Cultural Differences and Similarities in Perceptions of Artificial Social Agents Between German and Chinese Speakers

Emma Bokel[1]

Supervisors: 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 25, 2023

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

Abstract

The Artificial Social Agent(ASA) questionnaire serves as a tool to assess the interaction between humans and ASA[2]. In an effort to increase the number of people who can make use of the questionnaire, previous work has been done to translate this questionnaire to Mandarin Chinese[5]. To increase the number of people who can use this questionnaire even more, we have translated it to German using a similar iterative translating process. We had professionals translate the questionnaire and ran 3 rounds of surveys to see what questions needed improvement. Then, there was a final survey that evaluated the German version to see how well it performs and to see how it compares to the English version. In order to evaluate the cultural differences between the Chinese and German translations, the data of both translation evaluations were compared. From these results, it was shown that ASA developers who use the German version of the ASA questionnaire can make comparisons on a construct level to the English version. This was, however, only possible when using the full questionnaire, not the short version. Additionally, we have found that there are differences in how Mandarin Chinese speakers and German speakers rate ASAs.

1 Introduction

1.1 Motivation

After the design of the ASA Questionnaire and its Chinese translation, it is vital to continue translating the ASA Questionnaire into other languages and to be aware of the cultural differences between them. An ASA is a computer-based agent that autonomously interacts with humans in a range of social roles [8]. Several studies have looked into the way people perceive these ASAs to help improve them[3], however, culture and language may affect such perceptions [7]. A questionnaire has been designed to standardize the way ASAs are perceived [2]. After that, the questionnaire was systematically translated into Mandarin Chinese so that these Chinese speakers can also use it [5]. To further extend the usability of the questionnaire, it has now also been systematically translated into German. This translation has been analyzed. Performing this study is useful for creating an optimal translation and knowing how reliable the translations are. From there, the cultural differences and similarities between German and Chinese language speakers were researched so that ASA developers can make their creations more suitable for different Chinese and German speakers. This allows German speakers to be included when evaluating an ASA using the ASA Questionnaire. Moreover, it allows researchers to know how to interpret the German results compared to the English and Chinese results. Thus, in order to allow German speakers to use the ASA Questionnaire and to help understand the cultural differences between German and Chinese speakers, we have translated the questionnaire and evaluated its translation.

1.2 English ASA Questionnaire

The original questionnaire is a list of 90 questions, which can be used for evaluating human interaction with ASAs [2]. The process of creating this questionnaire was elaborate, it took multiple years and more than 100 Intelligent Virtual Agent researchers, to come up with this final list. The 90 questions are split into 19 measurement constructs.

1.3 Translation Steps to Get to the Final Translation

This paper follows a very similar procedure to the Chinese translation steps [5]. That research purposefully laid a foundation for the translation process for additional languages. Therefore, this paper has followed their translation steps. First, the English questionnaire was translated into German by experts, then a crowdsourcing study was run where bilingual participants answered the questions in English and German. From the resulting data, the similarity between the languages was measured. Two extra rounds of retranslation and reevaluation followed this. For the questions that had a weak correlation, more translations were created. Then, from there, the optimal translations of each question were compiled for a more extensive evaluation of the similarity between the German and English translations.

1.4 Comparing Chinese with German

After the final German translation was established, the English questions from the Chinese and German surveys were compared to see if they were similar on both a construct basis. Which leads to the research question: **What are the differences and similarities in perceptions of human-ASA interactions between German and Chinese speakers?**

2 Questionnaire Translation

This paper follows a very similar procedure to the Chinese translation[5] steps where multiple rounds of translation are followed by an analysis of the final results. Rabin et al[6] give a step-by-step approach for translating a questionnaire. Some of the steps they recommend are: 1) Forward translation, where the questionnaire is translated from the original language to a second language. 2) Reconciliation of the forward translation, here the two translations are compared and the translations get improved. 3) Pilot testing the questionnaire to see how well it performs. Steps 1 and 2 are explained in this section, step 3 is described in the methods section.

Round 1: Translators Translate English Questionnaire to German

First, the English questionnaire was translated into German by Native German speakers who are experts in virtual agents. Three people from RWTH Aachen translated each question, along with some attention checks. A fourth expert looked at the three translations and combined them into a new survey. It was difficult to decide on a single translation for some questions, so a few questions had two or three translations.

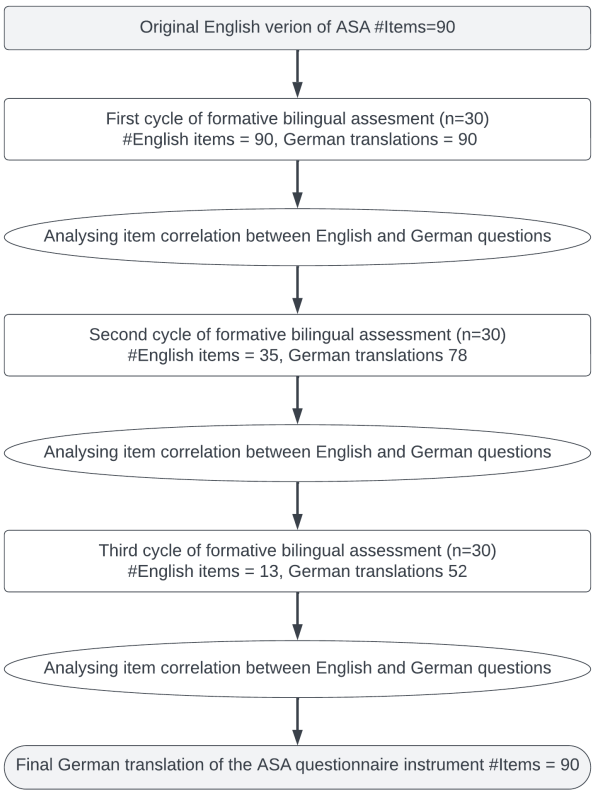

Figure 1: Flowchart of the questionnaire translation procedure

Round 1: Survey

With the translated questions, a survey was created on Qualtircs to verify how similar the German questions were to the corresponding English questions. To minimize fatigue effects, the survey was split into two parts. Each part consists of 12 out of the 24 constructs. The survey was distributed on Prolific. In total, 30 men, 27 women, and 3 non-binary people answered the survey. Their ages ranged from 19 to 69, with a mean of 31 and a standard deviation of 10. The survey consisted of some questions to check if the participants were eligible to participate in the study and a video of an ASA interaction.

Round 1: Code analysis

Based on the code that was used in the Mandarin translation [5], the resulting data was processed. The columns that contained the data about the responses to the original English questions and their German translations were selected. Then, the two sets of data from the two surveys were combined. Next, the intraclass correlation coefficient(ICC) values were calculated, and all questions with an ICC value that was lower than 0.6 were extracted. The value of 0.6 is the same value that was used for the Chinese translation Chinese paper[5], who decided upon this based on the ratings from Cicchetti [1], which classify ICC values starting from 0.6 as good or excellent. The last question of the survey asked the participants whether or not they recommend using their data for a scientific study. There was no difference in the number of questions that needed retranslation if we only included the data of the participants who recommend using their data.

Round 2: Retranslation

Based on the calculated ICC values, 35 questions were retranslated. All ICC values were shared with the translators so that they had an indication of which translations worked well. From there, the translators made a new list of translations for all questions which did not have a good enough translation in the first round. To have a bigger chance of getting a good translation for each question, this time, the questions had two or three retranslations.

Round 2: Survey

The 35 low-ICC questions and the corresponding new translations were put in the second survey. The survey had the exact same setup as the first survey, only this time, there were fewer questions. Because of this, only one Qualtrics survey was needed. The participants consisted of 30 people: 15 men, 13 women, and 2 non-binary people. Their ages ranging from 22 to 70, with a mean of 35 and a standard deviation of 12.

Round 2: Code analysis

Again, the data from the survey was processed in order to get a list of the ICC values that belong to each question, and to extract the questions that had a poor translation. The required columns were selected. The ICC values were calculated and exported. The list of poor translations was only 10 questions long now. The last question of the survey asked the participants whether or not they recommend using their data for a scientific study. There was a difference in the number of questions that needed retranslation if we only included the data of the participants who recommend using their data. But, to follow the Chinese procedure, we included all data to determine which questions needed retranslation.

Round 3: Retranslation

Based on the results from round 2, the 10 remaining questions along with 3 other questions were retranslated. It turned out that there was a small rounding mistake in the code of the earlier rounds, so there were 2 questions that should have been retranslated but weren't. Therefore, these questions were included in this round. Additionally, one question was too similar to a question that already existed in the translation, so, that question was retranslated again as well. The translation process was done in the same way as round 2. Since this was the last round of the Translation Phase, each question had 4 translations.

Round 3: Survey

Based on the new translations, another Qualtrics survey was made. Since there were only 13 English questions, the number of English attention checks was lowered from 7 to 3. Again, 30 people did the survey: 15 men and 15 women. Their ages ranged from 21 to 46, had a mean of 31, and a standard deviation of 6.

Round 3: Code analysis

Based on the round 3 survey, the ICC values were calculated once more. The data was extracted and the ICCs were computed.

Gathering the Final List of Translations
From this iterative process, the best translations were compiled. A list of all translations and their accompanying ICC values was gathered. From there, the translations with the highest ICC values were picked out to be part of the final translation. These ICC values were based on all data, not just the data that was created by participants who recommend using their data. Now that the final translation was established, it was evaluated to see how well it performed and how it compared to the Chinese paper in the Evaluation Phase.

3 Methods

Before the study started, we asked for and received ethical approval from the Human Research Ethics Committee (HREC), the approval number is 3051. After that, the study was preregistered. Then, we created a survey with all the gathered final translations, which we ran and analyzed.

3.1 Design and Procedure

In line with Li et al [5], following the formative phase of creating the translation, a summative assessment was done to evaluate the similarity of the German and English questionnaire. With this data, we also compared the German and Chinese speakers in terms of their perceptions of ASAs by evaluating the differences in their English answers. To obtain said data, a questionnaire was made that asked German speakers to rate a randomly assigned ASA, based on a 30-second video clip. These video clips were the same clips that were used in the Chinese study [5] and in the construct validity analysis of the original questionnaire [2]. The questionnaire was split into two halves to reduce the effects of fatigue. Because the German translations differed based on the gender of the ASAs and users in the videos, there were multiple versions of each set of questions. To make sure the survey takers were paying attention, 14 attention checks were randomly scattered throughout the survey. The survey was run in June of 2023.

3.2 Participants

The survey was completed by 154 participants in total. The first half of the survey was answered by 72 people, and the second half by 82. The gender balance was the following: 72 men, 69 women, and 3 non-binary people. The code relies on the two halves of the survey being of equal length, so, 10 data points from the second half were blindly omitted. Each participant was payed according to Prolific's minimum wage, which is the norm.

3.3 Materials

Each participant was assigned one of 14 videos that represent an ASA. The 14 videos that were used in the videos were: AIBO, AMY, CHAPPIE, DEEPBLUE, DOG, FURBY, HAL 9000, iCAT, NAO, POPPIE, SIM SENSEI, SIRI, SARAH, MARCUS. These are the same videos that were used in the Chinese Survey which chose them because they were used to evaluate the original English version [5]. Just like that Chinese research, the questions were asked in a third-person perspective. This was chosen because it makes the most intuitive sense, and Fitriane et al[2] found few differences between asking first and third-person questions. The survey was administered on Qualtrics.

3.4 Data Preparation and Analysis

Like the Chinese translation[5], the analysis will consist of 3 parts: First, we looked at the similarity of the questions and their translations in terms of their ICC values. Then, we investigated the biases of the translation to see if any adjustments needed to be made. Lastly, we compared the English part of the German data with the previously established English part of the Chinese data [5].

Correlation Between the English and German ASAQ
ICC is widely used in studies of questionnaire/scale translation and validation as a reliability index in test-retest, intra-rater, and inter-rater reliability analyses. We analyzed the data collected using the ICC to determine the correlation between the translated German questionnaire and the original English questionnaire. Cicchetti[1] gave guidelines for the interpretation of ICC inter-rater agreement measures which are cited widely: ICC values less than 0.40 are poor, values between 0.40 and 0.59 are fair, values between 0.60 and 0.74 are good, and values between 0.75 and 1.00 are excellent [1].

Variation Between the English and German ASAQ
In addition, Bayesian paired t-test was utilized to calculate the difference between the English questionnaire and the German questionnaire. The Bayesian estimation of the two groups provided complete distributions of credible values for the effect size, group means and their difference, and the standard deviations plus their difference [4]. 95% CIs, that do not include zero, were regarded as a credible indication of a systematic positive or negative bias and need conversion. As a result, we reported the correlation and difference between the original English questionnaire and translated German questionnaire for the item level, construct level, and scale level.

Cultural Comparison Between Chinese Speakers and German Speakers
Lastly, we looked into the systematic differences between the English data from the German questionnaire scores with earlier obtained Mandarin Chinese data[5]. Following a Bayesian approach, we fitted a multilevel model using Gaussian distribution on each construct and dimension score using a linear model that included culture as a fixed effect, and agent as a varying effect with partial pooling. The analyses used uninformed priors. For the interpretation, we regarded 95% CI of the culture coefficient estimate that excludes zero to indicate a credible indication that there was a difference between the sample groups. Finally, we calculated the posterior probability of either positive or negative bias by taking the posterior distribution area that was either small or greater than zero, whichever will be the largest area.

The data and code can be found here:
https://zenodo.org/record/8079938
Additional data and code can be found at (this includes comparison between German and English Culture):
https://zenodo.org/record/8079245

Classification	ICC Range	90-item set	Construct/ Dimension	24-item set
Excellent	0.75 - 1.00	24	19	6
Good	0.60 - 0.74	36	3	13
Fair	0.40 - 0.59	21	2	3
Poor	0 - 0.39	6	0	2

Table 1: Correlation in terms of ICC value, classified based on Cicchetti [1]

4 Results

Roughly speaking, the German translation is similar to the English survey, just like the Chinese translation [5]

4.1 Correlation between the English and German ASAQ

The average ICC value of the 24 constructs is 0.80, which falls under the classification of excellent [1] (M = 0.80, SD = 0.10, range = [0.51-0.93]). The average ICC value for the 90 questions is 0.65, which falls under the classification of good [1] (M = 0.65, SD = 0.14, range = [0.27, 0.90]). Table 1 shows an overview of the classifications of the ICC values. To see the average ICC value of each construct, see Table 2, and to see the ICC values of the shortened questionnaire, refer to Table 3.

4.2 Variation Between the English and German ASAQ

The average difference in scores between the English and German questionnaires is another estimate of the equivalence of the two. The closer the number is to 0, the more similar they are. A positive difference indicates that the German translations scored higher, and a negative difference indicates that the English version had a higher score. To see these results on a construct basis, refer two Table 2. For the results of the shortened ASAQ, see Table 3. The questions with a significant bias from the entire questionnaire are shown in Table 4.

4.3 Cultural Comparison Between Chinese Speakers and German Speakers

In order to investigate the Cultural Differences and Similarities in Perceptions of Artificial Social Agents Between German and Chinese Speakers, the English data of both the German and Chinese studies have been compared. This way, the translation doesn't add a skew due to the different languages. Most constructs are pretty similar, however, there are 7 that do have a significant difference, where the 95% Credibility Interval does not cover zero, these are: AU (Agent's Usability), PF (Performance), APP (Agent's Personality Presence), UAA (User Acceptance of the Agent), AE (Agent's Enjoyability), AA (Agent's Attentiveness), and SP (Social Presence).

5 Responsible Research

We tried to keep this research responsible. The majority and goal of this research is responsible, but there were some smaller less ethical aspects to it too.

5.1 Ethical Aspects of the Research

The aim of this study was to make the Artificial Social Agent Questionnaire accessible to German speakers. This was to make this form of research more demographically diverse and inclusive. Because of this research, the evaluation of ASAs with German speakers can now be done using the translated ASA Questionnaire. We took great care into conducting this research responsibly. A number of measures were taken to promote study efficiency and warrant group liability. Before we even started the process of this study, an HREC form was filled out by our supervisors and it was approved. This ensures that the research plan is in line with the standards of TU Delft. Regarding the process of our research, an OSF form was filled before conducting any of the surveys. Here, the details of the procedure were documented. The OSF form prevents the research from being conducted in any other way than initially intended. which makes the procedure easily reproducible for future researchers. Additionally, there were more steps to ensure participant privacy. The users were never directly asked for personal information except for their Prolific ID. This ID can be traced back to personal information, so, the supervisor ran the study, replaced the Prolific IDs with made-up identifiers, and sent it to us. This way, the chance of any of their data leaking is incredibly small, especially since there are no other personal questions. Lastly, this research was done in full transparency. All the data and code was published.

5.2 Point of Improvement in the Ethical Aspects of the Research

Even though we made an effort to do this research responsibly, there were some ethical concerns that should be addressed. Firstly, to ensure a gender balance, 15 men and 15 women were asked to participate in the study. To give non-binary people the chance to participate too, they were grouped with women as there were fewer women available on Prolific. This was not ideal as it reinforced a gender binary by implying non-binary people are not a separate group of individuals. In doing so, the number of women in the study was also reduced. The alternative of having a third group for the surveys seemed flawed as that required us to know the ratio of non-binary people to both men and women in the human population. This data is difficult to compile, however. A potential solution for future studies could include grouping non-binary individuals in with both the men and women groups.

5.3 Funding

This work 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. The translation is further funded by the North Rhine-Westphalia state government in Germany.

Consturct/Dimension	ID	Item n	ICC	M De	M En	Δ M	Δ SD	CI 2.5%	CI 97.5%
Agent's Believability									
Human-Like Appearance	HLA	4	0.91	-1.18	-1.16	-0.02	0.10	-0.20	0.17
Human-Like Behavior	HLB	5	0.89	-0.29	-0.34	0.04	0.09	-0.15	0.22
Natural Appearance	NA	5	0.83	-0.38	-0.46	0.16	0.09	0.00	0.35
Natural Behavior	NB	3	0.91	-0.35	-0.47	0.14	0.08	-0.03	0.30
Agent's Appearance suita.	AAS	3	0.76	1.28	1.24	0.10	0.10	-0.10	0.30
Agent's Usability	AU	3	0.77	1.32	1.45	-0.13	0.09	-0.31	0.05
Performance	PF	3	0.73	1.39	1.40	0.00	0.09	-0.18	0.18
Agent's Likeability	AL	5	0.93	0.80	0.8	0.01	0.06	-0.11	0.13
Agent's Sociability	AS	3	0.58	0.87	0.32	0.52	0.14	0.26	0.80
Agent's Personality Presence	APP	3	0.85	-0.56	-0.52	-0.05	0.10	-0.25	0.14
User Acceptance of the A.	UAA	3	0.72	1.30	1.42	-0.12	0.10	-0.31	0.07
Agent's Enjoyability	AE	4	0.82	1.23	1.34	-0.10	0.09	-0.28	0.07
User's Engagement	UE	3	0.51	1.87	1.65	0.19	0.10	0.00	0.40
User's Trust	UT	3	0.76	0.43	0.31	0.07	0.09	-0.11	0.24
User-Agent Alliance	UAL	6	0.81	0.40	0.42	-0.03	0.08	-0.18	0.12
Agent's Attentieness	AA	3	0.68	1.76	1.84	-0.09	0.08	-0.26	0.07
Agent's Coherence	AC	4	0.83	1.69	1.64	0.09	0.07	-0.04	0.22
Agent's Intentionality	AI	4	0.80	0.24	0.43	-0.19	0.09	-0.36	0.00
Attitude	AT	3	0.92	1.41	1.32	0.10	0.07	-0.03	0.23
Social Presence	SP	3	0.81	-0.50	-0.49	0.00	0.10	-0.21	0.20
Interaction Impact on Self.	IIS	4	0.91	0.14	0.21	-0.08	0.06	-0.20	0.04
Emotional Experience									
Agent's Emotional Intellig.	AEI	5	0.86	-0.40	-0.66	0.22	0.10	0.02	0.41
User's Emotion Presen.	UEP	4	0.76	1.04	0.78	0.19	0.09	0.02	0.37
User-Agent Interplay	UAI	4	0.79	1.11	0.86	0.14	0.08	-0.01	0.28
Grand Mean	-	-	0.80	0.61	0.56	0.12	0.09	-	-

Table 2: Correlation between the average score of the 24 constructs, based on Li et al [5]

6 Discussion

6.1 Long Version of Translation

The results suggest that the best way to interpret the researchers who use the German version of the ASAQ should preferably use the long version, and then only measure their ASA on a construct basis. At this level 19 constructs are excellent, 3 are good, and 2 (Agent's Sociability and User's Engagement) are fair. On average, we found an approximate 0.12 difference in scores between the two languages. For questions that don't include 0 in the Credibility interval, we recommend using the mean difference in Table 2 to convert from one language to another.

6.2 Short Version of Translation

The translation of the short version of the ASA is not as similar to the English short version. The short version of the ASAQ consists of a subset of 24 items, each corresponding to a main construct. Two items had a poor correlation, meaning that these translations are not reliable. However, there were 18 items that had a good or even excellent rating. We recommend that if researchers do use the short version, they adjust their scores with the biases that are given in Table 3.

6.3 Cultural Differences

The results corroborate the findings of both intra- and cross-cultural studies on human-ASA interaction. Despite not being able to exclude other reasons for the observed differences (such as collecting the data at different times), there are some questions that have a clear difference in terms of culture. Salem et al [7] suggest that there is a difference between Arabic speakers and English speakers in the way they view ASAs, particularly, the way they see politeness. From our results, it seems that German speakers rate the Agent's Usability, Enjoyability, and Attentiveness higher. This can indicate that on average, Germans view ASAs as a very practical and fun thing than Chinese people. Generally, Chinese speakers rated the agents as better in Performance, Agent Personality Presence, User Acceptance of the Agent, and Social Presence. This could indicate that Chinese people care more about how well the agent performs its tasks, and how much they fit in socially.

6.4 Limitations

This study has some limitations. Firstly, unlike the Chinese study [5] this study only did forward translations, meaning that the German translations were never translated back to English. Doing this could have helped the translation be more accurate. Secondly, unlike the Chinese study [5] we tested

			Mean		Δ		CI	
ID	**Item**	**ICC**	**De**	**En**	**M**	**SD**	**2.50%**	**97.50%**
HLA2	[The agent] has the appearance of a human	0.90	-1.32	-1.31	0.00	0.00	0.00	0.00
HLB5	[The agent] has a human-like manner	0.80	-0.17	-0.17	0.00	0.15	-0.29	0.29
NA4	[The agent] seems natural from the outward appearance	0.62	-0.44	-0.56	0.24	0.13	-0.02	0.50
NB3	[The agent] reacts like a living organism	0.79	0.13	-0.01	0.16	0.15	-0.14	0.45
AAS1	[The agent]'s appearance is appropriate	0.61	1.33	1.18	0.05	0.09	-0.10	0.25
AU1	[The agent] is easy to use	0.61	1.39	1.44	0.00	0.00	0.00	0.00
PF1	[The agent] does its task well	0.71	1.54	1.35	0.19	0.12	-0.04	0.41
AL2	I like [the agent]	0.90	0.85	0.75	0.00	0.00	0.00	0.00
AS1	[The agent] can easily mix socially	0.31	0.71	-0.39	1.13	0.22	0.70	1.56
APP1	[The agent] has a distinctive character	0.56	-0.08	-0.28	-0.01	0.10	-0.21	0.18
UAA1	The user will use [the agent] again in future	0.66	1.49	1.54	0.00	0.00	0.00	0.00
AE1	[R] [The agent] is boring	0.80	1.00	0.81	0.00	0.00	0.00	0.00
UE2	The interaction captured the user's attention	0.49	1.82	1.92	0.00	0.00	0.00	0.00
UT3	The user can rely on [the agent]	0.69	0.54	0.51	0.06	0.13	-0.20	0.31
UAL1	[The agent] and the user have a strategic alliance	0.73	-0.13	-0.22	0.08	0.14	-0.19	0.35
AA2	[The agent] is attentive	0.38	1.69	1.67	-0.12	0.11	-0.33	0.10
AC1	[R] [The agent]'s behavior does not make sense	0.66	1.89	1.86	0.00	0.04	-0.08	0.10
AI3	[R] [The agent] has no clue of what it is doing	0.62	0.88	1.10	-0.18	0.15	-0.46	0.11
AT1	The user sees the interaction with [the agent] as something positive	0.76	1.35	1.18	0.06	0.09	-0.10	0.26
SP2	The agent is a social entity	0.71	-0.53	-0.56	0.05	0.16	-0.26	0.36
IIS2	Others would encourage the user to use [the agent]	0.74	0.13	0.40	0.00	0.00	0.00	0.00
AEI3	[R] [The agent] is emotionless	0.51	0.08	-0.44	0.30	0.22	-0.11	0.74
UEP3	The emotions the user feels during the interaction are caused by [the agent]	0.64	1.15	1.10	0.09	0.13	-0.17	0.34
UAI4	[The agent]'s and the user's emotions change to what they do to each other	0.69	0.63	0.46	0.17	0.15	-0.12	0.47
Grand Mean			0.66	0.56	0.12	0.09	-	-

Table 3: Correlation between the shortend ASA Questionnaire, based on Li et al [5]

	Mean		Δ		CI		Max{P(Δ>0),
Item	**De**	**En**	**M**	**SD**	**2.5%**	**97.5%**	P(Δ <0)}
NA2	-0.29	-0.50	0.30	0.15	0.02	0.59	0.98
AL5	2.13	2.04	0.43	0.12	0.20	0.68	>0.99
AS1	0.71	-0.39	1.32	0.22	0.69	1.56	>0.99
AS2	1.11	0.60	0.49	0.20	0.09	0.89	>0.99
APP2	-0.5	-0.11	-0.28	0.14	-0.56	0.00	0.98
UAA3	0.85	1.29	-0.36	0.17	-0.69	-0.03	0.99
AE4	1.18	1.65	-0.43	0.19	-0.79	-0.05	0.99
UAL3	0.15	-0.43	0.47	0.17	0.14	0.82	>0.99
UAL4	0.86	1.28	-0.41	0.13	-0.67	-0.15	>0.99
UAL5	0.19	0.60	-0.34	0.14	-0.61	-0.06	>0.99
AEI1	0.19	0.47	0.53	0.18	0.18	0.89	>0.99

Table 4: Bias between German translation and English version, based on Li et al [5]

Construct/ Dimenstion	M		Δ		CI		Max{P(Δ>0),
	German	**Chinese**	**M**	**SD**	**2.5%**	**97.5%**	P(Δ <0)}
Agent's Believability							
-*HLA*	-0.70	-1.16	0.28	0.19	-0.09	0.65	0.93
-*HLB*	-0.34	0.01	-0.24	0.21	-0.64	0.17	0.87
-*NA*	-0.22	-0.46	0.12	0.17	-0.22	0.46	0.75
-*NB*	-0.47	-0.19	-0.14	0.18	-0.50	0.22	0.78
-*AAS*	0.98	1.24	-0.29	0.16	-0.61	0.03	0.97
AU	1.45	1.04	0.39	0.16	**0.08**	**0.71**	>0.99
PF	1.07	1.40	-0.32	0.14	**-0.61**	**-0.04**	0.99
AL	0.80	0.61	0.21	0.18	-0.14	0.56	0.88
AS	0.60	0.32	0.24	0.17	-0.10	0.58	0.92
APP	-0.52	0.21	-0.65	0.18	**-0.99**	**-0.31**	>0.99
UAA	1.06	1.42	-0.34	0.15	**-0.64**	**-0.05**	>0.99
AE	1.34	0.95	0.40	0.16	**0.10**	**0.71**	>0.99
UE	1.59	1.65	-0.07	0.15	-0.35	0.22	0.75
UT	0.31	0.35	-0.05	0.15	-0.35	0.24	0.63
UAL	0.64	0.42	0.21	0.14	-0.07	0.49	0.93
AA	1.84	1.51	0.33	0.15	**0.05**	**0.62**	0.99
AC	1.39	1.64	-0.21	0.16	-0.52	0.10	0.91
AI	0.43	0.70	-0.26	0.17	-0.60	0.07	0.94
AT	1.15	1.32	-0.12	0.15	-0.41	0.17	0.79
SP	-0.49	-0.11	-0.36	0.18	**-0.70**	**0.00**	0.98
IIS	0.45	0.21	0.25	0.16	-0.07	0.57	0.94
Emotional Experience							
-*AEI*	-0.66	-0.42	-0.16	0.19	-0.53	0.21	0.80
-*UEP*	0.81	0.78	-0.06	0.17	-0.38	0.27	0.64
UAI	0.96	1.05	-0.05	0.15	-0.35	0.25	0.63

Table 5: Correlation between the English items of the German and Chinese speakers, based on Li et al [5]

more translations for a single question, the last round had 4 translations for each. It is possible that this led to overfitting because the formative rounds were solely based on one ASA. Additionally, this study only tested the questions from a third-person perspective, yet translated it from both a first and third-person perspective. Lastly, the data that was gathered on the Chinese speakers was done through Prolific who did not recruit in China, so the participants of the Chinses survey are likely to have more foreign influence.

7 Conclusions

The German translation of the ASA Questionnaire is pretty similar to the original English one. Especially if the long version is used, and the measurements are done on a construct level. There are some differences in terms of culture between German and Chinese speakers, but only 7 constructs are significantly different.

References

[1] Domenic Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychological Assessment*, 6:284–290, 12 1994.

[2] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire: Establishing the long and short questionnaire versions. *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022.

[3] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Azazi, and Willem-Paul Brinkman. What are we measuring anyway?: - a literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences. pages 159–161, 07 2019.

[4] John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 2013.

[5] 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. 2023.

[6] 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. *Value in Health*, 17(1):70–76, 2014.

[7] Maha Salem, Micheline Ziadee, and Majd Sakr. Marhaba, how may i help you? effects of politeness

and culture on robot acceptance and anthropomorphization. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 74–81, 2014.

[8] Ravi Vythilingam, Deborah Richards, and Paul Formosa. The ethical acceptability of artificial social agents. 2022.

A Overview of Who Did What

A.1 Boleslav Khodakov (German group)

- co-created formative OSF form
- co-created summative OSF form
- created transformation code for first formative round
- created transformation code for second formative round
- created transformation code for third formative round
- created evaluation code for first formative round
- created evaluation code for second formative round
- created evaluation code for third formative round
- created legend files for formative rounds
- created readme files for formative rounds
- created prolific statistics code for formative rounds
- created equalization code for summative round (with German-English data in mind)
- created culture-data creation code for summative round (with German-English data in mind)
- created transformation code for summative round (with German-English data in mind)
- created evaluation code for summative round (with German-English data in mind)
- created legend files for summative round
- created readme files for summative round (German-English version)
- created prolific statistics code for summative round
- Set up the first half of the Prolific study (round 1)
- prepared Qualtrics survey for first half of first round
- tested all Qualtrics surveys for bugs
- Created dummy data for questionnaires
- Co-created Excel documents to send to the translators for formative rounds

A.2 Emma Bokel (German group)

- co-created formative OSF form
- co-created summative OSF form
- Set up the second half of the Prolific study
- Created all Qualtics questionnaires except the first round, first half
- Created dummy data for said questionnaires
- Triple checked the questionnaires to make sure the labels were all correct
- Started a Python script to calculate ICC values in the first round, but Bolek figured out how to run the R code first, so this was never completed or used
- Adjusted Bolek's code to work for the first round, second half
- Helped transform the data in the analysis code
- Created Excel documents to send to the translators
- Created the full document of the translated ASA questionnaire in German

A.3 Kriss Tesink (Dutch group)

- Co-created formative OSF form
- Co-created Qualtrics rounds 1 and 2
- Assisted in creation of dummy data for Qualtrics questionnaires
- Tested Qualtrics rounds 1 and 2 for bugs
- Created prolific codes
- Created Excel documents to be sent to translators
- Created comments for R code for round 1 and 2
- Created evaluation code for round 2
- Wrote R code to find the best alternative translations in round 2
- Assisted in evaluation R code for round 1
- Created the full document of the translated ASA questionnaire in Dutch
- Created code for summative assessment.

A.4 Johan Hensman (Dutch group)

- Co-created formative OSF form
- Co-created Qualtrics rounds 1 and 2
- Created dummy data for the Qualtrics questionnaires
- Created evaluation code for round 1
- Assisted in evaluation code for round 2
- Tested Qualtrics first half of round 1 and round 2 for bugs
- Created legend files for the translation rounds
- Created Readme files for the translation rounds
- Assisted in the Excel files that were sent to the translators
- Created data transformation code for summative assessment (for Dutch-Chinese version)
- Created evaluation code for summative assessment (for Dutch-Chinese version)
- Created legend files for the summative assessment (for Dutch-Chinese version)
- Created Readme files for the summative assessment (for Dutch-Chinese version)
- Provided comments for code for the summative assessment (for Dutch-Chinese version)