abilities with respect to the interaction.

Flashcard Quiz

Overview

Flashcard Quiz is an effective and engaging way of allowing a group to collectively take part in a question and answer exercise.

Embedding

It is difficult to envisage how to embed this instrument into the primary scenario as it requires the existence of a quiz-master character who, in principal, knows everything about the scenario (since, as an evaluation instrument, this needs to be able to ask about an aspect of the scenario.) This is discussed in detail in Technical Note I, but to summarise the solution is to conjure an *all-seeing alien* character who has observed everything that has happened but has questions. This character is, by its nature, separate from the primary scenario and so can ask questions about it without interfering with it.

Procedure

The instrument can be conducted as a group activity or as a single user activity. In the former case, a quizmaster asks the whole group a question. Each user picks the correct answer by pressing one of four buttons. In the latter case the quiz is presented as part of the application.

Analysis

As responses all relate to objective facts about the scenario they can be marked and tallied, giving an overall level of comprehension. Further, scores are paired back to individuals and so correlations between comprehension of different aspects of the interaction can be examined.

Validity

Although it is possible for users to copy the responses of other participants, the anonymous nature of the quiz makes it more likely to give rise to meaningful results. An important consideration is that the questions cover a good sample of the content of the scenario, so a large volume of questions is preferable. The Flashcard Quiz allows this volume as it can be performed in a rapid-fire fashion.

3.5 Embedded Evaluation: TRAVELLER

3.5.1 Social

INCA for Adults

Overview

The Intercultural Competence Assessment (INCA) has been developed to assess the users interactions with people in other cultures, along with their general acceptance of cultural difference. It is comprised of three parts, questionnaires, scenarios and role-play that have been made age appropriate.

Embedding

The questionnaires will be provided in a form that is visually engaging for the users.

Procedure

The user will be given the questionnaires before interacting with the application.

Analysis

The key strength of this instrument is that it is flexible and generic. The nature of the data it gathers depends on the nature of the protocol it implements. This is important in the context of TRAVELLER as the details of the interaction are a matter of on-going development.

Validity

In using this assessment portfolio, a user's perception of different social events before and after an interaction can be measured. This includes those found within their own culture and that of another.

Social Connectedness

Overview

A question based instrument that measures participants' perception of their relation to the social and cultural groups they live within.

Embedding

This is a difficult instrument to fully embed as it requires the participant to answer as themselves (rather than while playing a role from the primary interaction) and features content and themes specific to it's own purpose that can't be easily replaced with content congruent with the primary scenario. The instrument can be themed with a visual style that the users will find appealing, perhaps presented as an interactive application.

Procedure

The Social Connectedness instrument will be used as a base-line measure and so be given during the pre-interaction stage.

Analysis

Results from this instrument are in the form of Likert scales. Previous work with the Social Connectedness instrument involved data analysis of results; the same methods can be reused here.

Validity
The instrument has been validated and is directly applicable to the social evaluation theme.

3.5.2 Agents

TRAVELLER Character Evaluation Questionnaire
Overview
The Character Evaluation Questionnaire (CEQ) for TRAVELLER is based on the same protocol as the CEQ for MIXER, as is the process by which the instrument is embedded. Some specifics must be changed to match the context of TRAVELLER. There will be some differences in the questions as the functional requirements of the characters differ, for example the implementation of culture in MIXER focuses on group level interaction while TRAVELLER operates on higher-level cultural dynamics.

Motivational Analysis
Overview
The Motivational Analysis instrument for TRAVELLER is the same as in MIXER, the only difference is the content of the scenario it implements.

3.5.3 Engagement

Real-time Engagement State
Overview
The instrument is identical to the MIXER version. The application powering this instrument might use a different UI design, given the different audience, but as the application is very simple this might not be necessary.

3.5.4 Comprehension

Progression
Overview
As comprehension is a requirement for progression through TRAVELLER, the stage to which the user gets can be seen as an ambient measure of progression.
Embedding
No explicit embedding strategy is needed for this instrument.
Procedure
No explicit procedure is required for this.
Analysis
In order for progression through the scenario to be illustrative of comprehension, a mapping has to be made from points in the scenario to knowledge outcomes. This mapping is contingent in the content of the scenario, which is under development.

4 Appendices

Appendix I: Matrix of Instruments

The following instruments are being used in the evaluation of MIXER and TRAVELLER (Table 1)

	Social	Agent	Engagement	Comprehension
MIXER	• *Baseline: ICNA** • *Baseline: Empathy Index* • Remote inquisitor • Card sort	• MIXER CEQ • Empathy Index • Mot. Analysis (Story)	• RT Engagement State	• *Baseline: Att. levels* • Buzz
TRAVELLER	• *Baseline: INCA* • *Baseline: Empathy Index* • Physiological measures	• TRAVELLER CEQ • Empathy Index • Mot. Analysis • Physiological measures	• RT Engagement State • Physiological measures	• Progression

Table 1: Matrix of Instruments.

Appendix II: Bryant's Empathy Index for Children

Answer Yes or No to each of the questions.

UNDERSTANDING FEELINGS (F1)

9 Girls who cry because they are happy are silly.
3 Boys who cry because they are happy are silly
20 I think it is funny that some people cry during a sad movie or while reading a sad book
2 People who kiss and hug in public are silly
21 I am able to eat all my cookies even when I see someone looking at me wanting one
16 It's silly to treat dogs and cats as though they have feelings like people
18 Kids who have no friends probably don't want any
10 It's hard for me to see why someone else gets upset
17 I get mad when I see a classmate pretending to need help from the teacher all the time

FEELINGS OF SADNESS (F2)

12 It makes me sad to see a boy who can't find anyone to play with
1 It makes me sad to see a girl who can't find anyone to play with
6 I get upset when I see a girl being hurt
14 I get upset when I see a boy being hurt
11 I get upset when I see an animal being hurt
4 I really like to watch people open presents, even when I don't get a present myself

TEARFUL REACTION (F3)

19 Seeing a girl who is crying makes me feel like crying
5 Seeing a boy who is crying makes me feel like crying
13 Some songs make me so sad I feel like crying
8 Sometimes I cry when I watch TV
7 Even when I don't know why someone is laughing, I laugh too
15 Grown-ups sometimes cry even when they have nothing to be sad about
22 I don't feel upset when I see a classmate being punished by a teacher for not obeying school rules

Appendix III: Social: Baseline: INCA

This set of tools for cultural assessment contains a battery of instruments including video based scenarios and role play, the full set of tools is available at: http://www.incaproject.org/tools.htm

Appendix IV: Social Baseline: McGill Friendship Questionnaire

Please choose the option that applies to you the most. Answer all the questions.

1) Do you have a best friend?
    a) Yes, I have one or two best friends with whom I share almost everything.
    b) Yes, I have several friends whom I consider to be my best friend.
    c) No, I don't have a best friend.
2) Why do we need a friend?
    a) We need someone to confide into.
    b) We need someone who can listen to all our tantrums.
    c) We need someone with whom we can have fun.
    d) All of the above.
    e) We don't really need friends.
3) Dou you mingle with people well?
    a) I am an extrovert; I love to make new friends.
    b) I am an introvert; I keep my distance.
4) Are you a very enthusiastic friend?
    a) Of course, I call up my friends just to chat with them.
    b) Not really, I only call my friends to make some specific arrangements.
5) When you have a personal problem what do you do?
    a) I would work it out on my own.
    b) I would share it with my friend and be comforted.
    c) I would try to forget it.

Appendix V: Social Connectedness in Mainstream Society (SCMN) & Social Connectedness in the Ethnic Community (SCETH)

Please indicate your agreement with the following items using the 1-7 scale below. There are no right or wrong answers. Please be open and honest in your responding.

1	→	4	→	7
Strongly Agree		Neither agree nor Disagree		Strongly Agree

1. _____ I feel a sense of closeness with U.S. Americans.
2. _____ I feel a sense of belonging to U.S. Society.
3. _____ I feel accepted by U.S. Americans.
4. _____ I feel like I fit into U.S. Society.
5. _____ I feel connected with U.S. society.

1. _____ I feel a sense of closeness with _____ Americans.
2. _____ I feel a sense of belonging with _____ American community.
3. _____ I feel accepted by _____ Americans.
4. _____ I feel like I fit into the _____ American community.
5. _____ I feel connected with the _____ American community.

Appendix VI: *GEQ- Game Engagement Questionnaire questions*

1. I lose track of time
2. Things seem to happen automatically
3. I feel different
4. I feel scared
5. The game feels real
6. If someone talks to me, I don't hear them
7. I get wound up
8. Time seems to kind of stand still or stop
9. I feel spaced out
10. I don't answer when someone talks to me
11. I can't tell that I'm getting tired
12. Playing seems automatic
13. My thoughts go fast
14. I lose track of where I am
15. I play without thinking about how to play
16. Playing makes me feel calm
17. I play longer than I meant to
18. I really get into the game
19. I feel like I just can't stop playing

Appendix VII: PQ Items

1. How much were you able to control events?
2. How responsive was the environment to actions that you initiated (or per- formed)?
3. How natural did your interactions with the environment seem?
4. How completely were *all* of your senses engaged?
5. How much did the visual aspects of the environment involve you?
6. How much did the auditory aspects of the environment involve you?
7. How natural was the mechanism, which controlled movement through the environment?
8. How aware were you of events occurring in the real world around you?
9. How aware were you of your display and control devices?
10. How compelling was your sense of objects moving through space?
11. How inconsistent or disconnected was the information coming from your various senses?
12. How much did your experiences in the virtual environment seem consistent with your real-world experiences?
13. Were you able to anticipate what would happen next in response to the actions that you performed?
14. How completely were you able to actively survey or search the environment using vision?
15. How well could you identify sounds?
16. How well could you localize sounds?
17. How well could you actively survey or search the virtual environment using touch?
18. How compelling was your sense of moving around inside the virtual environment?
19. How closely were you able to examine objects?
20. How well could you examine objects from multiple viewpoints?
21. How well could you move or manipulate objects in the virtual environment?
22. To what degree did you feel confused or disoriented at the beginning of breaks or at the end of the experimental session?
23. How involved were you in the virtual environment experience?
24. How distracting was the control mechanism?
25. How much delay did you experience between your actions and expected out- comes?
26. How quickly did you adjust to the virtual environment experience?
27. How proficient in moving and interacting with the virtual environment did you feel at the end of the experience?
28. How much did the visual display quality interfere or distract you from per- forming assigned tasks or required activities?
29. How much did the control devices interfere with the performance of assigned tasks or with other activities?
30. How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those tasks or activities?
31. Did you learn new techniques that enabled you to improve your performance
32. Were you involved in the experimental task to the extent that you lost track of time?

5 Technical Annex

Technical Note I: In-role Flashcard Quiz

While working on the embedding strategy for MIXER and TRAVELLER the question of how to embed the Flashcard Quiz instrument was raised. In a shallow sense it is easy to embed the instrument in the scenario: by having the scenario involve a quiz in which the users are asked to take part. However, of the questions are aimed at general comprehension of the interaction or scenario, the content of questions can feature any proposition related to the scenario. This means that in order for the instrument to be embedded and in role the feature of a 'Quiz Master' who knows everything about the scenario world must be perceived by the user as realistic. For example, a question could ask about an event or proposition that is private to the user and one other character, this question being raised would require the Quiz Master to have an unrealistic insight into events couldn't have witnessed. The Quiz Master could be a character from the interaction, e.g. a teacher, but this would strictly limit the domain of possible questions to those things that that character could be realistically aware of. This is a severe limitation for an instrument designed to examine overall comprehension as such an instrument needs to be able to gather data on any and all aspects of the scenario.
The natural solution to this is to invoke a contextually supernatural character, for example an alien, who has overseen everything. This character, while being aware of all the events in the scenario has questions that are presented to the user as a multiple-choice quiz during discrete episodes in the scenario. Such a character is, in principal, benign with respect to the primary scenario, in that invoking it doesn't alter the user's perception of the mechanics of the scenario.

Technical Note II: Character Evaluation and the Magic Circle

During the course of a virtual world interaction there exists a reality that we present to the user and hope that they feel that they are part of that reality in a convincing and continuous way. If there are moments during which that isn't the case, for example some aspect of the interaction seems momentarily to no fit into the reality, then the magic circle has been broken. This creates a problem for some times of evaluation as it may be necessary to intentionally breach the magic circle.

If an interaction contains a character capable of autonomous behaviour, the character and their behaviour should be perceived as congruent with the magic circle. An evaluation of the effectiveness of the characters mind architecture could pose questions like "was the character's behaviour believable?" or "did the character have a mind of their own?", as demonstrated in D2.1. The problem with these types of questions is that they require the user to suppose two situations; one in which the character and its mind are, or belong, in the magic circle, the other in which the mind of the character (as this is what generates its behaviour) is outside, or doesn't belong, in the magic circle. The question requires the user to suppose that the magic circle is in breach, therefor breaching it.

Conceptually it seems difficult to solve this problem, you can't ask the user to suppose the character doesn't belong without them allowing the possibility that something doesn't belong. There is, however, a way around the problem. The solution is to create a 'cheap' extension to the magic circle. Cheap in the sense that it requires little design or development and can be instantiated during the evaluation without requiring corresponding components in the primary scenario to ever exist (although such components might be useful.) The key is that in the case that the character's behaviour is not believable, this is analogous to the proposition that the character is in some sense an *imposter*. So in the case of the question about whether the character's behaviour is 'realistic,' the question could be reinterpreted to "Could the character be an alien in disguise?" Intuitively, an alien in disguise would try to act convincingly, which is by definition the same as acting 'realistically' in the context of the original question, but could give away signs that it is not acting realistically. The user could be asked to consider these signs and attribute them to the character belonging to the extended magic circle rather than being in breach of it.

As stated above, this requires no corresponding components in the primary scenario. The alien imposters do not need to be mentioned before the evaluation step. This *imposters* component can be mentioned during the primary interaction, leading to the evaluation, but it is important that imposter behaviour always corresponds to how the character would behave if it were behaving unrealistically. This is because it is important that the 'alienness' of the characters is not perceived as valid behaviour, relative to the primary scenario, as being so would make alien behaviour a candidate for *realistic* behaviour. If, for example, the scenario called for an agent to replicate 'alien' behaviour then it doing so would be a success not a failure.

Technical Note III: Culture and Evaluation

Clearly, the measurement of an individual's cultural perspective is primary in detecting the impact of any artifact on that perspective. There is a key weakness with many of the methods typically chosen to measure such cultural perspectives. Cultural perspective measures and instruments measure something that is long-term. For instance, take measurement of an individual's view of other cultures. It is possible that learning some salient fact about that culture could change that person's view of another culture in a very short time. Will this change be very permanent? How deep is this change? If bestowing some facts upon a person, with only a shallow pedagogical approach, were enough to transform a person's cultural perspective then we wouldn't have any problems with intercultural relations and the work of this project would be unnecessary. It is intuitively obvious that the changes we would like to foster and measure are deep and slow.

In the context of a discreet interaction, such slow modulations are problematic. Firstly, a short term measure of a persons cultural perspective seems pointless as we wouldn't expect a measureable difference over such a short time, let alone a significant one. This seems to enforce the use of a more long-term measure, but actually, this introduces further problems. Take an interaction and evaluation that happens over the course of six weeks. Given that the participants aren't contained in a laboratory setting, anything that happens during that six-week period could significantly skew the results. For example, a user going on holiday overseas would represent a massive bias to the data, drowning out any measureable signal generated by the experiment.

Even if the exposure to bias could be corrected for somehow this, ironically, would not eliminate bias if the measurement is to be done using standard cultural perspective measurement tools / instruments. The reason is that such tools often rely what will be referred to as *auto-simulation*. The participant is asked report *how they would act* in a situation by simulating being in that situation and simulating acting in it in their imagination (e.g. "When in another culture I…".) There are immediate reasons to doubt the results of such a test, firstly, how 'realistic' is their simulation and secondly, even if it is 'realistic' how sure are we that what the participant says they would do is similar to reality? Much of the psychological content of an intercultural interaction is founded on unconscious responses; can someone simulate having these unconscious responses? Linking this back to the example above, the long-term evaluation, if we are asking users to tell us how they would act in a particular situation, if they haven't had an opportunity to be in that situation how do they know how they would act? So either, during the course of the long-term evaluation, the participant comes in contact with other cultures, thereby biasing the results, or they don't, in which case we have to wonder how reliable their account of how they *would* behave is. Add to this other basic biases, e.g. child participants' tendency to report what they think the tester wants to hear or is the 'right answer', and the method seems have far too many biases to be useful.

All the changes that we would care to measure must arise from some change in cognition. and empathic reactions. Behaviour is highly abstracted from cognition, especially in the case of cultural perspective, so measuring via behaviour seems impractical (measuring behaviour while simulating behaviour is even more impractical.) The important question seems to be: what is changing in a person's thought processes? Further, how can we measure this change?

Technical Note IV: Embedded Evaluation

Generally the concept of embedding evaluation refers to the integration of evaluation instruments into the primary interaction. It is useful to think about what an instrument is at a most fundamental level. An instrument is effectively a *query* made to the participant. Evaluation is the design, delivery and analysis of the results of these queries. Embedding evaluation is the process of putting the query and its delivery into the 'Magic Circle'. The Magic Circle is a conceptual boundary between the real world and the fictitious world that the primary scenario and interaction are intended to simulate. The Magic Circle permits the idea that it can go beyond the virtual world in that it contains not just what is presented to the user via the interaction modality but also structures generated by the user and even aspects of the physical world. Indeed, this magic circle is largely a figment of the imagination of the user, which is why the users perceptions are so important, because these perceptions are the arbiter for what is congruent and what is not. The Magic Circle also has the useful property of being dynamic and extensible. The user can and will incorporate new ideas and propositions into this magic circle thereby extending it, if it is seen as congruent. Embedding evaluation is really about putting evaluation in the magic circle (or by allowing the magic circle to extend to accommodate it.)

There is a spectrum of possible levels at which the instruments can be embedded:

Altering the Appearance of the Query

The shallowest way to embed an instrument is to harmonise the presentation of instrument with the presentation style of the primary interaction (e.g. making a questionnaire look like it belongs in the scenario world.) The instrument then has the 'look and feel' of something that belongs but its content, or the implications raised by its content, may well be incongruent with the scenario. For some types of instrument this approach is enough, for example gaining demographic data by means of a form. Such a form could well be something a member of the fictitious world would do and so giving the form a congruent appearance would be enough to embed it.

Altering the Delivery of the Query

In some cases the delivery of an instrument is at odds with the reality in which it is to be embedded. An example would be if the instrument were a question as to the user's feelings about a particular proposition. The instrument would likely be founded on some specific theory or model that specifies its exact wording so that it targets precisely the data needed. It might not be possible to simply alter the appearance of this instrument if, for example, the user would not expect to be periodically asked, in a direct way, about their feelings. Embedding such an instrument would involve finding some mechanism in the current scenario by which to administer the query.

Altering the Query

Some instruments are of a form that simply altering their delivery mechanism isn't enough to embed the instrument. The instrument has been altered in appearance, further, it has been given a more congruent delivery mechanism involving some aspect of the primary scenario, but still it doesn't fit into the scenario. This could be because the nature of query that the

instrument represents cannot be made in a congruent way (e.g. the Character Evaluation Questionnaire involves asking questions that breach the magic circle. In this case it is necessary to alter the query itself such that it gives rise to an analogous response. This is discussed in detail in the technical note #.

Technical Note V: Measuring Engagement

While it is possible to, in principal, measure the level of a user's feelings of engagement using post hoc questioning of some sort, our evaluation approach seeks to look at the temporal characteristics of user engagement state. For example, is a continuously moderately high engagement level better than an engagement level is dynamic but with high peaks? Does the user have to experience a low level of engagement in order for higher levels to 'feel' high? If so what is the optimal time-span for peaks and troughs? In order to begin to understand these temporal characteristics we need an instrument that delivery good temporal resolution. For this, post hoc measurements are lacking.

A second problem with post hoc measurement of user experience is that it suffers from perception biases of retrospection. A user's retrospective perception of the events over a time-span are influenced by aspects of the experience such that a user's report from memory will differ from their real-time report of the same time-span (O'Brien & Toms, 2010).

The above means that in order to measure engagement effectively a trade-off has to be made between eliminating retrospective biases and the bias caused by the engagement instrument interfering with the interaction it is measuring from. To say the instrument is applied in real-time is to say that it is in some way distracting the user from the primary interaction. This measurement paradox is, fundamentally, part of all measurement activities. The important question is: Can this *interference bias* be minimized such that it is less than the corresponding *retrospective bias*?

6 References

Babbitt, B. (1989). Questionnaire construction manual annex. Questionnaires: Literature survey and bibliography. *Operations Research Associates*, (June). Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA037815

Bosse, T., Memon, Z. A., & Treur, J. (2011). A RECURSIVE BDI AGENT MODEL FOR THEORY OF MIND AND ITS APPLICATIONS. *Applied Artificial Intelligence*, *25*(1), 1-44. doi:10.1080/08839514.2010.529259

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., Mcbroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*(4), 624-634. Elsevier Inc. Retrieved from http://dx.doi.org/10.1016/j.jesp.2009.02.016

Bryant, B. (1982). An index of empathy for children and adolescents. *Child development*, *53*(2), 413-425.

Hofstede, G. (1994). Business cultures. *Unesco Courie*, *47*(4), 12-16.

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, *66*(9), 641-661. doi:10.1016/j.ijhcs.2008.04.004

Killgore, W. D. (1998). The Affect Grid: a moderately valid, nonspecific measure of pleasure and arousal. *Psychological Reports*, *83*(2), 639-642.

Lunzer, E. A., & Gardner, W. K. (1979). The effective use of reading. *London: Heinemann Educational.*

Mc Breen, J., Di Tosto, G., Dignum, F., & Hofstede, G. J. (2011). Linking Norms and Culture. *2011 Second International Conference on Culture and Computing* (pp. 9-14). IEEE. doi:10.1109/Culture-Computing.2011.11

Oatley, K. (1995). A taxonomy of the emotions of literary response and a theory of identification in fictional narrative. *Poetics*, *23*(1-2), 53–74. Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/0304422X94P4296S

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69. Wiley Online Library. doi:10.1002/asi.21229.1

Salen, K., & Zimmerman, E. (2004). *Rules of Play: Game Design Fundamentals* (Vol. 2004, p. 672). MIT Press. Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0262240459

Schoenau-Fog, H. (2011). Hooked!–Evaluating Engagement as Continuation Desire in Interactive Narratives. *Interactive Storytelling*, 219–230. Springer. Retrieved from http://www.springerlink.com/index/J1044L140159153W.pdf

Turner, J. C., & Oakes, P. J. (1986). The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology* (Vol. 25, pp. 237-258).

Witmer, B. G., & Singer, M. J. (1998). Measuring Presence in Virtual Environments : A Presence. *Technology*, *7*, 225-240.

Wood, J., & Wood, L. (2008). Card Sorting : Current Practices and Beyond. *Most*, *4*(1), 1-6. Retrieved from http://www.upassoc.org/upa_publications/jus/2008november/JUS_Wood_Nov2008.pdf

Yoon, E., Jung, K. R., Lee, R. M., & Felix-Mora, M. (2012). Validation of social connectedness in mainstream society and the ethnic community scales. *Cultural diversity & ethnic minority psychology*, *18*(1), 64-73. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22250899