Appendix for "German and Dutch Translations of the Artificial-Social-Agent Questionnaire Instrument for Evaluating Human-Agent Interactions"

NELE ALBERS*, Delft University of Technology, Netherlands
ANDREA BÖNSCH* and JONATHAN EHRET, RWTH Aachen University, Germany
BOLESLAV A. KHODAKOV and WILLEM-PAUL BRINKMAN, Delft University of Technology, Netherlands

This document contains the appendix for our paper "German and Dutch Translations of the Artificial-Social-Agent Questionnaire Instrument for Evaluating Human-Agent Interactions." Using the same section names as in the paper, we provide (more) information on:

- The questionnaire translation steps.
- The characteristics of the participants of the Dutch and German summative assessment studies (Table A1).
- The correlation and variation between the original English and the translated questionnaires per construct/dimension (Table A3 and Table A4), representative item from the short version of the ASAQ (Table A5 and Table A6), and item from the full ASAQ with credible bias indication compared to the original English questionnaire (Table A7 and Table A8).
- The mean rating differences for the 24 constructs/dimensions between the three pairs of sample groups on the SLQ (Table A9, Table A10 and Table A11, as well as visualized in Fig. A1).
- Translation-related issues and lessons learned.

QUESTIONNAIRE TRANSLATIONS

Below we provide more information on the translation steps.

Step 1: First forward translation

Two bilingual Dutch-English and four bilingual German-English researchers with expertise in ASAs translated the original English ASAQ independently. One additional researcher per language reconciled the translations, sometimes selecting more than one translation to be assessed.

Step 2: First cycle of formative bilingual assessment

For each language, we recruited 60 bilingual participants with Dutch/German as their first and primary language and English as a fluent language from the crowdsourcing platform Prolific. After watching a 30-second video clip showing an interaction between the Honda robot ASIMO and a human [3][1], each participant rated both the SLQ and the TLQ of either the first 12 or last 12 ASAQ constructs/dimensions together with seven attention check questions per language. Submissions in which a participant failed one or more attention check questions were removed. It is important to note that the ASAQ comprises 19 overarching constructs, which can be further categorized into 24

*Both authors contributed equally to this research.
[1]While participants rated the items, they could rewatch the video clip at any time.

constructs/dimensions, including 5 dimensions under the Agent's Believability construct, 2 dimensions under the Emotional Experience construct, and 17 individual constructs [2]. Each of the 90 items was thus rated by 30 participants. At the end of the questionnaires, participants were asked if they had answered the survey carefully and recommended using their data for scientific purposes.

Based on the entire collected datasets, 34 Dutch and 33 German items displayed ICC values below 0.6, signaling inadequate translation and warranting subsequent refinement. Additionally, considering only the data recommended for use by participants, 3 more Dutch and 2 more German items exhibited ICC values below 0.6. For the Dutch TLQ, the researchers formulated new translations for the 37 items, which another independent researcher again reconciled to select the translations to be assessed. For the German TLQ, five new bilingual translators with expertise in computer science, acoustics, and psychology revised the translations of the 35 inadequate items, which were then synthesized by the same translation coordinator with expertise in ASAs for consistency.

Step 3: Second cycle of formative bilingual assessment

We recruited a new group of 30 bilingual participants per TLQ to evaluate the human-ASIMO interaction using all English items that did not yet have a sufficiently good translation (ICC < 0.6) and their respective new translations. The ICC values of 20 Dutch and 10 German items stayed below 0.6, with one additional German item when using only the data of participants who recommended using their data. The researchers formulated new translations for these 20 Dutch items, while four new bilingual translators with expertise in computer science revised the translations for the 11 German items, as well as for one German item whose ICC value was erroneously considered to be at least 0.6 in the first cycle. Again, the individual translations were synchronized by a translation coordinator.

Step 4: Third cycle of formative bilingual assessment

In this final cycle, we assessed all English items identified as being insufficiently translated in the last assessment and their respective new translations with new groups of 30 bilingual participants per language. For the pool of 9 Dutch and 5 German items that continued to have ICC values below 0.6, we selected the translations with the highest ICC values across all three assessment cycles.

METHODS SUMMATIVE ASSESSMENT

Participants

Table A1 shows the characteristics of study participants for the Dutch and German summative assessments.

Table A1. Characteristics of the participants of the Dutch and German summative assessment studies.

Characteristic	Value	
	Dutch	German
Number		
- n	240	240
Age (in years)		
- Mean (SD)	29 (8)	31 (10)
- Range	18 – 64	19 – 69
Gender		
- Female, n (%)	120 (50%)	113 (47%)
- Male, n (%)	109 (45%)	120 (50%)
- Other, n (%)	11 (5%)	7 (3%)
German as first language		
- No, n (%)		93 (39%)
- Yes, n (%)		147 (61%)
Highest completed education level		
- No formal qualifications, n (%)	/	1 (0%)
- Secondary education (e.g. GED/GCSE), n (%)	4 (2%)	5 (2%)
- High school diploma/A-levels, n (%)	44 (18%)	74 (31%)
- Technical/community college, n (%)	11 (5%)	17 (7%)
- Undergraduate degree (BA/BSc/other), n (%)	109 (45%)	74 (31%)
- Graduate degree (MA/MSc/MPhil/other), n (%)	64 (27%)	63 (26%)
- Doctorate degree (PhD/other), n (%)	7 (3%)	5 (2%)
- Don't know/not applicable, n (%)	1 (0%)	1 (0%)

Abbreviations: SD, Standard deviation; GED, General educational development; GCSE, General certificate of secondary education; BA, Bachelor of Arts; BSc, Bachelor of Science; MA, Master of Arts; MSc, Master of Science; MPhil, Master of Philosophy; PhD, Doctor of Philosophy.

RESULTS

Table A2. ID-to-construct/dimension lookup table.

ID	Construct/Dimension	ID	Construct/Dimension
	ASA's Believability	UE	User's Engagement
HLA	• Humanlike Appearance	UT	User's Trust
HLB	• Humanlike Behavior	UAL	User-ASA Alliance
NA	• Natural Appearance	AA	Attentiveness
NB	• Natural Behavior	AC	Coherence
AAS	• Appearance Suitability	AI	Intentionality
AU	Usability	AT	User's Attitude
PF	Performance	SP	Social Presence
AL	Likeability	IIS	Interaction Impact on Self.
AS	Sociability		Emotional Experience
APP	Personality Presence	AEI	• Emotional Intelligence
UAA	User Acceptance	UEP	• User Emotion Presence
AE	Enjoyability	UAI	User-ASA Interplay

Correlation and Variation between SLQ and TLQs

Table A3 and Table A4 show the correlation and variation between the English and the two translated questionnaires for the 24 constructs/dimensions, Table A5 and Table A6 for the 24 items of the short version of the ASAQ, and Table A7 and Table A8 for items from the full ASAQ with credible bias indication compared to the original English questionnaire. The mean score differences (Δ) thereby are estimates for score equivalence between English and Dutch/German as well as for positive (i.e., the Dutch/German score is higher than the English score) and negative (i.e., the Dutch/German score is lower than the English score) biases. The coloring of the ICC-values is based on the guidelines by Cicchetti [1] as introduced in Table 1 of our paper. The abbreviations of the constructs/dimensions are explained in Table A2.

Table A3. ICC values and mean score differences between the English and Dutch versions of the 24 constructs/dimensions.

		M		Δ		CI	
ID	**ICC**	**Du**	**En**	**M**	**SD**	**2.5%**	**97.5%**
HLA	0.89	-1.34	-1.44	0.08	0.06	-0.03	0.21
HLB	0.89	-0.83	-0.90	0.08	0.07	-0.05	0.21
NA	0.89	-1.05	-1.01	-0.04	0.06	-0.16	0.09
NB	0.83	-0.79	-0.91	0.12	0.08	-0.04	0.28
AAS	0.78	1.20	1.24	-0.06	0.07	-0.20	0.07
AU	0.83	1.31	1.29	0.01	0.06	-0.11	0.13
PF	0.78	1.25	1.22	0.03	0.07	-0.10	0.16
AL	0.58	0.97	0.29	0.68	0.09	**0.49**	**0.85**
AS	0.83	-0.08	-0.04	-0.03	0.08	-0.20	0.13
APP	0.87	-0.41	-0.38	-0.03	0.07	-0.17	0.11
UAA	0.71	1.15	0.95	0.13	0.09	-0.04	0.30
AE	0.86	0.69	0.81	-0.14	0.06	**-0.25**	**-0.02**
UE	0.77	1.98	1.95	0.04	0.05	-0.06	0.13
UT	0.81	0.31	0.26	0.04	0.06	-0.07	0.16
UAL	0.84	0.36	0.29	0.05	0.05	-0.06	0.15
AA	0.79	1.53	1.68	-0.16	0.06	**-0.28**	**-0.03**
AC	0.84	1.62	1.61	0.02	0.06	-0.10	0.13
AI	0.87	0.63	0.65	0.00	0.06	-0.12	0.13
AT	0.85	1.29	1.28	0.00	0.06	-0.11	0.12
SP	0.85	-0.53	-0.47	-0.06	0.07	-0.20	0.08
IIS	0.80	0.16	0.19	-0.03	0.06	-0.15	0.09
AEI	0.93	-0.98	-0.96	0.00	0.06	-0.12	0.12
UEP	0.79	0.94	0.87	0.08	0.07	-0.05	0.22
UAI	0.72	1.21	0.99	0.20	0.07	**0.07**	**0.34**
Mean	0.82	0.44	0.39	0.09	0.07	/	/

Bold CI values exclude 0 and hence provide a credible indication of systematic bias: CI > 0 implies a positive bias (i.e., Dutch score is higher), CI < 0 implies a negative bias (i.e., Dutch score is lower).
Abbreviations: M, Mean; CI, Credible interval; ICC, Intraclass correlation coefficient; SD, Standard deviation.

Table A4. ICC values and mean score differences between English and German versions of the 24 constructs/dimensions.

		M		Δ		CI	
ID	**ICC**	**Ge**	**En**	**M**	**SD**	**2.5%**	**97.5%**
HLA	0.92	-1.18	-1.17	0.00	0.06	-0.11	0.11
HLB	0.88	-0.32	-0.34	0.03	0.07	-0.10	0.16
NA	0.87	-0.51	-0.53	0.07	0.06	-0.06	0.20
NB	0.91	-0.44	-0.53	0.09	0.06	-0.03	0.21
AAS	0.75	1.26	1.24	0.04	0.07	-0.10	0.19
AU	0.77	1.39	1.49	-0.09	0.06	-0.21	0.03
PF	0.74	1.28	1.25	0.02	0.07	-0.11	0.14
AL	0.94	0.65	0.66	0.00	0.04	-0.09	0.08
AS	0.58	0.80	0.32	0.47	0.10	**0.26**	**0.67**
APP	0.87	-0.58	-0.58	0.00	0.07	-0.15	0.13
UAA	0.77	1.26	1.27	-0.01	0.07	-0.15	0.13
AE	0.86	1.10	1.19	-0.08	0.06	-0.20	0.03
UE	0.60	1.83	1.58	0.20	0.08	**0.04**	**0.35**
UT	0.74	0.51	0.40	0.09	0.07	-0.05	0.23
UAL	0.81	0.41	0.43	-0.01	0.06	-0.13	0.10
AA	0.71	1.83	1.84	0.00	0.06	-0.13	0.11
AC	0.83	1.75	1.72	0.03	0.05	-0.08	0.13
AI	0.80	0.44	0.57	-0.08	0.07	-0.22	0.07
AT	0.92	1.38	1.30	0.08	0.05	-0.02	0.18
SP	0.84	-0.62	-0.58	-0.06	0.08	-0.21	0.09
IIS	0.85	0.22	0.21	-0.05	0.05	-0.14	0.06
AEI	0.91	-0.50	-0.72	0.17	0.06	**0.05**	**0.30**
UEP	0.81	0.91	0.73	0.12	0.07	-0.02	0.26
UAI	0.80	0.96	0.92	0.02	0.06	-0.09	0.14
Mean	0.81	0.58	0.53	0.08	0.07	/	/

Bold CI values exclude 0 and hence provide a credible indication of systematic bias: CI > 0 implies a positive bias (i.e., German score is higher), CI < 0 implies a negative bias (i.e., German score is lower).
Abbreviations: M, Mean; CI, Credible interval; ICC, Intraclass correlation coefficient; SD, Standard deviation.

Table A5. ICC values and mean score differences between the English and Dutch versions of the short ASAQ.

		M		Δ		CI	
ID	**ICC**	**Du**	**En**	**M**	**SD**	**2.5%**	**97.5%**
HLA2	0.39	-1.32	-1.44	0.06	0.07	-0.07	0.22
HLB5	0.65	-0.28	-0.69	0.26	0.12	**0.02**	**0.49**
NA4	0.71	-1.32	-1.02	-0.06	0.07	-0.22	0.04
NB3	0.69	-0.37	-0.45	0.09	0.13	-0.17	0.35
AAS1	0.58	1.04	1.36	-0.26	0.12	**-0.49**	**-0.02**
AU1	0.65	1.26	1.18	0.00	0.00	0.00	0.00
PF1	0.71	1.26	1.13	0.10	0.09	-0.08	0.28
AL2	0.41	0.93	0.22	0.69	0.16	**0.38**	**0.99**
AS1	0.77	-0.55	-0.58	0.01	0.04	-0.05	0.09
APP1	0.68	-0.36	-0.28	-0.10	0.11	-0.32	0.13
UAA1	0.70	0.93	1.03	0.00	0.01	-0.02	0.02
AE1	0.74	0.13	0.35	-0.23	0.11	**-0.45**	**-0.03**
UE2	0.51	1.84	1.92	0.00	0.01	-0.02	0.02
UT3	0.74	0.48	0.36	0.09	0.08	-0.07	0.26
UAL1	0.57	0.23	-0.16	0.00	0.01	-0.02	0.03
AA2	0.53	0.94	1.18	-0.22	0.11	**-0.44**	**0.00**
AC1	0.52	1.79	1.61	0.09	0.10	-0.12	0.29
AI3	0.75	0.78	0.88	0.00	0.00	0.00	0.00
AT1	0.74	1.26	1.24	0.00	0.00	0.00	0.00
SP2	0.82	-0.63	-0.51	-0.12	0.09	-0.30	0.05
IIS2	0.73	0.25	0.28	0.00	0.00	0.00	0.00
AEI3	0.82	-0.82	-0.77	0.00	0.00	0.00	0.00
UEP3	0.61	1.02	0.84	0.16	0.13	-0.10	0.40
UAI4	0.64	0.58	0.52	0.01	0.09	-0.18	0.19
Mean	0.65	0.38	0.34	0.11	0.07	/	/

Bold CI values exclude 0 and hence provide a credible indication of systematic bias: CI > 0 implies a positive bias (i.e., Dutch score is higher), CI < 0 implies a negative bias (i.e., Dutch score is lower).
Abbreviations: M, Mean; CI, Credible interval; ICC, Intraclass correlation coefficient; SD, Standard deviation.

Table A6. ICC values and mean score differences between the English and German versions of the short ASAQ.

		M		Δ		CI	
ID	**ICC**	**Ge**	**En**	**M**	**SD**	**2.5%**	**97.5%**
HLA2	0.91	-1.32	-1.25	0.00	0.00	0.00	0.00
HLB5	0.76	-0.26	-0.15	-0.06	0.11	-0.28	0.15
NA4	0.67	-0.50	-0.64	0.18	0.11	-0.02	0.40
NB3	0.76	0.02	-0.16	0.17	0.12	-0.06	0.41
AAS1	0.62	1.32	1.24	-0.01	0.07	-0.16	0.14
AU1	0.61	1.35	1.51	0.00	0.00	0.00	0.00
PF1	0.75	1.30	1.15	0.02	0.05	0.00	0.16
AL2	0.89	0.61	0.62	0.00	0.00	0.00	0.00
AS1	0.30	0.57	-0.31	0.89	0.17	**0.56**	**1.22**
APP1	0.62	-0.22	-0.38	0.06	0.09	-0.13	0.24
UAA1	0.73	1.31	1.37	0.00	0.00	0.00	0.00
AE1	0.78	0.81	0.72	0.00	0.00	0.00	0.00
UE2	0.46	1.76	1.77	0.00	0.02	-0.04	0.05
UT3	0.63	0.71	0.52	0.12	0.10	-0.07	0.32
UAL1	0.76	-0.03	-0.17	0.12	0.10	-0.06	0.32
AA2	0.44	1.60	1.57	-0.02	0.05	-0.14	0.08
AC1	0.65	1.96	1.90	0.00	0.00	0.00	0.00
AI3	0.59	0.98	1.15	-0.07	0.10	-0.28	0.13
AT1	0.78	1.33	1.24	0.00	0.00	0.00	0.00
SP2	0.73	-0.62	-0.59	-0.02	0.08	-0.19	0.14
IIS2	0.76	0.23	0.34	0.00	0.00	0.00	0.00
AEI3	0.63	-0.07	-0.48	0.05	0.08	-0.10	0.22
UEP3	0.63	0.91	0.95	0.06	0.09	-0.11	0.24
UAI4	0.69	0.40	0.37	0.07	0.11	-0.14	0.28
Mean	0.67	0.59	0.51	0.08	0.06	/	/

Bold CI values exclude 0 and hence provide a credible indication of systematic bias: CI > 0 implies a positive bias (i.e., German score is higher), CI < 0 implies a negative bias (i.e., German score is lower).
Abbreviations: M, Mean; CI, Credible interval; ICC, Intraclass correlation coefficient; SD, Standard deviation.

Table A7. Items from the full Dutch ASAQ with credible bias indication compared to the English ASAQ.

		M		Δ		CI	
ID	**ICC**	**Du**	**En**	**M**	**SD**	**2.5%**	**97.5%**
HLB1	0.69	-1.35	-1.16	-0.22	0.10	**-0.42**	**-0.03**
HLB5	0.65	-0.28	-0.69	0.26	0.12	**0.02**	**0.49**
AAS1	0.58	1.04	1.36	-0.26	0.12	**-0.49**	**-0.02**
AL1	0.72	0.52	0.06	0.41	0.12	**0.19**	**0.64**
AL2	0.41	0.93	0.22	0.69	0.16	**0.38**	**0.99**
AL3	0.47	1.43	0.69	0.71	0.15	**0.41**	**1.02**
AL5	0.22	0.58	-1.02	1.59	0.16	**1.28**	**1.92**
UAA3	0.29	1.89	1.06	0.35	0.22	**0.00**	**0.78**
AE1	0.74	0.13	0.35	-0.23	0.11	**-0.45**	**-0.03**
UAL3	0.64	0.15	-0.16	0.27	0.11	**0.05**	**0.49**
AA2	0.53	0.94	1.18	-0.22	0.11	**-0.44**	**0.00**
UEP2	0.56	0.76	0.38	0.36	0.14	**0.09**	**0.64**
UEP4	0.54	0.83	1.14	-0.30	0.11	**-0.51**	**-0.09**
UAI2	0.26	1.87	1.04	0.64	0.12	**0.41**	**0.88**

Bold CI values exclude 0 and hence provide a credible indication of systematic bias: CI > 0 implies a positive bias (i.e., Dutch score is higher), CI < 0 implies a negative bias (i.e., Dutch score is lower).
Abbreviations: M, Mean; CI, Credible interval; ICC, Intraclass correlation coefficient; SD, Standard deviation.

Table A8. Items from the full German ASAQ with credible bias indication compared to the English ASAQ.

		M		Δ		CI	
ID	**ICC**	**Ge**	**En**	**M**	**SD**	**2.5%**	**97.5%**
AS1	0.30	0.57	-0.31	0.89	0.17	**0.56**	**1.22**
AS2	0.33	1.05	0.43	0.58	0.15	**0.28**	**0.88**
APP2	0.75	-0.46	-0.20	-0.21	0.10	**-0.40**	**-0.02**
AE4	0.58	1.18	1.56	-0.31	0.13	**-0.56**	**-0.06**
UAL5	0.66	0.33	0.54	-0.20	0.09	**-0.38**	**-0.02**
AEI1	0.71	-0.32	-0.80	0.47	0.13	**0.22**	**0.72**

Bold CI values exclude 0 and hence provide a credible indication of systematic bias: CI > 0 implies a positive bias (i.e., German score is higher), CI < 0 implies a negative bias (i.e., German score is lower).
Abbreviations: M, Mean; CI, Credible interval; ICC, Intraclass correlation coefficient; SD, Standard deviation.

Cross-language experience comparison

Table A9, Table A10 and Table A11 show the mean rating differences for the 24 constructs/dimensions between the three pairs of sample groups on the SLQ. Fig. A1 further depicts the mean scores for the three sample groups. Most differences are observed for constructs/dimensions related to the enjoyability (e.g., Likeability (AL), Enjoyability (AE)) and believability (e.g., Natural Appearance (NA), Humanlike Behavior (NLB)) of the ASAs.

Table A9. Construct/dimension differences (Δ) between mixed-international English-speaking and bilingual Dutch groups.

	M		**Δ**		**CI**		**Max{P(Δ > 0), P(Δ < 0)}**
ID	**Du**	**En**	**M**	**SD**	**2.5%**	**97.5%**	
HLA	-1.44	-0.75	-0.68	0.13	**-0.93**	**-0.43**	>0.99
HLB	-0.90	0.04	-0.94	0.14	**-1.21**	**-0.66**	>0.99
NA	-1.01	-0.24	-0.74	0.12	**-0.99**	**-0.51**	>0.99
NB	-0.91	-0.29	-0.60	0.13	**-0.85**	**-0.35**	>0.99
AAS	1.24	1.35	-0.10	0.12	-0.33	0.13	0.80
AU	1.29	1.23	0.05	0.11	-0.16	0.27	0.68
PF	1.22	1.31	-0.09	0.11	-0.30	0.13	0.79
AL	0.29	0.77	-0.46	0.12	**-0.69**	**-0.23**	>0.99
AS	-0.04	0.32	-0.33	0.13	**-0.59**	**-0.08**	0.99
APP	-0.38	0.20	-0.56	0.13	**-0.82**	**-0.31**	>0.99
UAA	0.95	1.31	-0.35	0.11	**-0.57**	**-0.13**	>0.99
AE	0.81	1.25	-0.43	0.11	**-0.64**	**-0.21**	>0.99
UE	1.95	1.81	0.13	0.10	-0.05	0.31	0.92
UT	0.26	0.43	-0.18	0.11	-0.39	0.04	0.94
UAL	0.29	0.51	-0.23	0.11	**-0.45**	**-0.02**	0.98
AA	1.68	1.65	0.03	0.11	-0.19	0.25	0.61
AC	1.61	1.55	0.06	0.10	-0.15	0.27	0.73
AI	0.65	0.69	-0.06	0.12	-0.29	0.18	0.69
AT	1.28	1.43	-0.16	0.11	-0.38	0.06	0.92
SP	-0.47	-0.16	-0.32	0.14	**-0.59**	**-0.05**	0.99
IIS	0.19	0.65	-0.47	0.11	**-0.68**	**-0.26**	>0.99
AEI	-0.96	-0.67	-0.31	0.14	**-0.59**	**-0.03**	0.98
UEP	0.87	0.62	0.23	0.11	**0.01**	**0.44**	0.98
UAI	0.99	0.79	0.19	0.11	-0.03	0.41	0.96

Bold CI values exclude 0 and hence provide a credible indication of a difference between the two sample groups.
Abbreviations: M, Mean; CI, Credible interval; SD, Standard deviation.

Table A10. Construct/dimension differences (Δ) between mixed-international English-speaking and bilingual German groups.

	M		Δ		CI		Max{P(Δ >
ID	**Ge**	**En**	**M**	**SD**	**2.5%**	**97.5%**	**0), P(Δ < 0)}**
HLA	-1.17	-0.75	-0.36	0.13	**-0.62**	**-0.11**	>0.99
HLB	-0.34	0.04	-0.36	0.14	**-0.63**	**-0.09**	>0.99
NA	-0.53	-0.24	-0.27	0.12	**-0.51**	**-0.04**	0.99
NB	-0.53	-0.29	-0.21	0.13	-0.46	0.04	0.95
AAS	1.24	1.35	-0.11	0.11	-0.33	0.12	0.83
AU	1.49	1.23	0.26	0.11	**0.05**	**0.46**	0.99
PF	1.25	1.31	-0.05	0.11	-0.26	0.15	0.69
AL	0.66	0.77	-0.11	0.12	-0.34	0.12	0.83
AS	0.32	0.32	0.00	0.13	-0.25	0.25	0.50
APP	-0.58	0.20	-0.76	0.13	**-1.01**	**-0.51**	>0.99
UAA	1.27	1.31	-0.04	0.11	-0.26	0.17	0.66
AE	1.19	1.25	-0.06	0.11	-0.28	0.15	0.72
UE	1.58	1.81	-0.23	0.10	**-0.42**	**-0.04**	0.99
UT	0.40	0.43	-0.04	0.11	-0.26	0.18	0.63
UAL	0.43	0.51	-0.09	0.11	-0.30	0.12	0.79
AA	1.84	1.65	0.18	0.11	-0.03	0.40	0.95
AC	1.72	1.55	0.17	0.10	-0.04	0.36	0.95
AI	0.57	0.69	-0.12	0.12	-0.35	0.12	0.84
AT	1.30	1.43	-0.13	0.11	-0.35	0.09	0.87
SP	-0.58	-0.16	-0.41	0.14	**-0.68**	**-0.13**	>0.99
IIS	0.21	0.65	-0.43	0.11	**-0.65**	**-0.21**	>0.99
AEI	-0.72	-0.67	-0.05	0.14	-0.33	0.22	0.65
UEP	0.73	0.62	0.10	0.11	-0.12	0.32	0.82
UAI	0.92	0.79	0.12	0.11	-0.09	0.34	0.87

Bold CI values exclude 0 and hence provide a credible indication of a difference between the two sample groups.
Abbreviations: M, Mean; CI, Credible interval; SD, Standard deviation.

Table A11. Construct/dimension differences (Δ) between bilingual German and bilingual Dutch groups.

	M		Δ		CI		Max{P(Δ >
ID	**Ge**	**Du**	**M**	**SD**	**2.5%**	**97.5%**	**0), P(Δ < 0)}**
HLA	-1.17	-1.44	0.31	0.16	0.00	0.62	0.97
HLB	-0.34	-0.90	0.58	0.17	**0.25**	**0.91**	>0.99
NA	-0.53	-1.01	0.46	0.14	**0.19**	**0.74**	>0.99
NB	-0.53	-0.91	0.39	0.15	**0.10**	**0.68**	>0.99
AAS	1.24	1.24	0.01	0.14	-0.28	0.29	0.52
AU	1.49	1.29	0.19	0.13	-0.07	0.46	0.92
PF	1.25	1.22	0.04	0.13	-0.23	0.30	0.61
AL	0.66	0.29	0.36	0.14	**0.09**	**0.63**	0.99
AS	0.32	-0.04	0.34	0.15	**0.06**	**0.63**	0.99
APP	-0.58	-0.38	-0.17	0.16	-0.48	0.15	0.85
UAA	1.27	0.95	0.31	0.14	**0.03**	**0.58**	0.98
AE	1.19	0.81	0.37	0.14	**0.10**	**0.63**	>0.99
UE	1.58	1.95	-0.36	0.12	**-0.59**	**-0.13**	>0.99
UT	0.40	0.26	0.14	0.14	-0.14	0.40	0.83
UAL	0.43	0.29	0.14	0.14	-0.13	0.41	0.85
AA	1.84	1.68	0.15	0.13	-0.11	0.41	0.88
AC	1.72	1.61	0.11	0.13	-0.15	0.38	0.79
AI	0.57	0.65	-0.08	0.16	-0.39	0.23	0.70
AT	1.30	1.28	0.03	0.13	-0.22	0.28	0.60
SP	-0.58	-0.47	-0.11	0.17	-0.44	0.21	0.75
IIS	0.21	0.19	0.04	0.14	-0.23	0.32	0.62
AEI	-0.72	-0.96	0.24	0.17	-0.09	0.57	0.93
UEP	0.73	0.87	-0.11	0.14	-0.38	0.17	0.78
UAI	0.92	0.99	-0.06	0.12	-0.30	0.18	0.69

Bold CI values exclude 0 and hence provide a credible indication of a difference between the two sample groups.
Abbreviations: M, Mean; CI, Credible interval; SD, Standard deviation.

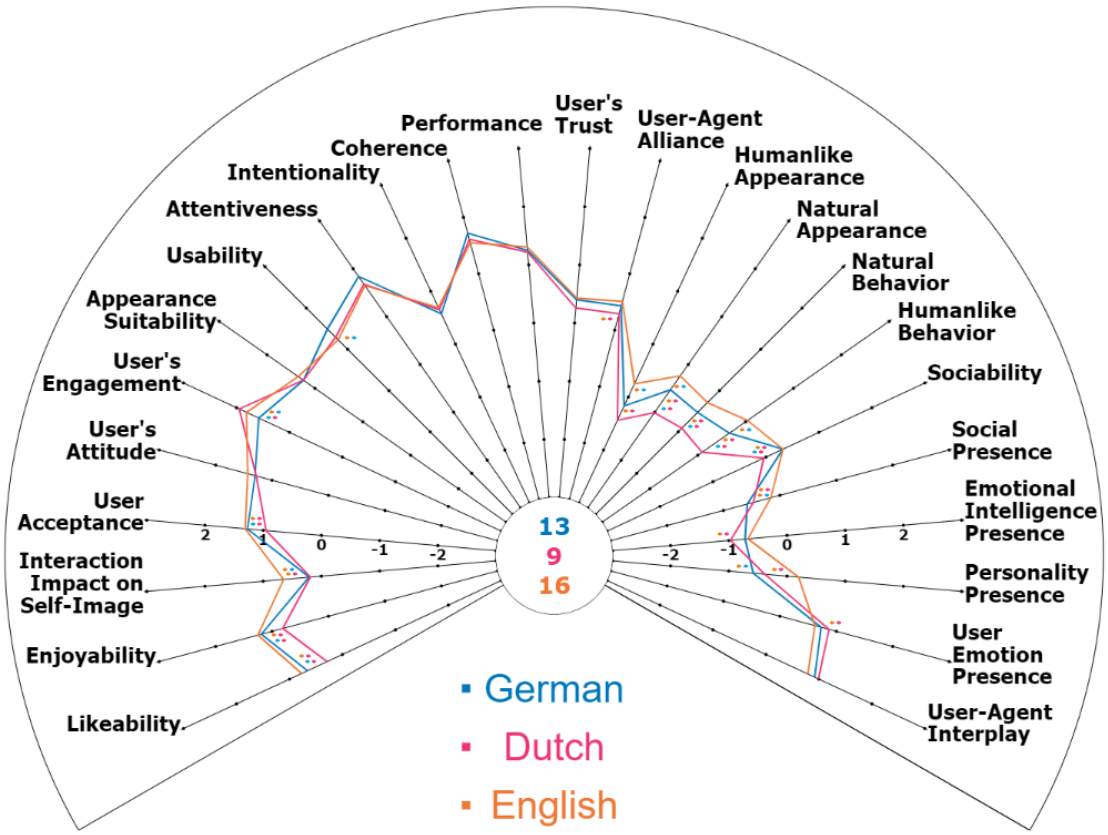

Fig. A1. Mean construct/dimension ratings for the bilingual German, bilingual Dutch, and mixed-international English samples. Two dots highlight differences with credible bias indication. The circle in the center shows the sums across all constructs/dimensions for the three samples.

While considering potential confounding factors such as when the samples were collected (July 2021, second half of 2023), differences in the educational composition of the samples, and variation in language selection criteria (both first and primary language, or only primary language), we found different patterns in the ratings of the ASA's enjoyability (scales Likeability (AL) and Enjoyability (AE)) and believability (Humanlike Appearance (HLA), Natural Appearance (NA), Humanlike Behavior(NLB), Natural Appearance (NA), except Appearance Suitability (AAS)). The original mixed-international English sample group tends to provide the highest ratings, and the bilingual Dutch sample group has the lowest ratings. In contrast, the other constructs were rated rather comparably, with an interesting strong consistency on the performance construct (PF).

SYNTHESIS AND OUTLOOK

Below we provide more in-depth discussions regarding the translation.

Translation-related discussion

Gendered TLQs. Building upon the research by Harkness and Schoua-Glusberg [4], the translation of questionnaires requires meticulous consideration of both linguistic and grammatical elements, specifically for third-person point-of-view formulation in the questionnaire, i.e., referring to observed users who interact with an agent. In the German and Dutch TLQ, we emphasized integrating grammatical gender, which is absent in the English SLQ and is, thus, recognized as a grammatical gain according to Harkness and Schoua-Glusberg [4]. This addition aligns with the cultural commitment to an inclusive language and was guided by two primary categories of concerns. As the discussion about gender-sensitive language is much stronger in Germany, we provide examples for the German TLQ first, followed by Dutch examples.

Firstly, the gender of the ASA is classified as male, female, or neutral. For instance, the SLQ item "[The agent] does its task well" (item PF1), one out of five items requiring an additional pronoun in German and Dutch, is translated to *"[Der Agent/Die Agentin] erledigt seine/ihre Aufgabe gut."* for the German TLQ, with *"Der Agent . . . seine"* for male, *"Die Agentin . . . ihre"* for female, and *"[Der Agent/Die Agentin] . . . seine/ihre"* for the neutral ASA entity, considering the lack of a neutral form for "the agent" as well as the absence of explicit possessive pronouns for things in the German language. While "the agent" can be translated to *"de agent"* for all three genders in Dutch, the possessive pronouns also translate differently for all three cases resulting in *"[De agent] weet wat hij doet"* for a male ASA, *"[De agent] weet wat die doet"* for a neutral ASA, and *"[De agent] weet wat ze doet"* for a female ASA. In total five items had to be adapted based on ASA gender.

Secondly, the gender configuration of the observed human interaction partners is categorized into single male, single female, or ungendered plural. In the German translation, 31 items of the ASAQ are affected by this, for instance, item UT3, which states "[The user] can rely on [the agent]." The translation is tailored based on the gender of the interacting human(s): *"Der Nutzer kann sich auf [den Agent/die Agentin] verlassen"* for a single male, *"Die Nutzerin kann sich auf [den Agent/die Agentin] verlassen"* for a single female, and *"Die Nutzer können sich auf [den Agent/die Agentin] verlassen"* for plural interactants. This design offers flexibility, allowing researchers to align the item's wording to the interactant configuration observed by their participants.

While the three categories employed represent an initial step towards inclusion, it is acknowledged that they do not encompass all possible genders, thereby falling short of fostering a communication environment that fully respects diverse gender identities. However, this decision was made to prioritize simplicity in reading. Exploring further inclusive language alternatives, such as gender-neutral terms (e.g., *"Nutzende"* without a pronoun for plural, not applicable for singular cases) or gender-inclusive pronouns (e.g., *"Der/Die Nutzer*in"* or *"Der/Die NutzerIn"*), was halted to prevent statements from becoming excessively verbose, particularly important for complex questionnaire items (e.g., item IIS4 "People would look favorably at the user because of their interaction with [agent]" translating to *"Andere Menschen würden positiv auf den/die Nutzer*in schauen, aufgrund seiner/ihrer Interaktion mit [dem Agenten/der Agentin]"*. In the Dutch translation, translators opted for the gender non-specific word *"de persoon,"* which can be used for both female and male persons. The direct translation *"de gebruiker/gebruikster"* was not further considered as it would have required more gender-specific adaptations and expressed a specific human-ASA relationship, specifically one in which the human person "uses" the ASA. With *"de persoon,"* the translators preferred a translation that was more neutral with regard to

this relationship. For the Dutch translation, two items had to be adapted for single female users, e.g., *"Vrienden van de persoon zouden haar aanraden om [de agent] te gebruiken"* instead of *"Vrienden van de persoon zouden hem aanraden om [de agent] te gebruiken"*. Additionally, 37 items (e.g., *"Vrienden van de personen zouden hen aanraden om [de agent] te gebruiken"*) had to be adapted for plural users.

The intricate linguistic framework of the two TLQs is designed to accommodate the dual aspects of gender representation, ensuring precision and cultural relevance in communication dynamics. This approach recognizes the grammatical structure of the German and Dutch languages and the significance placed on gender nuances, contributing to a more nuanced and contextually accurate portrayal of social interactions, without causing irritations due to assumed genders.

Maintaining construct consistency amidst linguistic challenges. In addition to achieving grammatical accuracy in the German and the Dutch TLQs, we also encountered the challenge of linguistic loss. The potential loss or distortion of nuanced meanings during the translation of the SLQ to TLQs is a critical concern in cross-cultural research. This challenge is exemplified by the specific case presented, where the English SLQ articulates a positive perception of interactions with an ASA in a nuanced manner. Specifically, the English SLQ item, "The user views the interaction as something favorable"(AT2), subtly conveys that the user regards the interaction with the ASA as something advantageous, beneficial, or even desirable. The German translation *"Die Nutzer sehen die Interaktion als vorteilhaft an"* (for multiple human interactants) aligns with the original nuance but was assessed with a suboptimal ICC of 0.43. Conversely, the German translation *"Die Nutzer bewerten die Interaktion mit [dem Agenten/der Agentin] positiv,"* yielded a higher and closer to optimal ICC of 0.83. Despite affirming positive evaluation, this translation falls short in capturing the broader spectrum of favorable attributes associated with the interaction and thus notably diverges from the nuanced SLQ item. Instead, it is close to another item "The user sees the interaction with [the agent] as something positive"(AT1), translated as *"Die Nutzer sehen die Interaktion mit [dem Agenten/der Agentin] als etwas Positives,"* (ICC = 0.72) which captures the intended meaning more accurately.

REFERENCES

[1] Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6, 4 (1994), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

[2] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. 2022. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *IVA '22: ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, September 6 - 9, 2022*, Carlos Martinho, João Dias, Joana Campos, and Dirk Heylen (Eds.). ACM, 1–8. https://doi.org/10.1145/3514197.3549612

[3] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, and Willem-Paul Brinkman. 2021. Questionnaire Items for Evaluating Artificial Social Agents - Expert Generated, Content Validated and Reliability Analysed. In *IVA '21: ACM International Conference on Intelligent Virtual Agents, Virtual Event, Japan, September 14-17, 2021*. ACM, 84–86. https://doi.org/10.1145/3472306.3478341

[4] Janet Harkness and Alicia Schoua-Glusberg. 1998. Questionnaires in Translation. In *Cross-cultural survey equivalence*, Janet Harkness (Ed.). ZUMA-Nachrichten Spezial, Vol. 3. Zentrum für Umfragen, Methoden und Analysen -ZUMA-, Mannheim, 87–126.