‘What I see is not what you get’: why culture-specific behaviours for virtual characters should be user-tested across cultures

Nick Degens, Birgit Endrass, Gert Jan Hofstede, Adrie Beulens, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Degens, Nick, Birgit Endrass, Gert Jan Hofstede, Adrie Beulens, and Elisabeth André. 2017. “‘What I see is not what you get’: why culture-specific behaviours for virtual characters should be user-tested across cultures.” *AI & Society: Journal of Knowledge, Culture and Communication* 32 (1): 37–49. https://doi.org/10.1007/s00146-014-0567-2.

‘What I see is not what you get’: why culture-specific behaviours for virtual characters should be user-tested across cultures

Nick Degens · Birgit Endrass · Gert Jan Hofstede · Adrie Beulens · Elisabeth André

Abstract Integrating culture into the behavioural models of virtual characters requires knowledge from very different disciplines such as cross-cultural psychology and computer science. If culture-related behavioural differences are simulated with a virtual character system, users might not necessarily understand the intent of the designer. This is, in part, due to the influence of culture on not only users, but also designers. To gain a greater understanding of the instantiation of culture in the behaviour of virtual characters, and on this potential mismatch between designer and user, we have conducted two experiments. In these experiments, we tried to simulate one dimension of culture (Masculinity vs. Femininity) in the behaviour of virtual characters. We created four scenarios in the first experiment and six in the second. In each of these scenarios, the same two characters interact with each other. The verbal and non-verbal behaviour of these characters differs depending on their cultural scripts. In two user perception studies, we investigated how these differences are judged by human participants with different cultural backgrounds. Besides expected differences between participants from Masculine and Feminine countries, we found significant differences in perception between participants from Individualistic and Collectivistic countries. We also found that the user’s interpretation of the character’s motivation had a significant influence on the perception of the scenarios. Based on our findings, we give recommendations for researchers that aim to design culture-specific behaviours for virtual characters.

N. Degens (✉) · G. J. Hofstede · A. Beulens
Information Technology, Wageningen University,
Hollandseweg 1, 6706 KN Wageningen, The Netherlands
e-mail: nick.degens@wur.nl

B. Endrass · E. André
Human Centered Multimedia, Augsburg University,
Universitätsstrasse 6a, 86159 Augsburg, Germany

1 Introduction

Virtual characters hold great potential for the field of intercultural training. As intercultural training can be expensive and difficult to organize, virtual characters could replace real-life actors and trainers to create educational tools that can be used by trainees whenever and wherever they want.

Such characters need to be programmed to show appropriate culture-specific behaviour. To do this, one must understand how culture influences behaviour and be able to apply that knowledge to create computational models of culture-specific behaviour that determine the behaviour of those characters.

Even if the implemented culture-specific behaviours are based on extensively validated theories from cultural psychology, there is no guarantee that these behaviours will be perceived by human participants as appropriate; the designer’s intention might not be perceived or understood by the users. This is, in part, due to the fact that culture affects the perception and interpretation of both designers and users. Such problems are especially likely to occur when designing for people with different cultural backgrounds than those of the designers.

For this article, we conducted two experiments that provide us with a deeper insight into two processes. The first process is about the modelling and instantiation of culture in the behaviour of virtual characters. The second

process is about the perception of this instantiated behaviour by human participants. Specifically, we investigated how the behaviour of virtual characters is judged by participants from different cultural backgrounds.

This article is structured as follows: In the *related work section,* we give a brief overview of research that integrates culture in the behaviour of virtual characters. In the *theoretical background section,* we give a description of the underlying model of culture that we use in the rest of the article. The next two sections describe, respectively, *the first and the second experiment*. In both of these sections, we show how we have applied the model of culture to create different scenarios featuring virtual characters with different synthetic culture scripts, describe the evaluation procedure, present and discuss the results. We end the article with some *conclusions and recommendations* for designing culture-specific behaviours for virtual characters.

2 Related work

To create virtual characters that show appropriate culture-specific behaviour, different aspects of human behaviour need to be considered. Some researchers have focused on formalizing verbal behaviour. For example, the tactical language training system (TLTS) (Johnson et al. 2004) focuses on training competencies for military purposes and training skills in foreign languages such as Iraqi Arabic, Dari or Pashto (Johnson and Valente 2008). In the TLTS, trainees have to learn a foreign language in order to complete the tasks provided by the system.

Kim et al. (2009) introduce the BiLAT system that focuses on teaching intercultural skills in order to communicate with people from Iraq. The user has to adapt to interaction rules that are appropriate in Iraq and use those interaction rules to successfully negotiate with simulated Iraqi characters.

Another example of integrating verbal aspects of culture-specific behaviour is work carried out by Endrass et al. (2011b). Focusing on topic selection and dialogue structure in small talk situations, they investigated culture-related differences between participants from Germany and Japan.

Other researchers have focused on non-verbal behaviours such as facial expressions, gesture selection, expressivity, spatial behaviour and gaze. Jan et al. (2007), for example, take into account eye-gaze, proxemics and turn-taking to point out differences in behaviour between people from Arab countries and people from the USA. This is done by having users observe a group of virtual characters interacting.

Koda et al. (2008) investigate the perception of virtual characters' appearance by studying the facial expressions of avatars. In a follow-up study (Koda et al. 2010), they focused on different regions of the face (eyes and mouth) and conducted a cross-cultural study in Hungary and Japan. In their results, the authors report that Japanese participants found facial cues in the eye region more important than Hungarian participants, who concentrated more on facial cues in the mouth region.

Mascarenhas et al. (2009) focus on rituals, which are described as symbolic social activities which are carried out in a predetermined fashion dependent on cultural background. For their simulation, two groups of characters were created that follow goals and create plans in order to carry out different rituals.

Aylett et al. (2009) use fantasy cultures (i.e. not based on existing cultures) for their virtual characters. With their work, they take an educational approach to create intercultural empathy in the user.

In our work, we take a more fundamental approach of culture-related research by not focusing on specific learning objectives, but by primarily investigating whether participants perceive the implemented culture-specific behaviours and whether they consider that behaviour appropriate. This effect is likely to occur according to the similarity principle (Byrne 1971), which states that interaction partners who perceive themselves as being similar are more likely to like each other.

3 Theoretical background

A well-established model of culture is described by Hofstede et al. (2010). Their model consists of six dimensions of values, which describe common issues that every society faces. Participants from different countries were found to have different solutions for these issues, which is represented by the score of their country on those dimensions. Over 70 countries were categorized using this model by assigning a value for each dimension. This model shows, according to Smith (2006), superior continued validation over other dimensional frameworks.

The current version of Hofstede's model consists of the following six dimensions: Power Distance, Individualism versus Collectivism, Masculinity versus Femininity, Uncertainty Avoidance versus Uncertainty Tolerance, Long-term versus Short-term Orientation, Indulgence versus Restraint. We will only discuss two of these dimensions in greater detail, as they are most relevant to our experiments.

Individualism versus Collectivism describes the degree to which individuals are integrated into a group. On the Individualist side, ties between individuals are loose, and everybody is expected to take care of him or herself. On the Collectivist side, people are integrated into strong, cohesive groups.

Masculinity versus Femininity describes the distribution of roles between males and females. On the masculine side, people are supposed to be assertive or competitive. On the feminine side, people are supposed to be caring and modest.

The above dimensions deal with values, which are one manifestation of culture. There are other manifestations, such as practices, which are comprised of elements that are more easily visible to an outsider. Examples of practices are rituals (collective activities that serve a relational purpose) and symbols (messages that carry a similar meaning for all those who belong to a group). These practices are important to consider when designing culture-specific behaviours for virtual characters.

Hofstede and colleagues introduce *synthetic cultures* (Hofstede et al. 2002), which help to clarify the relationship between the dimensions of culture and practices. These synthetic cultures are based on the extremes of each of the dimensions of culture, and each synthetic culture contains the culture's values, core distinctions, key elements as well as words with positive or negative connotations. These descriptions can be used to create extreme examples of the culture clashes that can occur in real life. They also serve as a good basis for designing virtual characters' behaviour, since clear behavioural trends are provided for synthetic cultures.

Due to the influences of the dimensions on each other, we had to isolate one dimension. We were interested in the user's perception of the distinction between performing and caring, as such distinctions are very important in an educational or work environment. These situations would be a manifestation of the Masculinity versus Femininity dimension. The synthetic culture scripts based on this dimension are called *Mascu* and *Femi*; the following ideas are taken from the description of these two synthetic cultures (Hofstede et al. 2002).

The *Mascu* synthetic culture has as core value *winning* and as core distinction the distinction between *men and women*. Key elements are statements such as "material success and progress are dominant values", "bigger and faster are better" and "failing (at school, at work, in sports or wherever) is a disaster". *Mascu's* are described as being loud and verbal, with a tendency to argue with others. Non-verbally they like physical contact, direct eye contact and animated gestures. Words with a positive connotation are as follows: career, competition, fight, aggressive, success, winner, force, fast, big, power and action.

The *Femi* synthetic culture, which is located on the opposite side of the same dimension, has as core value *caring for others, especially the weak* and as core distinction the distinction between *caring and needing care*. Key elements are statements such as "small and slow are beautiful", "everybody is supposed to be modest, soft-spoken and empathetic—men and women alike", and "conflicts are resolved through compromise and negotiation". Typically, *Femi*'s do not raise their voice and like small talk and agreement. Non-verbally they do not take much space and are warm and friendly in conversation. Words with a positive connotation are as follows: caring, solidarity, modesty, compromise, help, love, soft, slow, tender and touch.

4 Experiment 1

4.1 Method

As the basis for a scenario, we chose to focus on a conversation between a professor and a student. In this conversation, a female student needs an extension for a deadline and asks the male professor whether this is possible. To show different culture-specific behaviours in this scenario, we designed two scripts for each character. These scripts contain the verbal and non-verbal behaviours of these characters, and they were based on the description of the *Mascu* and *Femi* synthetic cultures described in the previous chapter. Since both characters have their own script, there are four different scenarios in total (see Table 1).

The student with the *Mascu* script wants to *perform* the best she can and needs an extension to improve the assignment. In contrast, the student with the *Femi* script cares more about her family members than the assignment and needs an extension because she had to attend an important family event and was unable to finish the assignment on time.

The professor with the *Mascu* script considers improving the assignment a good reason and will give the student an extension. The same professor considers a family event a weak excuse and will not give an extension. In contrast, the professor with the *Femi* script considers a family event a good reason and will give the student an extension. The same professor does not consider it important to improve the assignment, but he will still give an extension if it is important to the student with the *Mascu* script (for all the outcomes, see Table 1).

The wording for *Mascu* characters is designed in a direct way, e.g. "No. Everybody knew it had to be handed in today". Vice versa, the utterances of the *Femi* professor's speech focus on caring towards the students and uses soft wording, e.g. "Oh nice, a family event. What was it?"

The non-verbal behaviour also differs for *Mascu* and *Femi* characters. Previous research (Endrass et al. 2011a) describes how participants from different nationalities differ in their prototypical body postures. Based on these findings, we chose more moderate and soft body postures

Table 1 The student's reason, the professor's response and the scenario outcome for each of the four scenarios (the student gets no extension in the scenario that is italicized)

Scenario			Reason for needing an extension	Response of professor	Does the student get an extension?
No.	Student	Professor			
1	*Femi*	*Mascu*	Family event	Needing an extension because of attending a family event is a weak excuse	No
2	Mascu	Mascu	Performance	Wanting to perform better is a good reason for needing an extension	Yes
3	Femi	Femi	Family event	Attending a family event is good reason for needing an extension	Yes
4	Mascu	Femi	Performance	Professor does not understand, but will give an extension if its important to the student	Yes

for the *Femi* characters (e.g. folding hands in front of the body, or touching facial regions with the hands, see the left screenshot in Fig. 1, while we chose upright body postures that use more space for *Mascu* characters (e.g. Arms Akimbo—hands on the hips and elbows bowed outward or fold arms in front of the body, see the right screenshot in Fig. 1).

According to previous work on culture-specific non-verbal behaviour (Endrass et al. 2011a), differences in gestural expressivity of virtual characters such as spatial extent or speed can be found in people from different nationalities. Based on these findings, we used gestures with a larger spatial extent and higher speed for the *Mascu* characters compared to the *Femi* characters.

Communicative gestures also differ between the characters in our scenarios. A pointing gesture performed by a *Mascu* character, for example, can contain pointing at the interlocutor, while a pointing gesture carried out by a *Femi* character points at an imaginary point in space, e.g. behind one's shoulder, to refer to the family event, prototypical culture-specific gestures for virtual characters can be observed here.[1]

To realize the scenarios described above, we used the Virtual Beergarden scenario running in the AAA application (Damian et al. 2011). In this application, characters can be loaded that are able to move around in the scenario freely, use gestures and communicate with each other. Verbal behaviour is realized by a text-to-speech component, while for non-verbal behaviour a set of over 70 animations is available.

Each of the four scenarios lasts for about half a minute (between 23 and 32 s) and contains between 6 and 10 dialogue turns. In order to avoid side effects evoked by the gender of the characters, we left the genders of the virtual characters constant over all four scenarios. Figure 1 shows the virtual scenario including our professor–student setup with a female (gender) student and a male (gender) professor, showing *Femi* (culture) or *Mascu* (culture) behaviour.

[1] http://www.hcm-lab.de/projects/animations.

4.1.1 Evaluation procedure and hypotheses

To evaluate the perceptions of human participants, we recorded four videos showing each of the four scenarios. After answering demographic questions, such as age, gender and nationality, participants were able to view all four videos. They were given the opportunity to watch the videos multiple times. Our aim for the evaluation was to discover how the implemented culture-related differences are perceived by participants from different countries. Based on the contents of the scenarios, we identified three hypotheses:

Participants from countries that score higher on the dimension of Masculinity will be more likely to…

1.1 …consider the behaviour of characters with the *Mascu* script more appropriate than characters with the *Femi* script;
1.2 …consider getting an extension less fair than not getting an extension, because they will be unforgiving towards underperforming students;
1.3 …like the characters with the *Mascu* script more than characters with the *Femi* script.

To test these hypotheses, we created the following questions, which the participants had to answer after watching each video:

- (1.1) Do you think the student acted appropriately?
- (1.1) Do you think the professor reacted appropriately?
- (1.2) Do you think the professor's decision was fair towards the student who asked for the extension?
- (1.2) Do you think the professor's decision was fair towards other students who did not ask for an extension?
- (1.3) Would you like to have this professor as a teacher?
- (1.3) Would you like to have this student as a friend?

Fig. 1 Example of the characters interacting in our scenario (*left side*: *Femi* student and *Femi* professor; *right side*: *Mascu* student and *Mascu* professor)

Participants were able to rate these questions on a 7-graded Likert scale, rating their agreement with "yes, absolutely", "yes", "somewhat yes", "neither yes or no", "somewhat no", "no" or "no, not at all". After answering the above questions, participants were also able to further clarify their choices in a comment box.

Using the recorded videos, we created an online survey and circulated the link to universities of different countries, and we also circulated the link to people interested in culture. For further information on the study setup, introduction, dialogues and videos, please visit the online study.[2]

4.2 Results

In total, 75 participants of 10 different nationalities took part in our study. Since we only collected enough data for statistical analysis of participants from four countries, we only considered the data from those participants. In that manner, 15 people from Germany (seven females; mean age: 27.8; SD age: 3.57), 11 people from Japan (five females; mean age: 27; SD age: 6.4), 19 people from the Netherlands (seven females; mean age: 23.1; SD age: 3.27) and 20 people from Thailand (11 females; mean age: 28.5; SD age: 3.06) were included for analysis, while 10 participants from six other countries were excluded. The scores for the four participating nationalities on Hofstede's dimensions are provided in Table 2. As the data were not normally distributed, we used nonparametric tests in all cases.

4.2.1 Appropriateness of behaviour

The dialogue was different depending on the scripts of the characters; it might thus be that the appropriateness of one of the characters was influenced by the other character. This is why we compared each scenario to every other scenario using the Wilcoxon signed-rank test. We found that the participants from Germany and the Netherlands found the *Mascu* student significantly more appropriate (two comparisons for participants from Germany and two comparisons for participants from the Netherlands) and that participants from Thailand found the *Femi* student significantly more appropriate (two comparisons for participants from Thailand). See Table 3 for an overview of these results.

To account for differences due to culture, we used the Mann–Whitney U test to compare groups comprised of people from different nationalities. This distribution of participants was done for the Masculinity versus Femininity dimension (Germany and Japan vs. the Netherlands and Thailand). We found no significant differences for this dimension. After looking at the data in more depth, we discovered that the Individualism versus Collectivism dimension (Germany and the Netherlands vs. Japan and Thailand) had a large effect on perception. With this configuration, we found significant differences with regard to the appropriateness of the characters (see Table 4).

Table 2 Number of participants from each country and the scores for these countries on Hofstede's dimensions (from Hofstede et al. 2010)

	N	PDI	IND	MAS	UAI	LTO	IvR
Germany	15	*35*	67	55	65	83	*40*
Japan	11	54	46	*95*	*92*	*88*	42
The Netherlands	19	38	*80*	*14*	*53*	67	*68*
Thailand	20	*64*	*20*	34	64	*32*	45
Difference between highest and lowest		29	60	81	39	56	28

Highest and lowest scores are italicized

4.2.2 Fairness of professor's decision

To determine whether people from certain countries perceived getting an extension as significantly more fair in one of the scenarios, we used the Wilcoxon signed-rank test to compare each scenario to every other scenario.

For the participants from the Netherlands, the fairness of the extension to the student was significantly higher for the

[2] http://mm-werkstatt.informatik.uni-augsburg.de/survey/index.php?sid=21954&lang=en.

Table 3 Comparison of student appropriateness between different scenarios for the participants from different countries

Scenarios to compare				Country for which the scenarios were compared			
Scenario A		Scenario B		Germany	The Netherlands	Thailand	Japan
Stud	Prof	Stud	Prof				
Femi	*Mascu*	Mascu	Mascu	Scenario B $p = 0.002$	Scenario B $p = 0.026$	–	–
Femi	*Mascu*	Femi	Femi	–	–	Scenario B $p = 0.001$	Scenario B $p = 0.033$
Femi	*Mascu*	Mascu	Femi	–	Scenario B $p = 0.008$	–	–
Mascu	Mascu	Femi	Femi	Scenario A $p = 0.002$	–	Scenario B $p = 0.033$	–
Mascu	Mascu	Mascu	Femi	Scenario A $p = 0.033$	–	–	–
Femi	Femi	Mascu	Femi	–	–	Scenario A $p = 0.008$	–

Named scenario was rated significantly higher (the student gets no extension in the scenarios that are italicized)

Table 4 Appropriateness of student and professor judged by participants from IND countries, Germany and the Netherlands versus COL countries, Japan and Thailand (the student gets no extension in the scenario that is italicized)

Stud	Prof	Mean IND countries	Mean COL countries	p
Appropriateness of student				
Femi	*Mascu*	4.03	3.94	0.767
Mascu	Mascu	5.29	4.42	0.013
Femi	Femi	4.26	5.16	0.040
Mascu	Femi	4.94	4.03	0.015
Appropriateness of professor				
Femi	Femi	3.39	4.90	<0.001

Mascu student (Mdn = 5.5) than with *Femi* student (Mdn = 4.0), $T = 15$, $p = 0.030$, in the scenario with the *Femi* professor.

We found significant differences with regard to the fairness of an extension towards other students (see Table 5). In particular, we found that participants from every country found it less fair towards other students if the student was granted the requested extension. In addition, we found that participants from Germany and Thailand each found the extension significantly fairer in two of the other comparisons (see the bottom three rows in Table 5).

4.2.3 Affective reaction to the characters

We analysed whether people from certain countries would like the professor as a teacher or the student as a friend. To determine whether this was more likely if the character had a certain synthetic culture script, we used the Wilcoxon signed-rank test to compare each scenario to every other scenario.

The results for liking the professor as a teacher can be found in Table 6. Participants from Germany liked the professor as a teacher significantly more in two of the comparisons. Participants from the Netherlands liked the professor as a teacher significantly more in four of the comparisons. Participants from Thailand liked the professor as a teacher significantly more in one of the comparisons. Participants from Japan liked the professor as a teacher significantly more in two of the comparisons. Eight of these nine significant differences can be found when two scenarios are being compared in which the student gets an extension versus the student does not get an extension. In all of these, participants have a preference for the professor as a teacher in the scenarios in which the professor does not grant an extension.

The results for liking the student as a friend can be found in Table 7. Participants from Thailand liked the *Femi* student as a friend significantly more in two of the comparisons. They also liked the *Femi* student more when they were interacting with the Femi professor. Participants from Japan liked the student as a friend significantly more in one of the comparisons. They also liked the *Femi* student more when they were interacting with the Femi professor. All of these significant differences occur when there is a comparison with the scenario with a *Femi* professor and a *Femi* student (which is also the preferred scenario).

Table 5 Comparison of the fairness of the extension to other students between different scenarios for the participants from different countries (the student gets no extension in the scenarios that are italicized)

Scenarios to compare				Countries for which the scenarios were compared			
Scenario A		Scenario B		Germany	The Netherlands	Thailand	Japan
Stud	Prof	Stud	Prof				
Femi	*Mascu*	Mascu	Mascu	Scenario A $p = 0.011$	Scenario A $p < 0.001$	Scenario A $p = 0.001$	Scenario A $p = 0.017$
Femi	*Mascu*	Femi	Femi	Scenario A $p = 0.003$	Scenario A $p < 0.001$	Scenario A $p < 0.001$	Scenario A $p = 0.028$
Femi	*Mascu*	Mascu	Femi	Scenario A $p = 0.005$	Scenario A $p < 0.001$	Scenario A $p < 0.001$	Scenario A $p = 0.007$
Mascu	Mascu	Femi	Femi	Scenario A $p = 0.011$	–	–	–
Mascu	Mascu	Mascu	Femi	–	–	Scenario A $p = 0.023$	–
Femi	Femi	Mascu	Femi	Scenario B $p = 0.013$	–	Scenario A $p = 0.021$	–

Table 6 Comparison of liking the professor as a teacher between different scenarios for the participants from different countries (the student gets no extension in the scenarios that are italicized)

Scenarios to compare				Countries for which the scenarios were compared			
Scenario A		Scenario B		Germany	The Netherlands	Thailand	Japan
Stud	Prof	Stud	Prof				
Femi	*Mascu*	Mascu	Mascu	Scenario B $p = 0.012$	Scenario B $p = 0.002$	Scenario B $p = 0.044$	Scenario B $p = 0.011$
Femi	*Mascu*	Femi	Femi	–	Scenario B $p = 0.038$	–	Scenario B $p = 0.021$
Femi	*Mascu*	Mascu	Femi	Scenario B $p = 0.018$	Scenario B $p = 0.015$	–	–
Mascu	Mascu	Femi	Femi	–	–	–	–
Mascu	Mascu	Mascu	Femi	–	–	–	–
Femi	Femi	Mascu	Femi	–	Scenario B $p = 0.047$	–	–

4.3 Discussion

We expected to find that the respondents from countries that score high on Masculinity considered the behaviour of the characters with the *Mascu* script more appropriate than the behaviour of the characters with the *Femi* script (Hypothesis 1.1).

This hypothesis was not confirmed by our results. They do show, in some of the comparisons, that participants from countries that score high on Individualism (the Netherlands and Germany) considered the behaviour of the student with the *Mascu* script more appropriate, and participants from countries that score low on Individualism (Japan and Thailand) considered the behaviour of the student with the *Femi* script more appropriate (see Tables 3 and Table 4).

For three out of the four scenarios, the participants from Individualistic and Collectivistic countries differ significantly in their perception of the student's appropriateness. The only exception constitutes the scenario with the *Femi* student and the *Mascu* professor, in which an extension was not granted.

We expected to find that participants from countries that score high on Masculinity would think that it is less fair if the student gets an extension (Hypothesis 1.2).

We did not find any significant results with regards to the fairness of an extension for the student who requested an extension.

Table 7 Comparison of liking the student as a friend between different scenarios for the participants from different countries (the student gets no extension in the scenarios that are italicized)

Scenarios to compare				Countries for which the scenarios were compared			
Scenario A		Scenario B		Germany	The Netherlands	Thailand	Japan
Stud	Prof	Stud	Prof				
Femi	*Mascu*	Mascu	Mascu	–	–	–	–
Femi	*Mascu*	Femi	Femi	–	–	Scenario B $p = 0.006$	Scenario B $p = 0.006$
Femi	*Mascu*	Mascu	Femi		–	–	–
Mascu	Mascu	Femi	Femi	–	–	Scenario B $p = 0.013$	–
Mascu	Mascu	Mascu	Femi	–	–	–	–
Femi	Femi	Mascu	Femi	–	–	Scenario A $p = 0.010$	Scenario A $p = 0.044$

We did find that participants from each country tested found it less fair towards the other students if the student received an extension (Table 5). This suggests that the actual outcome of the scenario, e.g. whether an extension is granted or not, had a large influence on the perception of the participants.

We expected that participants from countries that score high on Masculinity would like the *Mascu* professor as a teacher and the *Mascu* student as a friend (Hypothesis 1.3).

Our results did not confirm this hypothesis. We did find that in two of the comparisons, participants from Thailand would like the *Femi* student significantly more as a friend, and in one of the comparisons, participants from Japan would like the *Femi* student as a friend. We believe this may be due to the importance of modesty in these countries; the *Femi* student showed more respect to the professor than the *Mascu* student [which is also reflected in the comments: for example, "The reason was personal, but the student acted respectfully" (participant from Japan)].

While the quantitative data do not confirm our original expectations, the qualitative data do show information which is largely aligned with those expectations. Some participants from countries that score high on Masculinity stated that the *Mascu* professor "acted according to the rules" (participant from Germany), and "the professor made a fair decision" (participant from Japan). Participants from countries that score high on Femininity stated that the *Mascu* professor "is a bit rude" (participant from the Netherlands) and "should not judge too soon" (participant from Thailand). In comparison, the *Femi* professor was judged "kind" and a "nice man" by participants from countries that score high on Femininity, while he was judged "too soft" and "not fair" from participants from countries that score high on Masculinity.

This discrepancy between the qualitative and quantitative data may be due to certain elements, unknown to us, that have a large effect on the perceptions of the participants, but are not captured by the closed questions. By going through the qualitative data, we found that certain comments were made quite often and by participants from every country. They mainly had to do with the student's reason for needing an extension and the decision of the professor:

1. The student should have known in advance that she would need to attend a family event/need more time, so she should have asked sooner;
2. Giving an extension is not fair towards others; the professor should give everybody an extension if he gives it to a single student.

Some examples of the types of comments that frequently appeared are as follows: "I think he and she acted appropriately. Her reason is good for extending the deadline, so I feel his decision is OK. But his decision is not fair to other students" (participant from Japan); "While his decision is nice, it is not really fair towards the others, especially since she did not ask in advance but confronted him with the problem after the deadline" (participant from Germany); "The student should have asked for an extension earlier, not on the day the project is due. In that case, she could have worked around it if the professor said no. I understand the professor does not extend her deadline because that would be unfair towards other students. In particular, since the student is pretty late with asking for an extension, I think he is right" (participant from the Netherlands).

The above discrepancies show a clear mismatch between our intentions and the participants' perceptions. Table 8 shows this divide in terms of differences in perception and intention. We consider the researcher in our case to be the designer and the users to be the participants. This table is inspired by "Johari windows", a simple two-by-two matrix by Joseph Luft and Harry Ingram, originally

Table 8 Designer versus user: differences in perception or intention

	Designer	
User	Intended	Unintended
Perceived	Known by both	Hidden user-context
Unperceived	Unperceived by user	Unknown by both

created to better understand misunderstandings in interpersonal relations (Luft 1970).

Ideally, the "known by both" area should contain as many elements as possible. This reduces the risk that the results will diverge from the initial expectations. However, when dealing with the culture-specific appropriateness of behaviours, it is unlikely that a designer is able to guarantee this. This is, in part, due to the influence of culture on both the designer and the user and is especially true when there is a difference in cultural background between designer and user.

In our experiment, we expected that certain behaviours would be representative of prototypical Masculine and Feminine behaviour. Instead, we found that we might have targeted a different dimension of culture: Individualism versus Collectivism. This represents a typical problem that occurs in the "*hidden user-context*" and "*unperceived by user*" areas; users perceived elements that we did not think were important (for example, the modesty of the student) and might not have perceived elements that we thought were important (for example, the specific non-verbal behaviours of the characters, which were not mentioned in the comments).

The student's reason for needing an extension and the professor's decision whether or not to give an extension were two important factors influencing perception. In particular, the fact that the student should have told the professor in advance was an element that we did not consider to be important (*hidden user-context*). Even though we expected that the fairness of an extension would be an important element of the interaction (*known by both*), we did not expect that the participants from each country would perceive them similarly.

It is possible that the effect of culture does not apply as strongly to the situation in the scenarios: personal experiences of the participants may have influenced their judgements, or these situations might feel unnatural to the participants. The fact that we did not vary the gender of the characters was to keep the results stable. As Masculinity versus Femininity has a large effect on the perceptions of gender roles, we would have to include another set of four scenarios. Since we are not interested in the specific perceptions of gender roles, we decided to keep the gender of the character static. However, we found that none of the participants remarked on the sex of the protagonists in the written comments.

5 Experiment 2

The first experiment showed that participants perceived the characters significantly different if the student did or did not get an extension. Our intention was to see whether people from different cultures would perceive the student significantly different if she had a culturally appropriate reason for needing an extension. We found that the participants did mention the student's reason in the comments, but the quantitative data did not reflect this.

Another element that may have had an influence on perception could have been the reference to family (Femi student). In Collectivistic countries, people are integrated into strong, cohesive groups, and in Individualistic countries, people are supposed to take care of themselves. By removing the reference to family, we hope to find differences between participants that score high and low on Masculinity.

In short, for this second experiment, we decided to do a follow-up study investigating two elements:

- The influence of the student's reason for needing an extension on the perception of the entire interaction;
- Whether we can target a different dimension of culture (Masculinity vs. Femininity) by changing the reason for needing an extension (by removing the emphasis on family).

5.1 Method

We added two new scenarios to the original four. In terms of behaviour, these two scenarios are identical to scenarios with the *Femi* student. There was only one difference: instead of the student needing an extension because she had to attend a family event, the student in the two new scenarios needs an extension because of a computer breakdown. In the remainder of this article, we will refer to this *Alternative Femi* student as the student with the *FemiAlt* script. For more information, see the online study.[3]

5.2 Evaluation procedure and hypotheses

The same experimental setup as in the first experiment was used. Participants saw six videos (the original four videos and the two new videos, see Table 9).

Our hypotheses for the second experiment were:

[3] http://mm-werkstatt.informatik.uni-augsburg.de/survey/index.php?sid=44443&lang=en.

Table 9 The student's reason, the professor's response and the scenario outcome for each of the six scenarios (the student gets no extension in the scenarios that are italicized)

Scenario			Reason for needing an extension	Response of professor	Does the student get an extension?
No.	Student	Professor			
1	*Femi*	*Mascu*	Family event	Needing an extension because of attending a family event is a weak excuse	No
2	Mascu	Mascu	Performance	Wanting to perform better is a good reason for needing an extension	Yes
3	Femi	Femi	Family event	Attending a family event is a good reason for needing an extension	Yes
4	Mascu	Femi	Performance	Professor does not understand, but will give an extension if its important to the student	Yes
5	*FemiAlt*	*Mascu*	Computer breakdown	Needing an extension because of a computer breakdown is a weak excuse	No
6	FemiAlt	Femi	Computer breakdown	A computer breakdown is a good reason for needing an extension	Yes

2.1 Participants from countries that score higher on Masculinity will be more likely to consider the behaviour of characters with the *Mascu* script more appropriate than characters with the *FemiAlt* script.

2.2 The scenarios with the *FemiAlt* student will be perceived significantly different from the scenarios with the *Femi* student.

In addition to the questions used in the first experiment, we included two open questions to gain a greater understanding of what the participants consider appropriate behaviour:

- What do you think a good teacher would have done?
- Do you think the student could have finished the project on time?

5.3 Results

In total, 81 participants of 31 different nationalities took part in our second study. Since we only collected enough data of participants from six countries for statistical analysis, we only considered the data from those participants. In that manner, five people from France (two females; mean age: 34.20; SD age: 7.6), five people from Egypt (three females; mean age: 21; SD age: 0.71), 10 people from Germany (five females; mean age: 34.00; SD age: 7.8), six people from Russia (four females; mean age: 29.83; SD age: 13.26), nine people from the UK (three females; mean age: 42.11; SD age: 14.06) and 14 people from the USA (10 females; mean age: 45.64; SD age: 16.23) were included for analysis, while 32 participants from 25 different countries were excluded. The scores for the six participating nationalities on Hofstede's dimensions are provided in Table 10. As the data were not normally distributed, we used nonparametric tests in all cases.

Table 10 Number of participants from each country and the scores for these countries on Hofstede's dimensions (from Hofstede et al. 2010)

	N	PDI	IND	MAS	UAI	LTO	IvR
France	5	68	71	43	86	63	48
Germany	10	*35*	67	55	65	*83*	40
Russia	6	*93*	*39*	*36*	*95*	81	20
UK	9	*35*	89	*66*	*35*	51	*69*
USA	14	40	*91*	62	46	26	68
Egypt	5	80	38	53	–	*7*	*4*
Difference between lowest and highest		58	52	30	60	76	65

Highest and lowest scores are italicized

To determine whether the influence of a different reason for needing an extension created significant differences in user perceptions, we used the Wilcoxon signed-rank test to compare the two new scenarios (*FemiAlt* student and *Mascu* professor; *FemiAlt* student and *Femi* professor) to the original scenarios with the *Femi* student (*Femi* student and *Mascu* professor; *Femi* student and *Femi* professor).

We only found significant differences for participants from the UK (Table 11) and Germany (Table 12). Participants from the UK found the student and the professor more appropriate in the scenario with the *FemiAlt* student and the *Femi* professor than in the scenario with the *Femi* student and the *Femi* professor. Participants from Germany found the extension fairer to others with the *Femi* student and the *Mascu* professor. They found the *FemiAlt* student more appropriate than the *Femi* student in the scenarios with the *Femi* professor. They also thought the extension was fairer to other students with the *FemiAlt* student and *Femi* professor combination.

Table 11 Comparison of selected questions between the *Femi* student and the *FemiAlt* student for participants from the UK

Scenarios to compare				Country for which the scenarios were compared	
Scenario A		Scenario B		UK	
Stud	Prof	Stud	Prof	Student appropriateness	Professor appropriateness
Femi	*Mascu*	*FemiAlt*	*Mascu*	–	–
Femi	Femi	FemiAlt	Femi	Scenario B	Scenario B
				$p = 0.016$	$p = 0.010$

Named scenario was rated significantly higher (the student gets no extension in the scenario that is italicized)

Table 12 Comparison of selected questions between the *Femi* student and the *FemiAlt* student for participants from Germany

Scenarios to compare				Country for which the scenarios were compared	
Scenario A		Scenario B		Germany	
Stud	Prof	Stud	Prof	Student appropriateness	Extension fair to others
Femi	*Mascu*	*FemiAlt*	*Mascu*	–	Scenario A
					$p = 0.039$
Femi	Femi	FemiAlt	Femi	Scenario B	Scenario B
				$p = 0.026$	$p = 0.034$

Named scenario was rated significantly higher (the student gets no extension in the scenario that is italicized)

To determine whether the behaviour of the *Mascu* characters was considered more appropriate than the *FemiAlt* characters, we used the Wilcoxon signed-rank test to compare the new scenarios to the old scenarios (*FemiAlt* student and *Mascu* professor versus *Mascu* student and *Mascu* professor; *FemiAlt* student and *Femi* professor versus *Mascu* student and *Femi* professor).

We only found significant differences for participants from Egypt. They considered the appropriateness of the student significantly higher with the *FemiAlt* script ($Mdn = 6$) than with the *Mascu* script ($Mdn = 3$) $T = 0$, $p = 0.042$.

Looking at the participants' comments on the scenarios with the *FemiAlt* student gave additional insight into the participants' choices. Interestingly, four out of the five French participants stated explicitly in the scenario with the *FemiAlt* student and *Femi* professor that the professor should not have given an extension due to a computer breakdown (e.g. "A good teacher cannot give an extension for no reason. Here the teacher cannot be certain of the reason the student gave"). Interestingly, the same four French participants gave a similar reasoning for the *Femi* student and *Femi* professor combination. In comparison, four out of six Russian participants, as well as four out of five Egyptian participants, argued that in this scenario the professor was correct in giving an extension.

Russian participants were quite consistent on their comments on the scenario with the *FemiAlt* student and the *Mascu* professor; four out of six participants stated that the professor should have given an extension (e.g. "He could understand everything and offer to redo the project", or "he would give her more time"). The same trend can be observed in the Egyptian data, where four out of five participants stated that the professor should have given the extension (e.g. "he would have extended the deadline as it is a technical problem, the student has no hand in it").

5.4 Discussion

We expected that participants from countries that score high on Masculinity would consider the behaviour of characters with the *Mascu* script more appropriate than characters with the *FemiAlt* script (Hypothesis 2.1).

We were unable to confirm this hypothesis. The qualitative data do show that Individualism versus Collectivism still plays a strong role (Egypt and Russia vs. France).

We expected that the scenarios with the *FemiAlt* student will be perceived significantly different from the scenarios with the *Femi* student (Hypothesis 2.2).

We found that there is a significant difference in perception of appropriateness between participants from Germany and the UK. In particular, the participants from Germany found the *FemiAlt* student more appropriate than the *Femi* student when the student was interacting with the *Femi* professor. The participants from the UK found both the student and the professor more appropriate with the *FemiAlt* student and the *Femi* professor combination than the *Femi* student and *Femi* professor combination. These results suggest that the appropriateness of behaviour is not judged primarily based on the visible behaviour of the characters, but more on the user's interpretation of the character's motivation, notably the student's reason for needing an extension, and the professor decision whether or not to give an extension.

We are aware of the small sample size, in combination with the many judgements each participant had to do. However, our aim was not to do theory testing, but to explore the difference between the perception of users and the intentions of designers.

6 Conclusion and recommendations

In this article, we considered culture-related differences in behaviour to create four, and later six, different scenarios in which two virtual characters interact. The behaviour of these characters was intended to resemble prototypical behaviour from countries that score high or low on the cultural dimension of Masculinity (Hofstede et al. 2010). By showing these scenarios to participants of different nationalities, we investigated their perceptions.

Results from our first experiment indicate that participants did judge the behaviour in the scenarios to be significantly different from each other, but not as we expected. We found in the first experiment that participants from countries that score high on Individualism judged the behaviour of the characters significantly different from participants from countries that score high on Collectivism. In the second experiment, we introduced two more scenarios that are less likely to be influenced by the Individualism dimension. We found that participants from Masculine countries considered the characters in the new scenarios significantly more appropriate than the Feminine characters in some of the old scenarios.

The study allowed us to formulate recommendations for researchers that aim to design culture-specific behaviour for virtual characters. This was possible despite small sample sizes, because of the variety of countries and continents, and the answers to the open-ended questions. A larger-scale user test would be valuable but costly and not necessarily more productive. Moreover, a new test is required for every modification, as experiments 1 and 2 have shown.

The recommendations are based on elements that appear in the "Hidden user-context" (unintended by designer) and "Unperceived by user" areas of Table 8. We consider it important to…

1. …test whether participants from different nationalities perceive the behaviour of virtual characters with different cultural scripts differently;
2. …test hypotheses with a wide variety of cultures represented, instead of a large number of participants from a limited variety of cultures;
3. …include open-ended questions in user tests to discover hidden-user-context issues that may not become apparent from closed-ended questions;
4. …test whether the intended appropriate culture-specific behaviour is actually considered to be appropriate by the target audience;
5. …test whether different elements of the content and context, even those inconspicuous to the designer's mind, affect the users' perception of the scenario as a whole.

Acknowledgments This work was partially supported by European Community (EC) and was funded by the ECUTE (ICT-5-4.2 257666) project. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC or the FCT, which are not responsible for any use that might be made of data appearing therein. We would like to thank John Mc Breen for proofreading the article.

References

Aylett R, Vannini N, Andre E, Paiva A, Enz S, Hall L (2009) But that was in another country: agents and intercultural empathy. In: Proceedings of 8th international conference on autonomous agents and multi agent systems. International Foundation for Autonomous Agents and Multiagent Systems, Budapest, Hungary, pp 10–15. Retrieved from http://dl.acm.org/citation.cfm?id=1558058

Byrne D (1971) The attraction paradigm. Academic, New York, p 474

Damian I, Endrass B, Huber P, Bee N, Andre E (2011) Individualized agent interactions. In: Proceedings of the 4th international conference on motion in games, pp 15–26

Endrass B, André E, Rehm M, Lipi AA, Nakano Y (2011a) Culture-related differences in aspects of behavior for virtual characters across Germany and Japan. In: Proceedings of the 10th international conference on autonomous agents and multiagent systems, pp 441–448

Endrass B, Rehm M, André E (2011b) Planning small talk behavior with cultural influences for multiagent systems. Comp Speech Lang 25(2):158–174

Hofstede GJ, Pedersen P, Hofstede G (2002) Exploring culture: exercises. Nicholas Brealey Publishing, Stories and Synthetic Cultures

Hofstede G, Hofstede GJ, Minkov M (2010) Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival. McGraw-Hill, New York

Jan D, Herrera D, Martinovski B, Novick D, Traum D (2007) A computational model of culture-specific conversational behavior. In: Proceedings of 7th international conference on intelligent virtual agents. Springer, Berlin. doi:10.1007/978-3-540-74997-4_5

Johnson WL, Valente A (2008) Tactical language and culture training systems: using artificial intelligence to teach foreign languages and cultures. In: Proceedings of 20th national conference on innovative applications of artificial intelligence. AAAI Press, Chicago, Illinois

Johnson WL, Marsella S, Vilhjálmsson H (2004) The DARWARS tactical language training system. In: Interservice/industry training, simulation and education conference, pp 1–11

Kim JM, Hill RW, Durlach PJ, Lane HC, Forbell E, Core MG, Marsella SC, Pynadath D, Hart J (2009) BiLAT: a game-based environment for practicing negotiation in a cultural context. Int J Artif Intell Educ 19:289–308

Koda T, Rehm M, André E (2008) Cross-cultural evaluations of avatar facial expressions designed by western designers. In:

Proceedings of 8th international conference on intelligent virtual agents, pp 245–252

Koda T, Ruttkay Z, Nakagawa Y, Tabuchi K (2010) Cross-cultural study on facial regions as cues to recognize emotions of virtual agents. Cult Comput 6259:16–27

Luft J (1970) The Johari Window: a graphical model of awareness in interpersonal relations. In: Group processes: an introduction to group dynamics. National Press Books, Palo Alto, pp 11–20

Mascarenhas SF, Dias J, Afonso N, Enz S, Paiva A (2009) Using rituals to express cultural differences in synthetic characters. In: Proceedings of 8th international conference on autonomous agents and multi agent systems. International Foundation for Autonomous Agents and Multiagent Systems, Budapest, Hungary

Smith PB (2006) When elephants fight, the grass gets trampled: the GLOBE and Hofstede projects. J Int Bus Stud 37(6):915–921