Perceptions of Artificial Social Agents

The cultural similarities and differences between Dutch and Chinese speakers in their perception of artificial social agents

Johan Hensman[1]

Supervisor(s): Willem-Paul Brinkman[1]**, Nele Albers**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 27, 2023

Name of the student: Johan Hensman
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Nele Albers, Odette Scharenborg

Abstract

Artificial social agents (ASAs) are systems designed to interact with humans in a socially intelligent manner. As the field of robotics is rapidly advancing, some studies focused on creating more effective agents by analysing how people perceive them. However, culture affects people's perception of ASAs. So, cultural aspects must be examined in order to create more effective ASAs. This study aims to contribute to the research of cultural influence on ASAs by discovering the cultural similarities and differences in the perception of ASAs from Dutch and Chinese speakers. An English questionnaire has been made to evaluate ASAs and also has been translated into Chinese. In this study, the questionnaire is translated into Dutch and validated using both the original and the translated questionnaire. While assessing the translation, data is gathered from Dutch speakers on a human-ASA interaction. This data is then compared to the previously collected data from the Chinese translation study by calculating their mean, standard deviation and t-distribution of the differences between both cultural groups. On an item level, the results show a satisfactory level of correlation (ICC M = 0.67, SD = 0.12, range [0.36, 0.92]). On a construct/dimension level, results show a good level of correlation (ICC M = 0.80, SD = 0.11, range [0.52, 0.93]). Correction values for the translation have been advised for converting item scores between the two questionnaires. Cultural differences have been found and reported between Dutch and Chinese speakers, which could be used in future research on creating more effective ASAs.

1 Introduction

Artificial social agents (ASAs) are systems designed to interact with humans in a socially intelligent manner. These agents can range from text-based conversational agents to humanoid robots and are capable of simulating human-like behaviour and holding meaningful conversations [11]. As the field of robotics is rapidly advancing, ASAs are demonstrating potential in various fields, such as education, healthcare and entertainment [19]. They can provide virtual coaching in teaching a new language or trying to help you stop smoking [1], [16]. Some studies focused on creating more effective agents by analysing how people perceive them [1]. However, researchers have seen that culture has a significant influence on an individual's perception of ASAs [21], [22], [24]. So, cultural aspects should be examined to create more effective ASAs.

Previous research focused on creating more effective agents by analysing people's perceptions. So it is apparent that there is a need to evaluate how people perceive ASAs. In pursuit of this, Fitrianie et al. [11] constructed a standardized manner to evaluate ASAs. From empirical studies, Bruijnes et al. [4] found multiple questionnaires encompassing at least 100 different constructs. Because it is not feasible to evaluate all of them, Fitranie et al. [13] fused the constructs into 19 distinctive ones which covered 80% of them. They also generated a set of items for each of the constructs and validated them by conducting surveys and calculating a construct validity analysis [11], [12].

This questionnaire is sufficient for a standardized manner to evaluate ASAs, but is not fully capable of assessing the cultural differences in people's perceptions of ASAs as it is only available in English. For research conducted in another language or research designed for comparing perceptions of different cultures, the questionnaire should be translated into their respective languages [5]. Using these questionnaires we can evaluate how people perceive ASAs from multiple countries and a standardized way to compare cultural differences can be achieved.

So, translating the original questionnaire to another language is needed for cross-cultural comparisons [5]. The questionnaire is already translated and validated into Chinese [18], but the Dutch version is not available. This paper will therefore first focus on the translation of the original English questionnaire to Dutch. Secondly, as culture has an influence on the perception of ASAs [21], the comparison between the perception of Dutch and Chinese speakers will be explored. The main research question for this paper is formulated as follows: What are the cultural differences and similarities between Dutch and Chinese speakers in their perceptions of ASAs? This research question can be divided into the following sub-questions: What is the quality of the questionnaire that was developed? To what extent do the ratings differ from Dutch and Chinese speakers on a construct level? Together, the answers to these sub-questions will provide more knowledge on cultural similarities and differences between Dutch and Chinese speakers in their perception of ASAs, which will be key contributions to future research.

2 Related background

This section elaborates on research that has been previously done. This includes further information on the questionnaire that was created, translation strategies and existing research on Dutch and Chinese speakers' views on ASAs.

2.1 ASA questionnaire

As previously stated in the introduction, Fitrianie et al. [11] focused on making a standardized method to validate human-ASA interaction with a questionnaire. This questionnaire contained 90 items that were categorized into 19 constructs with two constructs having multiple dimensions. Dimensions are aspects of a construct that can be evaluated independently, e.g., construct 19 of the questionnaire 'Emotional Experience' is divided into the 'Agent's Emotional Intelligence presence' and the 'User's Emotion Presence'. Each construct/dimension has three to six items, and each item is rated on a 7-point scale [-3,3]. A short version of the questionnaire has also been created to acquire a general idea of an agent. This version of the questionnaire consists of 24 items representing each construct or dimension [12].

2.2 Translation strategy

A questionnaire should be translated to be used properly in cross-cultural research [5]. A strategy should be implemented to create a high-quality questionnaire instead of using a machine translation method. Mondal et al. [20] showed that translations from 'Google Translate' showed more errors than the translations made by human experts. Rabin et al. [23] documented the evolution of creating language adaptations of questionnaires and the more recent procedures. This study follows the first steps of the more recent procedure by creating forward translations and reconciling them. However, backward translation is omitted and the reconciled forward translations are immediately tested.

2.3 Cultural influence on the perception of ASAs

Researchers have seen that the perception of ASAs is influenced by an individual's cultural background [21]. Mascarenhas et al. [19] conducted cross-cultural research between the Netherlands, an individualistic country, and Portugal, a collectivistic country. They showed that people from the Netherlands had a more negative view of collectivistic agents than people from Portugal. People from both countries reacted similarly positively to individualistic agents. In addition, other research from Diana et al. [8] showed that Dutch individuals had a less positive attitude towards robots compared to Japanese individuals.

Bartneck et al. [2] showed that the perception of ASAs of Chinese individuals is similar to the perception of Dutch individuals. The attitudes of the Dutch and Chinese participants towards the interaction and social influence of robots and the emotions in interaction with robots were similar. On the contrary, in the same research, they showed that American participants had a more positive attitude towards robots, while Evers et al. [9] indicated that Chinese individuals reacted more positively to robots than American individuals and even interacted with robots as if they were humans.

3 Method

In the following section, the method and the setup of our experiment are explained, including the participants, materials and data analysis.

3.1 Design and Procedure

The first part of this research focused on creating a Dutch translation for the English questionnaire that evaluates interaction with ASAs. As this paper is similar to the Chinese translation of the questionnaire, the approach is based on their approach [18]. Figure 1 illustrates the steps taken to obtain the translation. For the first step, experts from TU Delft translated the original English questionnaire into Dutch. The two translations were then submitted to a separate expert who examined both of them. From the two translations, they opted for each item in the questionnaire the translation that suits the dimension/construct the most.

For the validation of the translation, in step 2 a survey had been set up using both the newly translated and the original English questionnaire. The survey started with a 30-second video segment of an ASA, the Honda robot ASIMO, interacting with people. The bilingual participants rated this interaction twice using the original and the translated questionnaire. The original questionnaire consists of 90 items, so the translated Dutch version also contained 90 items, resulting in 180 questions in total. To reduce fatigue for the participants, the survey was split in half for the first evaluation round.

Both sub-questionnaires were set up the same way. All of the items in the original questionnaire are divided into 19 constructs, where two constructs have two and five dimensions [11]. To divide them equally into two parts for the sub-questionnaires, each of them contained half of the total constructs/dimensions. However, questions covered in the same construct were not split, which led to a slight difference in the number of questions in the two sub-questionnaires. Both questionnaires also incorporated attention checks in the Dutch and in English questionnaires to verify whether they were answering the questions attentively and truthfully. The specification for the number of questions can be found on the OSF form [15].

The correlation between the English items and their corresponding translation was calculated in step 3, which is explained in greater detail in subsection 3.4 "Data analysis". At the end of the survey, the participants were asked if they answered the questions carefully and would recommend their data to be used. We calculated the correlation twice, once using all the data and once using only the data the participants recommended. In both cases the items with a low correlation value needed to be translated again. This implies, whenever an item had a high correlation value in one analysis, but a low correlation value in the other analysis, the item would be classified as a poor translation and should be retranslated.

In step 4, the items that needed retranslation were sent back to the experts including the chosen translation and the correlation value from the first round of translation. The two experts could see how well the translation was received and adjust the translation accordingly. The new translations were then sent, similar to step 1, to a separate expert who reconciled the translations. The expert could either opt for one of the suggested options or give alternative translations. If more than one translation was chosen for an item, all selected translations were evaluated.

For the assessment in step 5, a survey had been set up similar to step 3. However, in this questionnaire, only the items with low correlation values and their translations were evaluated. The correlation values between the English and the Dutch items were then calculated. From the items that had multiple translations, only the alternative with the highest correlation value was picked. Finally, in step 7 the results from the first and second translation round were combined. This resulted in the final Dutch translation of the ASA questionnaire, which can be found in the OSF Form [15].

This study had been approved by the university's ethics committee for human research (ID: 116203).

3.2 Participants

To determine the sample size it is conventional to have an 80% power of analysis [3]. This means if an experiment is conducted 100 times, then the results will be significant for

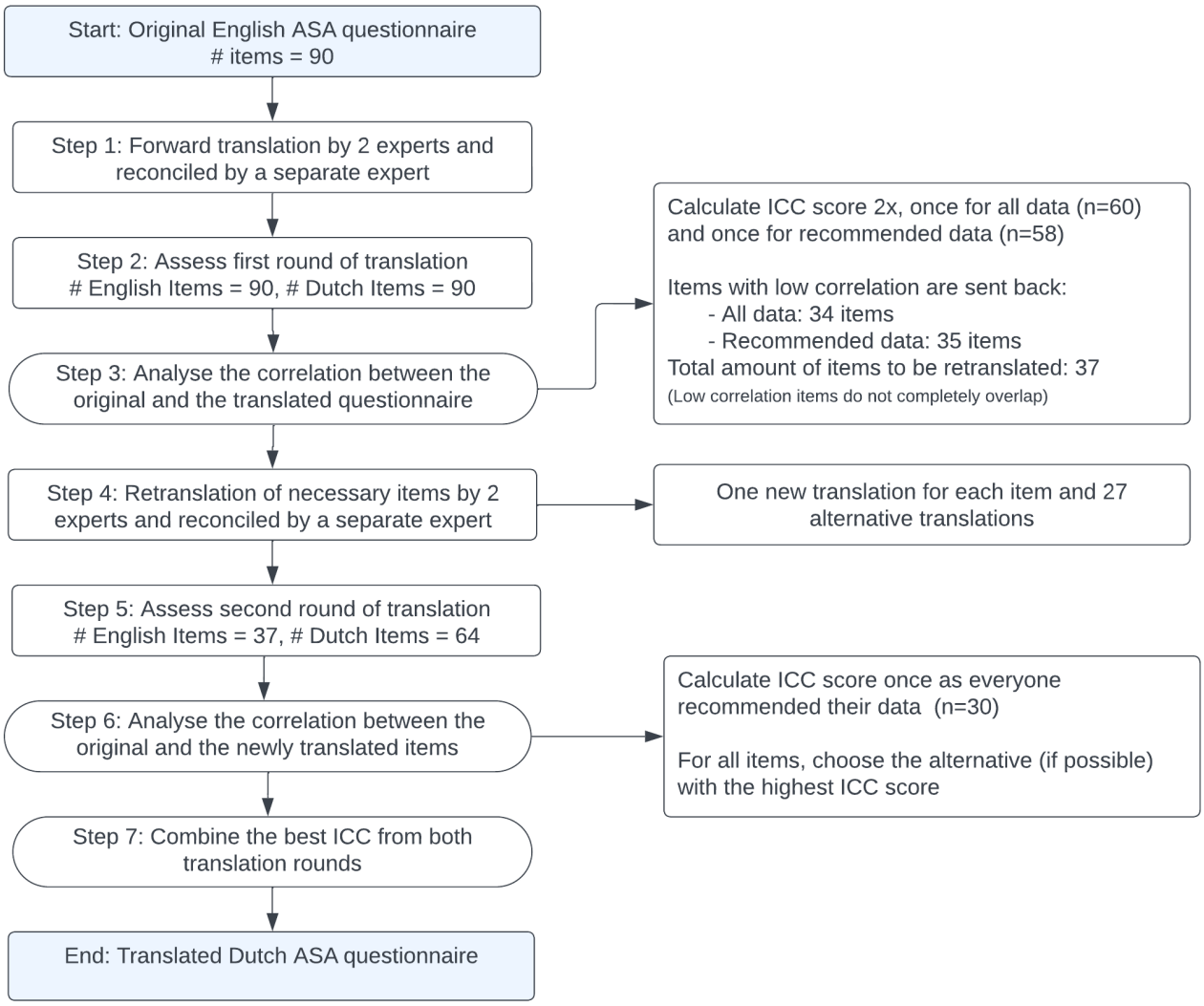

Figure 1: Flowchart for creating the translated questionnaire

at least 80 of them. The Chinese translation study found after running 1000 simulations that around 25 participants are needed to achieve more than 80% power of analysis [18]. To ensure this power is achieved, the sample size was set to 30 for each questionnaire. This means that there were 60 participants in the first translation round when the survey was split in half and 30 participants in the second translation round.

The participants are asked to take part in our study on an online crowdsourcing platform, Prolific. The participants are paid the crowd-platform minimum. The use of Prolific offers multiple benefits, but their pre-screening is the most significant one. As their workers are not able to see the requirements for particular studies and they cannot change their prescreeners immediately [25].

For this study, participants needed to meet specific criteria before being eligible to participate. These criteria included being bilingual, fluent in English and Dutch as their first and primary language. Furthermore, an equal gender ratio of male-female was sought in this study. This was done by creating separate prolific studies where one was intended for males and the other questionnaire for females and others. To be inclusive, transgender people were accepted in both prolific studies.

3.3 Material

The participants were gathered on the online crowdsourcing platform Prolific, where we collected data on their age, gender and highest level of education. The questionnaires are constructed and held on Qualtrics. So, the data of the questionnaire is collected through Qualtrics. The video of ASIMO that was picked for this survey is the same video that was used in the Chinese questionnaire [18]. Fitrianie et al. [13] presented a short description for a list of ASAs with short clips

for each ASA including ASIMO.

3.4 Data analysis

The intraclass correlation coefficient (ICC) is a reliability index that assesses the correlation and the agreements between measurements. It is widely used in test-retest, intra-rater and inter-rater reliability analyses [17]. The approach for this study followed the methodology from the Chinese translation study, which relied on the approach described by Finch et al. [10]. After fitting each item on the questionnaire to a model using the R package nlme v3.1.162, the ICC value was calculated by looking at the variability ratio between the participants' scores and within the participant's scores. The construct/dimension scores were calculated with the mean of the items' scores that belonged to their respective dimension/construct per participant. In the final assessment, ICC values of a construct can only be calculated if the whole construct is rated by the same participant.

Cicchetti [6] gave guidelines for the interpretation of the ICC values. The guidelines stated that ICC values below 0.40 indicate poor reliability, values between 0.40 and 0.59 indicate fair reliability, values between 0.60 and 0.74 indicate good reliability and values above 0.75 indicate excellent reliability [6]. During the translation rounds, the values that had fair or poor reliability were considered low correlation values. So, we set up a cut-off point of 0.6, where all items below this point had to be retranslated. The same guidelines from Cicchetti [6] were used in the summative assessment of the translation.

We also calculated in the same manner as Li et al. [18] the mean, standard deviation and 95% Credibility Intervals (CI) of the *t*-distribution of the mean differences between the Dutch and the English questionnaires. This was done by using the Bayesian paired *t*-test from the R package BayesianFirstAid v0.1. The 95% CI indicates the posterior distribution that contains the central 95% of the values [7]. We can speak of a bias if the interval of the mean difference does not contain zero, which means it needs conversion correction for future use.

Lastly, we calculate the differences in the English ASA questionnaire scores between the Chinese speakers and the Dutch speakers. The data from the Chinese speakers have been previously obtained by Li et al. [18], where participants of their survey rated human-ASA interaction using the same video of ASIMO as this study. Based on the approach of Li et al. [18] using the R package Rethinking v2.31, we can calculate the mean, standard deviation and a 95% CI of the mean differences of both groups by fitting a linear model to obtain the Gaussian distribution on each construct/dimension. We can again speak of a bias when the CI does not contain zero. If the zero is not included and the interval is negative, we can conclude there is a negative bias and if the interval is positive, there is a more positive bias.

All data sets, analysis scripts and outcomes files are online available on 4TU.ResearchData[1] [14]. The translated Questionnaire can be found on the OSF Form [15].

4 Results

In this section, the results of the experiment are covered.

4.1 Correlation between the Dutch and the English ASA Questionnaires on item level

In the first translation round, the mean ICC value of all 90 items and all participants showed a decent level of correlations with a mean of 0.61, a standard deviation of 0.17 and all values in a range of [0, 0.92]. As you can see in Table 1, around 62% of the items had an excellent or good correlation, while 38% had a fair or poor correlation that should be sent back to the translators. When we only look at the data that the participants recommended themselves, we can see a slight difference. The mean of the ICC values becomes 0.62, the standard deviation 0.17 and the values are in a range of [0.01, 0.92]. The percentage that should be sent back to the experts becomes 40%. However, multiple items shifted classifications when we used only the non-recommended data where some items moved from good to fair and vice versa. So, the number of items that should be translated increased from 34 to 37 (from 40% to 41%).

For the 37 items that needed retranslations in the second round, we received 37 new and 27 alternative translations, so 64 translations in total. From the survey, it follows that 34%

[1]For now the code for the summative assessment can be found here: DOI: 10.5281/zenodo.8079921

Table 1: Categories of ICC classifications and number of ICC values in classification category of the first translation round

Classification	ICC Range	90-item set (All data)	90-item set(Recommended data)
Excellent	0.75-1.00	20 (22.22%)	23 (25.56%)
Good	0.60-0.74	36 (40.00%)	32 (35.56%)
Fair	0.40-0.50	21 (23.33%)	26 (28.89%)
Poor	0-0.39	13 (14.44%)	9 (10.00%)

Table 2: Categories of ICC classifications and number of ICC values in classification category of the second translation round

Classification	ICC Range	90-item set (All alternatives)	90-item set(Best set)
Excellent	0.75-1.00	5 (7.81%)	4 (10.81%)
Good	0.60-0.74	17 (26.56%)	13 (35.14%)
Fair	0.40-0.50	28 (43.75%)	16 (43.24%)
Poor	0-0.39	14 (21.88%)	4 (10.81%)

Table 3: Categories of ICC classifications and number of ICC values in classification category after combining the rounds

Classification	**ICC Range**	**90-item set**	**Construct/ Dimension**	**24 item set**
Excellent	0.75-1.00	24 (26.67%)	18 (75.0%)	7 (29.17%)
Good	0.60-0.74	47 (52.22%)	4 (16.67%)	12 (50.0%)
Fair	0.40-0.59	16 (17.78%)	2 (8.33%)	5 (20.83%)
Poor	0-0.39	3 (3.33%)	0 (0.0%)	0 (0.0%)

had good or excellent correlation and 66% had either fair or poor correlation (Table 2). However, these percentages contain all data and not only the 37 we need. When only taking the best alternatives for each item, we obtain a slightly better distribution with 46% having good or excellent classification and 54% fair or poor classification. The set with the best ICC values had a mean of 0.57 and a standard deviation of 0.15.

After the second translation round, we combined the highest ICC values from the first and the second translation round. This resulted in a good correlation where 79% of the items had a good or excellent correlation and 21% a fair or poor correlation (Table 3) with a grand mean of 0.67, a standard deviation of 0.12 and a range of [0.36, 0.92]. For the construct/dimension level, the ICC values could only be calculated if the whole construct was rated by the same participant. Only one construct 'Interaction Impact on Self-Image' was surveyed in its entirety in round 2 and thus the only construct of round 2 that was used in the calculation of ICC values. The correlation of the ICC values on a construct level showed also a good level of correlation with a mean of 0.80, a standard deviation of 0.11 and a range of [0.52, 0,93]. However, two constructs had a poor correlation and four with only a fair correlation (Table 4). Moreover, the short version of the questionnaire showed a decent correlation with 79% having a good or excellent correlation and 21% with a fair correlation (Table 6).

Variation between Dutch and English ASA questionnaire

To look for the variation between the two questionnaires, we calculated the mean, standard deviation and 95% CI of the *t*-distribution of the mean score differences, as they represent how the two questionnaires are similar. This can be seen by the values of the CIs. Whenever zero is not included in the interval, we can speak of a positive or a negative bias. Ta-

Table 4: ICC values and correlation and difference values between the original ASA and the translated ASA questionnaire

		Item		**M**		**Δ**		**CI**	
Construct/Dimension	**ID**	***n***	**ICC**	**Du**	**En**	**M**	**SD**	**2.5%**	**97.5%**
Agent's Believability									
Human-like Appearance	HLA	4	0.83	-1.67	-1.55	-0.08	0.14	-0.36	0.18
Human-like Behaviour	HLB	5	0.89	-1.09	-1.09	-0.00	0.10	-0.20	0.20
Natural Appearance	NA	5	0.51	-1.75	-1.44	-0.29	0.16	-0.59	0.03
Natural Behaviour	NB	3	0.56	-1.90	-1.79	-0.08	0.14	-0.37	0.20
Agent's Appearance suit.	AAS	3	0.62	1.04	1.12	-0.16	0.18	-0.51	0.22
Agent's Usability	AU	3	0.80	0.36	0.68	-0.25	0.16	-0.58	0.07
Performance	PF	3	0.88	0.69	0.60	0.08	0.11	-0.15	0.30
Agent's Likeability	AL	5	0.90	0.40	0.38	-0.02	0.10	-0.22	0.18
Agent's Sociability	AS	3	0.81	-0.48	-0.31	-0.14	0.16	-0.44	0.17
Agent's Personality Prese.	APP	3	0.85	-1.06	-0.98	-0.09	0.12	-0.34	0.15
User Acceptance of the A.	UAA	3	0.84	0.67	0.76	-0.08	0.14	-0.36	0.20
Agent's Enjoyability	AE	4	0.80	0.58	0.95	-0.23	0.11	**-0.45**	**-0.01**
User's Engagement	UE	3	0.84	2.23	2.21	0.03	0.08	-0.12	0.18
User's Trust	UT	3	0.74	0.17	0.07	0.05	0.14	-0.21	0.33
User-Agent Alliance	UAL	6	0.81	-0.16	-0.34	0.18	0.12	-0.06	0.43
Agent's Attentiveness	AA	3	0.82	0.88	0.69	0.21	0.16	-0.11	0.54
Agent's Coherence	AC	4	0.93	0.38	0.48	-0.09	0.11	-0.31	0.13
Agent's Inentionality	AI	4	0.88	0.09	0.06	0.03	0.14	-0.25	0.30
Attitude	AT	3	0.71	1.36	1.28	0.16	0.13	-0.09	0.42
Social Precence	SP	3	0.86	-0.49	-0.26	-0.25	0.16	-0.55	0.07
Interaction Impact on Self.	IIS	4	0.79	0.18	0.21	-0.03	0.13	-0.28	0.23
Emotional Experience									
Agent's Emotional Intell.	AEI	5	0.90	-1.43	-1.56	0.13	0.11	-0.09	0.35
User's Emotional Prese.	UEP	4	0.80	1.21	1.20	-0.02	0.11	-0.24	0.20
User-Agent Interplay	UAI	4	0.74	0.14	0.26	-0.13	0.16	-0.47	0.19
Grand Mean	-	-	0.80	0.01	0.06	0.11	0.13	-	-

Table 5: Items with bias indication

	M		Δ		CI		Max{(PΔ ≥ 0),
Item	**Du**	**EN**	**M**	**SD**	**2.5%**	**97.5%**	**PΔ ≤ 0)}**
HLB4	-0.70	-1.37	0.67	0.26	0.15	1.17	≥0.99
AS1	-0.43	-1.13	0.71	0.26	0.22	1.22	≥0.99
AS3	1.00	0.37	0.57	0.21	0.15	0.97	≥0.99
UAL5	-0.47	-1.10	0.60	0.27	0.06	1.13	0.99
AA3	1.13	0.40	0.69	0.18	0.34	1.06	≥0.99
SP1	0.60	0.20	0.41	0.19	0.02	0.79	0.98
UAI3	-0.37	0.00	-0.36	0.16	-0.69	-0.05	0.99

ble 4 shows a grand mean difference of 0.11, a grand mean of standard deviation of 0.13 and a range of [-0.29, 0.21]. Moreover, it shows that there is almost no bias between the two languages on a construct level, except for one construct 'Agent's Enjoyability (AE)'. It has exclusively negative values in the interval and thus a credible indication of a negative bias. We also analysed the mean score differences on an item level where we found six positive biases and one negative bias indicated in Table 5. Similarly, the same analysis has been done on the short version of the questionnaire. It had the same results as the construct analysis, where there was only one negative bias, namely for 'Agent Enjoyability (AE)' (Table 6).

Comparison of human-ASA interaction between different cultural backgrounds

To compare how users view ASAs from different cultural backgrounds, we analysed the score differences between the two groups on a construct level. As before, whenever the credibility interval did not contain zero, there is a statistical indication of a bias. There were 13 constructs where biases were indicated which can be seen in Table 7. Dutch speakers gave a higher score for User's Engagement (UE), but they gave a lower score for Human-Like Behaviour (HLB), Natural Appearance (NA), Natural Behaviour (NB), Agent's Usability (AU), Performance (PF), Agent's Likeability (AL), Agent's Sociability (AS), Agent's Personality Presence (APP), User's Trust (UT), User-Agent Alliance (UAL), Agent's Attentiveness (AA) and Agent's Coherence (AC).

5 Discussion

This research showed that the long version of the newly translated Dutch questionnaire had a satisfactory correlation on an item level, where 80% had a good to excellent correlation and 20% had a poor to fair correlation. Findings also show that it is preferred that researchers should compare the results of their survey on a construct/dimension level with the English questionnaire. 91.7% of the constructs/dimensions showed good to excellent correlation and 8.3% had fair correlation.

Table 6: The short version of the ASA questionnaire

			M		Δ		CI	
Item	**Question**	**ICC**	**Du**	**EN**	**M**	**SD**	**2.5%**	**97.5%**
HLA2	[The agent] has the appearance of a human	0.42	-1.20	-1.67	0.29	0.28	-0.11	0.87
HLB5	[The agent] has a human-like manner	0.62	-0.37	-0.77	0.32	0.25	-0.18	0.80
NA4	[The agent] seems natural from the outward appearance	0.60	-1.27	-1.47	0.20	0.25	-0.29	0.69
NB3	[The agent] reacts like a living organism	0.43	-1.30	-1.43	0.13	0.31	-0.51	0.72
AAS1	[The agent]'s appearance is appropriate	0.62	1.10	1.10	0.02	0.22	-0.43	0.44
AU1	[The agent] is easy to use	0.75	0.53	0.63	-0.09	0.22	-0.52	0.35
PF1	[The agent] does its task well	0.84	0.37	0.33	0.05	0.16	-0.27	0.38
AL2	I like [the agent]	0.82	0.53	0.43	0.13	0.16	-0.18	0.44
AS1	[The agent] can easily mix socially	0.57	-0.43	-1.13	0.70	0.25	**0.19**	**1.21**
APP1	[The agent] has a distinctive character	0.72	-0.67	-0.80	0.14	0.21	-0.27	0.56
UAA1	The user will use [the agent] again in the future	0.63	0.07	0.33	-0.24	0.25	-0.73	0.25
AE1	[R] [The agent] is boring	0.69	0.27	0.33	-0.15	0.24	-0.62	0.33
UE2	The interaction captured the user's attention	0.80	1.80	1.67	+0.00	+0.00	-0.00	+0.00
UT3	The user can rely on [the agent]	0.79	-0.07	0.07	-0.00	+0.00	-0.00	+0.00
UAL1	[The agent] and the user have a strategic alliance	0.71	-0.37	-0.70	0.33	0.23	-0.12	0.78
AA2	[The agent] is attentive	0.55	0.40	0.53	-0.10	0.30	-0.71	0.49
AC1	[R] [The agent]'s behavior does not make sense	0.73	0.30	0.53	-0.13	0.24	-0.61	0.34
AI3	[R] [The agent] has no clue of what it is doing	0.80	-0.17	-0.07	-0.06	0.22	-0.51	0.36
AT1	The user sees the interaction with [the agent] as something positive	0.62	1.13	1.03	0.11	0.19	-0.27	0.50
SP2	[The agent] is a social entity	0.73	-0.53	-0.27	-0.22	0.25	-0.71	0.28
IIS2	Others would encourage the user to use [the agent]	0.70	0.10	0.33	-0.23	0.21	-0.65	0.18
AEI3	[R] [The agent] is emotionless	0.89	-1.37	-1.60	-0.00	0.00	-0.00	0.00
UEP3	The emotions the user feels during the interaction are caused by [the agent]	0.73	1.53	1.47	0.00	0.00	-0.00	0.00
UAI4	[The agent]'s and the user's emotions change to what they do to each other	0.41	-0.1	0.33	-0.37	0.28	-0.92	0.21
Grand mean		0.67	0.01	-0.03	0.17	0.20	-	-

Table 7: Construct/dimension rating difference between Chinese and Dutch speakers

	M		Δ		CI		Max{(PΔ >0),
Construct/Dimension	**Du**	**Ch**	**M**	**SD**	**2.5%**	**97.5%**	**PΔ <0)}**
Agent's Believability							
HLA	-1.55	-0.90	-0.57	0.39	-1.34	0.20	0.93
HLB	-1.09	0.57	-1.43	0.36	**-2.14**	**-0.71**	≥0.99
NA	1.44	-0.21	-1.10	0.33	**-1.75**	**-0.43**	≥0.99
NB	-1.79	0.42	-2.01	0.30	**-2.60**	**-1.41**	≥0.99
AAS	1.12	0.93	0.18	0.35	-0.50	0.87	0.71
AU	0.68	1.42	-0.64	0.32	**-1.27**	**-0.00**	0.98
PF	0.60	1.27	-0.60	0.29	**-1.17**	**-0.01**	0.98
AL	0.38	1.33	-0.83	0.33	**-1.48**	**-0.17**	0.99
AS	-0.31	1.37	-1.49	0.32	**-2.11**	**-0.86**	≥0.99
APP	-0.98	0.56	-1.36	0.33	**-2.00**	**-0.71**	≥0.99
UAA	0.76	1.22	-0.41	0.30	-1.00	0.17	0.92
AE	0.95	1.13	-0.15	0.32	-0.77	0.47	0.69
UE	2.21	1.29	0.89	0.23	**0.43**	**1.33**	≥0.99
UT	0.07	0.68	-0.57	0.26	**-1.07**	**-0.06**	0.99
UAL	-0.34	0.33	-0.62	0.26	**-1.14**	**-0.11**	0.99
AA	0.69	1.60	-0.78	0.34	**-1.44**	**-0.10**	0.99
AC	0.48	2.18	1.47	0.34	**-2.13**	**-0.80**	≥0.99
AI	0.06	0.39	-0.29	0.34	-0.95	0.38	0.80
AT	1.28	1.50	-0.19	0.28	-0.74	0.36	0.75
SP	-0.26	-0.26	-0.00	0.41	-0.81	0.80	0.50
IIS	-0.03	0.54	-0.53	0.29	-1.08	0.05	0.97
Emotional Experience							
AEI	-1.56	-0.78	-0.68	0.38	-1.43	0.09	0.96
UEP	1.20	0.54	0.59	0.35	-0.10	1.28	0.95
UAI	0.26	0.61	-0.30	0.35	-0.99	0.40	0.80

On average the mean for the ICC value was 0.80 with an average difference of 0.11, which confirms the good correlation level and that it can be used in future research.

Even though the correlation value shows that the translations are close to their corresponding English items, we still present conversion correction for the different items and constructs. If researchers wish to compare the results on an item level, the conversion values are presented in Table 5. For example, 0.67 should be added to item 'HLB4' to get a similar value as its English counterpart. If the researchers wish to compare their results on a construct level, only for the construct 'Agent's Enjoyability' conversion correction of 0.37 should be applied.

The comparison between the Dutch speakers and the Chinese speakers supports further research on inter- and cross-cultural studies on human-ASA interaction. The values in Table 7 also indicate which constructs Dutch speakers think more negatively or positively on. The two constructs 'Agent's Usability' and 'Performance' do not contain zero, but they are rather close to zero than other constructs. In other words, the Dutch speakers' view is close to the view of the Chinese speakers. When taking this data for further research, the extent of the differences must also be taken into account.

An important limitation of this study is the number of translation rounds. Initially, this study targeted three translation rounds, but due to time constraints, it was only possible to conduct two rounds. This means that the final translation of the Dutch questionnaire of this study is rather a preliminary version than a final version. In the Chinese translation study by Li et al. [18], the amount of poor or fair translations dropped significantly in each translation round. It cannot be assumed that the amount of poor/fair translations would also drop significantly if we conducted a third round for the Dutch translations, but this will likely be the case.

Another limitation of this study is the lack of a final survey using the Dutch translation. The ICC values on the construct level could only be calculated if the entire construct/dimension was rated by the same participant as stated in the section 'Method'. However, in round 2 of the translation, only the items with low correlation values were surveyed. A significant portion of those items had higher ICC values, but their impact on a construct level could not be calculated, as most of the items were only part of a construct.

Furthermore, this study was meant to create a translated version of the ASA questionnaire and compare the Dutch and Chinese speakers' views on ASAs. As stated in the introduction, ASAs encompass a wide range of agents, from text-based agents to virtual robots [11]. However, in this study, the methods and analysis were based on the data of one agent 'ASIMO'. Therefore, the results of the surveys and in particular the correlation values that were calculated are based on this agent. It is not guaranteed that the same results can be reproduced when the translated questionnaire is used on other agents.

For future research, we recommend that another round of translation should be conducted to increase the accuracy of the translation. Two constructs, 'Natural Appearance' and 'Natural behaviour', were considered fair according to the guidelines of Cicchetti [6]. Four items had a low correlation level in these constructs: NA2, NA5, NB2 and NB3 in the original questionnaire [12]. So, in the new translation round a special focus on these items must be brought.

Moreover, we recommend that another survey should be done in the same manner as the final summative assessment of Li et al. [18]. They surveyed participants using 14 different agents to validate their translation to get a more generalised view of ASAs instead of basing their results on one agent. We advise the same method as Li et al. [18] as they already collected data on Chinese speakers, so a comparison study between the Dutch and Chinese speakers can be done without having to collect data on Chinese speakers.

Lastly, we also recommend a study that focuses on the Dutch and Chinese cultures. In this study, we found the differences and similarities between the two cultures regarding their view on ASAs. However, no research has been done where the differences are explained. So, to create more effective ASAs, the reasoning behind these scores should be explored and clarified.

6 Conclusion

The first objective of this paper was to create and validate a Dutch translation of the ASA questionnaire. We provided a preliminary version of the translated questionnaire that already showed a satisfactory level of correlation on an item level and a good level of correlation on a construct level. However, in the same approach of this study, another round of translation would be advised with a special focus on the items: NA2, NA5, NB2 and NB3 of the original ASA questionnaire. Secondly, we wanted to see what constructs the Chinese and Dutch speakers would rate differently. Findings showed that Dutch speakers rated the ASA ASIMO more negatively on Human-Like Behaviour (HLB), Natural Appearance (NA), Natural Behaviour (NB), Agent's Usability (AU), Performance (PF), Agent's Likeability (AL), Agent's Sociability (AS), Agent's Personality Presence (APP), User's Trust (UT), User-Agent Alliance (UAL), Agent's Attentiveness (AA) and Agent's Coherence (AC) and more positively on User's Engagement (UE) compared to the Chinese speakers. Using these new findings and the questionnaire, a better understanding has been brought on the perception of ASAs from Dutch and Chinese speakers, which can be used in further research on ASAs.

7 Responsible Research

In this section, the ethical sides of this research are examined. The first section looks at the integrity of the projects and the second section looks at the reproducibility.

7.1 Integrity

This research is based on the data gathered during the experiment. During the whole process of collecting data, the integrity of the data is of utmost importance. That is why we ensured this integrity from collecting to storing to analysing the data. The gathering of data is randomised to protect this research from having selection bias. We only had select requirements for the participants, but we strived for an equal gender-balanced ratio in the participant pool. When the data was gathered, the names of the participants were not collected, but their prolific IDs were. These were only available to our supervisor and were anonymised before they were transferred to us. As transparency is important, the data was published on 4TU.ResearchData [14]. During the experiment, we used all the data that was available. However, the participants had the option to recommend that their data should be excluded from the analysis, as described in the methodology section. We did this to make sure that the data was not skewed in our favour. Furthermore, this research is very similar to the Chinese study [18].

7.2 Reproduciblity

Another important ethical aspect of this research and research in general is reproducibility. This means that the results and the experiment of this paper can be replicated by others. To guarantee this, an OSF form was created. This is an online repository where designs for funded experiments are stored and shared with the public. In this form, the methods are described in great detail before we executed them. In addition to the methodology, the code used for this project is also published on 4TU.ResearchData [14]. The code is extensively documented, including a README.md file and comments that explain each step of the code. Since the code and data are both publicly available, this contributes to the reproducibility of this research.

8 Acknowledgements

This research is part of the multidisciplinary research project Perfect Fit, which is supported by several funders organized by the Netherlands Organization for Scientific Research (NWO), program Commit2Data - Big Data & Health (project number 628.011.211). Besides NWO, the funders include the Netherlands Organisation for Health Research and Development (ZonMw), Hartstichting, the Ministry of Health, Welfare and Sport (VWS), Health Holland, and the Netherlands eScience Center.

References

[1] Nele Albers, Mark A. Neerincx, Kristell M. Penfornis, and Willem-Paul Brinkman. Users' needs for a digital smoking cessation application and how to address them: A mixed-methods study. 10:e13824.

[2] Christoph Bartneck, Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kato Kennsuke. Cultural differences in attitudes towards robots.

[3] Jason Brownlee. A gentle introduction to statistical power and power analysis in python.

[4] Merijn Bruijnes, Siska Fitrianie, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. What

are we measuring anyway? a literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences: 31st benelux conference on artificial intelligence and the 28th belgian dutch conference on machine learning, BNAIC/BENELEARN 2019. pages 1–2. Publisher: CEUR-WS.org.

[5] A. M. Chang, J. P. Chau, and E. Holroyd. Translation of questionnaires and issues of equivalence. 29(2):316–322.

[6] Domenic Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. 6:284–290.

[7] Federico Comotto. Statistics 101: Credible vs confidence interval.

[8] Fabiola Diana, Misako Kawahara, Isabella Saccardi, Ruud Hortensius, Akihiro Tanaka, and Mariska E. Kret. A cross-cultural comparison on implicit and explicit attitudes towards artificial agents.

[9] Vanessa Evers, Heidy Maldonado, Talia Brodecki, and Pamela Hinds. Relational vs. group self-construal: Untangling the role of national culture in HRI. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 255–262. ISSN: 2167-2148.

[10] W. Holmes Finch, Jocelyn E. Bolin, and Ken Kelley. *Multilevel Modeling Using R*. Chapman & Hall, 2nd edition.

[11] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire: Establishing the long and short questionnaire versions.

[12] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, and Willem Paul Brinkman. Questionnaire items for evaluating artificial social agents - expert generated, content validated and reliability analysed: 21st ACM international conference on intelligent virtual agents, IVA 2021. pages 84–86. Publisher: Association for Computing Machinery (ACM).

[13] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. The 19 unifying questionnaire constructs of artificial social agents: An IVA community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8. ACM.

[14] Johan Hensman, Kriss Tesink, Nele Albers, and Willem-Paul Brinkman. Dutch ASA questionnaire translation - translation and formative assessment: Rounds 1 and 2.

[15] Boleslav Khodakov, Emma Bokel, Kriss Tesink, Johan Hensman, Nele Albers, and Willem-Paul Brinkman. German and dutch ASA questionnaire translations - part 1: Translation and formative assessment. Publisher: OSF.

[16] Elly A. Konijn, Brechtje Jansen, Victoria Mondaca Bustos, Veerle L. N. F. Hobbelink, and Daniel Preciado Vanegas. Social robots for (second) language learning in (migrant) primary school children. 14(3):827–843.

[17] Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. 15(2):155–163.

[18] Fengxiang Li, Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, Fu Guo, and Willem-Paul Brinkman. Mandarin chinese translation of the artificial-social-agent questionnaire instrument for evaluating human-agent interaction.

[19] Samuel Mascarenhas, Nick Degens, Ana Paiva, Rui Prada, Gert Jan Hofstede, Adrie Beulens, and Ruth Aylett. Modeling culture in intelligent virtual agents (online first).

[20] Himel Mondal, Shaikat Mondal, and Sarika Mondal. Feasibility of using "google translate" in adaptation of survey questionnaire from english to bengali: A pilot study. 35(2):119.

[21] Mohammad Obaid, Maha Salem, Micheline Ziadee, Halim Boukaram, Elena Moltchanova, and Majd Sakr. Investigating effects of professional status and ethnicity in human-agent interaction. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 179–186. ACM.

[22] Chao Qu, Willem-Paul Brinkman, Yun Ling, Pascal Wiggers, and Ingrid Heynderickx. Human perception of a conversational virtual human: an empirical study on the effect of emotion and culture. 17(4):307–321.

[23] Rosalind Rabin, Claire Gudex, Caroline Selai, and Michael Herdman. From translation to version management: a history and review of methods for the cultural adaptation of the EuroQol five-dimensional questionnaire. 17(1):70–76.

[24] Maha Salem, Micheline Ziadee, and Majd Sakr. Marhaba, how may i help you?: effects of politeness and culture on robot acceptance and anthropomorphization. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 74–81. ACM.

[25] Prolific Team. How do i use prolific's demographic prescreening?

Appendix A

Reported data from Qualtrics and Prolific

Table 8: Reported data from the translation rounds

	Age range	Age mean	Male ratio	Female ratio	Non-Binary ratio	Date start	Date end
Round 1	18 - 63	30.8	0.5	0.5	0	29/5/23	16/6/23
Round 2	20 - 74	30	0.47	0.5	0.33	30/5/23	17/6/23