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

Albers, N.; Bönsch, Andrea; Ehret, Jonathan; Khodakov, B.A.; Brinkman, W.P.

DOI
10.1145/3652988.3673928

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents, IVA 2024

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

Nele Albers*
n.albers@tudelft.nl
Delft University of Technology
Delft, Netherlands

Andrea Bönsch*
boensch@vr.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Jonathan Ehret
ehret@vr.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Boleslav A. Khodakov
Delft University of Technology
Delft, Netherlands

Willem-Paul Brinkman
Delft University of Technology
Delft, Netherlands

ABSTRACT

Enabling the widespread utilization of the Artificial-Social-Agent (ASA) Questionnaire, a research instrument to comprehensively assess diverse ASA qualities while ensuring comparability, necessitates translations beyond the original English source language questionnaire. We thus present Dutch and German translations of the long and short versions of the ASA Questionnaire and describe the translation challenges we encountered. Summative assessments with 240 English-Dutch and 240 English-German bilingual participants show, on average, excellent correlations (Dutch ICC M = 0.82, SD = 0.07, range [0.58, 0.93]; German ICC M = 0.81, SD = 0.09, range [0.58, 0.94]) with the original long version on the construct and dimension level. Results for the short version show, on average, good correlations (Dutch ICC M = 0.65, SD = 0.12, range [0.39, 0.82]; German ICC M = 0.67, SD = 0.14, range [0.30, 0.91]). We hope these validated translations allow the Dutch and German-speaking populations to evaluate ASAs in their own language.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → **HCI design and evaluation methods**.

KEYWORDS

Artificial social agent, evaluation instrument, validation, culture

ACM Reference Format:
Nele Albers, Andrea Bönsch, Jonathan Ehret, Boleslav A. Khodakov, and Willem-Paul Brinkman. 2024. German and Dutch Translations of the Artificial-Social-Agent Questionnaire Instrument for Evaluating Human-Agent Interactions. In *ACM International Conference on Intelligent Virtual Agents (IVA '24), September 16–19, 2024, GLASGOW, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3652988.3673928

*Both authors contributed equally to this research.

IVA '24, September 16–19, 2024, GLASGOW, United Kingdom

ACM ISBN 979-8-4007-0625-7/24/09
https://doi.org/10.1145/3652988.3673928

1 INTRODUCTION

To systematically assess human interaction with Artificial Social Agents (ASAs), researchers commonly use post-exposure self-report measures [5] like rating scales and questionnaires [14]. One useful instrument is the ASA Questionnaire (ASAQ), developed collaboratively within the ASA research community, to ensure consistent evaluation across different ASAs. The ASAQ covers 19 constructs assessing the overall ASA quality [15], evaluated through either a concise 24-item (e.g., [1, 8, 10, 29]) or a comprehensive 90-item version. Moreover, researchers can focus on specific constructs relevant to their research (e.g., [11]) while maintaining comparability.

Ensuring widespread usability and accurate data collection in cross-cultural research involves translating research instruments such as the ASAQ, initially developed in English, into multiple languages [16, 27, 28]. This practice enhances accessibility, ensuring accurate comprehension by non-English-speaking participants while reducing errors and misinterpretations associated with ad-hoc translations. Proper translations uphold the instrument's robustness, validity, and reliability [4], enabling aggregated results for generalizable findings and meaningful cross-cultural comparisons.

To enable the ASAQ's broad applicability, the English version serves as the source language questionnaire (SLQ), translated into various target language questionnaires (TLQs). Li et al. pioneered this with a Mandarin Chinese TLQ [22]. Applying their translation approach, the primary contribution of this work is the development and assessment of a Dutch and German TLQ, guided by principles by Harkness and Schoua-Glusberg [16] to address linguistic and cultural nuances. Moreover, we compared samples obtained with the original English SLQ in English-Dutch, English-German, and mixed-international English sample groups, examining cultural nuances and the ASAQ's applicability in cross-cultural research.

2 QUESTIONNAIRE TRANSLATIONS

Figure 1 illustrates our procedures for obtaining the Dutch and German TLQs. Following the approach established by Li et al. [22] for creating a Mandarin Chinese translation of the ASAQ, we performed three rounds of *committee translations* [16] with multiple independent translators per language, each synthesized by a *translation coordinator* [16]. Response data for the SLQ and TLQs were collected from bilingual individuals via Prolific Academic. We assessed their reliability using the Intraclass Correlation Coefficient

(ICC) on item and construct levels to ensure the linguistic and cultural appropriateness of the translations. Thereby, we aimed for ICC values of at least 0.6 based on Cicchetti's guidelines [9]. Items falling below that threshold were reformulated in the next round. We preregistered the procedure [20] and published the underlying data and analysis code [3, 18, 19]. More information on the translation steps can be found in the Appendix [2].

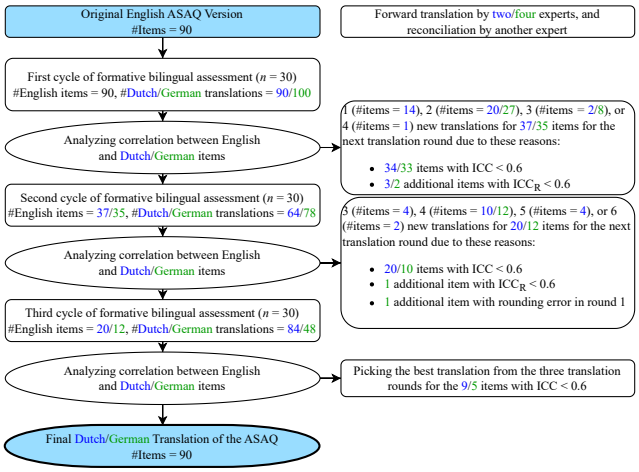

Figure 1: Dutch and German translation procedures. ICC_R denotes an ICC value based on only the data from participants who recommended using their data for scientific purposes.

3 METHODS SUMMATIVE ASSESSMENT

After these two parallel series of formative assessments, we conducted one additional study per finalized TLQ to obtain a summative assessment of the translated ASAQ. Because of fatigue concerns, we split the 24 constructs/dimensions into two parts. The Dutch assessment study was run between 10 October and 14 December 2023; the German assessment study was conducted between 19 June and 2 July 2023. The Human Research Ethics Committee of TU Delft granted ethical approval for the studies (approval number: 3051), which were preregistered [7]. Our finalized Dutch and German ASA questionnaires, data, and analysis code can be found online [2].

3.1 Participants

To be able to detect a small effect size (Cohen's d = 0.2) with a chance of 80% with a Bayesian pairwise t-test and with a small safety margin, we aimed for 120 participants rating each English item and its translation as proposed by Li et al. [22] for 12 constructs/dimensions. So, in total, we aimed for 240 participants per summative assessment. Participants were, again, recruited from Prolific. To be eligible, participants had not to have participated in the earlier formative assessment and to indicate being bilingual and fluent in English. Participants for the Dutch summative assessment further had to report having Dutch as their first and primary language; participants for the German summative assessment had to indicate having German as their primary language[1]. Participants who passed all 14 attention checks were paid based on the minimum payment rules on Prolific (i.e., 6 GBP per hour). Participant characteristics are listed in Table A1 in the Appendix [2].

[1]As we did not get enough participants with German as their first and primary language, we removed the requirement of having German as first language as preregistered [7].

3.2 Procedure

After providing informed consent, participants were first given a control question asking about the languages they were fluent in. If participants confirmed being fluent in English and Dutch/German, they next checked the compatibility of their browsers by watching a test video and answering a control question about the video's content. If participants passed the control question, they saw one of 14 30-second videos, each showing one human-ASA interaction. Afterward, participants rated the human-ASA interaction using the English and the translated items of the first or last 12 constructs/dimensions. Half of the participants first rated the English items; the other half first rated the translated items. All items of the same language were shown in random order together with seven attention check questions. While rating the human-ASA interaction, participants could rewatch the video as often as they wished.

3.3 Materials

We used the 14 human-ASA interaction videos used in the construct validity analysis of the SLQ [13] and the Chinese translation study [22]. For the English items, we referred to all ASAs with gender-neutral pronouns (e.g., "it"). For the Dutch and German items, we used male, female, and gender-neutral pronouns [7].

3.4 Data preparation and analysis

For calculating the ICC and mean score differences between the SLQ and the finalized TLQs, we followed the approach taken by Li et al. [22] when creating the Mandarin Chinese translation. Moreover, we performed Bayesian pairwise comparisons between the SLQ scores of our Dutch and German translation studies and the ones of a mixed-international English sample previously collected in July 2021 based on the same 14 videos [13]. Since participants were solely required to be fluent in English, we regarded this as a mixed international English-speaking sample. As done by Li et al. [22] for the cultural comparison of their bilingual Mandarin Chinese sample and the same mixed-international English sample, we fit a Bayesian multilevel model with uninformed priors, culture as a fixed effect, and agent as a varying effect with partial pooling using the rethinking package [23]. We regarded 95% credible intervals of the culture coefficient estimate that excluded zero as a credible indication of a difference between the two sample groups at hand. We further computed the posterior probability that the culture coefficient is either smaller or larger than zero, and report the largest of these posterior probabilities as the probability of a bias between the two sample groups.

4 RESULTS

4.1 Correlation between SLQ and TLQs

For both the Dutch and the German translations, we obtained on average good ICC levels for the 90 questionnaire items (Dutch: M = 0.65, SD = 0.13, range = [0.22, 0.86], German: M = 0.66, SD = 0.13, range = [0.30, 0.91]). For the 24 constructs and related dimensions,

the correlation is even excellent on average (Dutch: ICC M = 0.82, SD = 0.07, range = [0.58, 0.93], German: ICC M = 0.81, SD = 0.09, range = [0.58, 0.94]). Table 1 shows a good or excellent correlation for 76% of the Dutch and 73% of the German items and for all but 1 Dutch and 2 German constructs and related dimensions. For the 24 representative items of the short version of the ASAQ, we also obtained overall good correlations (Dutch: M = 0.65, SD = 0.12, range = [0.39, 0.82], German: M = 0.67, SD = 0.14, range = [0.30, 0.91]). For 17 (71%) Dutch and 20 (83%) German representative items the correlation is good or excellent and for a further 6 (25%) Dutch and 3 (13%) German items, the correlation can be classified as fair. One representative item per TLQ obtained a poor correlation.

Table 1: Categories of ICC classifications by Cicchetti [9] and number of ICC values per classification category, with the full ASAQ version (90 items) on item-level and construct-level, as well as the short version (24 items) on item-level.

Classification	ICC-Range	90-Item Set	24 Constructs/ Dimensions	24-Item Set
Dutch				
Excellent	0.75 – 1.00	18 (20.0%)	21 (87.5%)	4 (16.7%)
Good	0.60 – 0.74	50 (55.6%)	2 (8.3%)	13 (54.2%)
Fair	0.40 – 0.59	18 (20.0%)	1 (4.2%)	6 (25.0%)
Poor	0 – 0.39	4 (4.4%)	/	1 (4.2%)
German				
Excellent	0.75 – 1.00	25 (27.8%)	19 (79.2%)	9 (37.5%)
Good	0.60 – 0.74	41 (45.6%)	3 (12.5%)	11 (45.8%)
Fair	0.40 – 0.59	20 (22.2%)	2 (8.3%)	3 (12.5%)
Poor	0 – 0.39	4 (4.4%)	/	1 (4.2%)

4.2 Variation between SLQ and TLQs

We analyzed mean score differences between the SLQ and TLQs. These 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. For the 24 constructs/dimensions, we obtained a mean difference of 0.09 in absolute terms (SD = 0.07, range = [-0.16, 0.68]) for Dutch and 0.08 (SD = 0.07, range = [-0.09, 0.47]) for German. There is a credible indication of positive bias for two Dutch and three German constructs/dimensions, and a credible indication of negative bias for two Dutch constructs/dimensions (Enjoyability (AE) and Attentiveness (AA)). Performing a similar analysis for the 24 representative items of the short ASAQ, we observe a credible indication of positive bias for two Dutch and one German item as well as a credible indication of negative bias for three Dutch items. The mean difference for the 24 representative items is 0.11 (SD = 0.07, range = [-0.26, 0.69]) for Dutch and 0.08 (SD = 0.06, range = [-0.07, 0.89] for German. Looking at the 90 items of the full ASAQ, we observe a credible indication of bias for 14 Dutch and 6 German items. The mean difference here in absolute terms is 0.11 (SD = 0.07, range = [-0.30, 1.59]) for Dutch and 0.07 (SD = 0.06, range = [-0.31, 0.89]) for German. Results for all constructs/dimensions and items are given in our Appendix [2].

4.3 Cross-language experience comparison

We found a credible indication of a difference between the mixed-international English and Dutch sample groups for 14 constructs/dimensions, and between the other two pairs of sample groups for eight constructs/dimensions each. Plotting the mean scores for the three sample groups (see Figure A1 [2]) shows that 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. The mixed-international English sample group thereby tends to provide the highest and the bilingual Dutch sample group the lowest ratings. Detailed results can be found in our Appendix [2].

5 SYNTHESIS AND OUTLOOK

The 90-item sets of the Dutch and German ASAQs closely match the initial English version in construct and dimension scores, demonstrating strong equivalence, with only one or two constructs being classified as fair. While the short versions show good average correlations, the Dutch TLQ has a poor correlation for the Human-Like Appearance (HLA) item and, similarly, the German TLQ for the Sociability (AS) item. Consequently, researchers should be cautious when comparing these specific items with results obtained with the original English version. Similar caution applies to items of the long TLQs showing poor correlations with SLQ items. Validating these translations still enables effective study of human-ASA interactions in Dutch and German-speaking populations. Nevertheless, future work should focus on (i) refining Dutch and German TLQ items with poor correlations and (ii) expanding TLQ development to more languages to increase ASAQ accessibility, potentially leveraging large language models for improved translations.

While we delve into linguistic details related to the translation process–— introduction of grammatical genders in both TLQs and challenges related to linguistic loss––in the Appendix [2], our focus here is on two practical usage aspects. (i) With the ASAQ using a 7-point scale from -3 to +3, we found an average absolute score difference of around 0.1 for both languages and versions. Still, our findings indicate credible biases, necessitating conversion corrections when comparing results to those obtained from the SLQ. For example, converting a Dutch AE score of 0.70 to an equivalent English score involves subtracting -0.14 (ΔM), resulting in 0.84, while adding -0.14 (ΔM) to an English AE score yields an equivalent Dutch TLQ score (cp. Table A3 [2]). (ii) Practical usage also requires understanding cultural impacts. Figure A1 [2] illustrates our data comparison of three language groups: mixed-international English, Dutch, and German. We observed varied ratings for ASA attributes like enjoyability and believability, with the English sample generally giving higher ratings than the Dutch sample. Performance ratings, however, were consistent across all groups. These findings highlight the need to consider cultural and linguistic influences in ASA comparisons and emphasize the importance of expanding cross-cultural research to gain deeper insights, supported by existing research (e.g., [12, 24–26]).

ACKNOWLEDGMENTS

This work received funding from 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.

This work is partially based on the Bachelor's theses by Emma Bokel [6], Johan Hensman [17], and Boleslav A. Khodakov [21]. The authors further acknowledge the help they received from Merijn Bruijnes, Esra de Groot, Michaël Grauwde, and Ruben Verhagen in creating the Dutch translations as well as from Jan Delember, Viktor Wolf, Maike Paetzel-Prüsmann, Cosima Ermert, Carolin Breuer, Sara Nefo, Marcel Krüger, Karin Loh, and further dear colleagues in creating the German translation. The authors also appreciate Emma Bokel's assistance in evaluating the German translations.

REFERENCES

[1] Amal Abdulrahman, Katherine Hopman, and Deborah Richards. 2024. Do Not Freak Me Out! The Impact of Lip Movement and Appearance on Knowledge Gain and Confidence. *Multimodal Technologies and Interaction* 8, 3 (2024). https://doi.org/10.3390/mti8030022

[2] Nele Albers, Andrea Bönsch, Jonathan Ehret, Boleslav A Khodakov, and Willem-Paul Brinkman. 2024. German and Dutch Translations of the Artificial-Social-Agent Questionnaire Instrument for Evaluating Human-Agent Interactions: Final Questionnaires, Data, Analysis Code and Appendix. https://doi.org/10.4121/A1457CC7-424A-4BB1-AEAC-1288B5178FBE

[3] Nele Albers and Willem-Paul Brinkman. 2024. Dutch ASA Questionnaire Translation - Translation and Formative Assessment: Round 3. https://doi.org/10.4121/EF5BFD13-3EDD-4D58-972A-AADD667CE3F9

[4] E Berkanovic. 1980. The effect of inadequate language translation on Hispanics' responses to health surveys. *American Journal of Public Health* 70, 12 (Dec. 1980), 1273–1276. https://doi.org/10.2105/ajph.70.12.1273

[5] Timothy W. Bickmore and Daniel Schulman. 2009. A virtual laboratory for studying long-term relationships between humans and virtual agents. In *8th International Joint Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, 297–304.

[6] Emma Bokel. 2023. Cultural Differences and Similarities in Perceptions of Artificial Social Agents Between German and Chinese Speakers. *Bachelor's thesis* (2023). http://resolver.tudelft.nl/uuid:71c06e64-2112-4134-99fb-711361ae4e68

[7] Emma Bokel, Boleslav Khodakov, Nele Albers, Andrea Bönsch, Jonathan Ehret, and Willem-Paul Brinkman. 2023. German and Dutch ASA Questionnaire Translations - Part 2: Summative Assessment. https://doi.org/10.17605/OSF.IO/G3729

[8] Fabien Boucaud, Catherine Pelachaud, and Indira Thouvenin. 2023. "\"It patted my arm\": Investigating Social Touch from a Virtual Agent". In *International Conference on Human-Agent Interaction, HAI 2023, Gothenburg, Sweden, December 4-7, 2023*. ACM, 72–80. https://doi.org/10.1145/3623809.3623853

[9] 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

[10] Martin Dierikx, Nele Albers, Bouke L. Scheltinga, and Willem-Paul Brinkman. 2024. Collaboratively Setting Daily Step Goals with a Virtual Coach: Using Reinforcement Learning to Personalize Initial Proposals. In *Persuasive Technology - 19th International Conference, PERSUASIVE 2024, Wollongong, NSW, Australia, April 10-12, 2024, Proceedings (Lecture Notes in Computer Science, Vol. 14636)*. Springer Nature Switzerland, Cham, 100–115. https://doi.org/10.1007/978-3-031-58226-4_9

[11] Jonathan Ehret, Andrea Bönsch, Patrick Nossol, Cosima A. Ermert, Chinthusa Mohanathasan, Sabine J. Schlittmeier, Janina Fels, and Torsten W. Kuhlen. 2023. Who's next?: Integrating Non-Verbal Turn-Taking Cues for Embodied Conversational Agents. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents, IVA 2023, Würzburg, Germany, September 19-22, 2023*. ACM, 1–8. https://doi.org/10.1145/3570945.3607312

[12] Friederike Eyssel and Dieta Kuchenbrandt. 2011. Social Categorization of Social Robots: Anthropomorphism as a Function of Robot Group Membership. *British Journal of Social Psychology* 51, 4 (Nov. 2011), 724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

[13] 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*. ACM, 1–8. https://doi.org/10.1145/3514197.3549612

[14] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What are We Measuring Anyway?: - A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, IVA 2019, Paris, France, July 2-5, 2019*. ACM, 159–161. https://doi.org/10.1145/3308532.3329421

[15] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 Unifying Questionnaire Constructs of Artificial Social Agents: An IVA Community Analysis. In *IVA '20: ACM International Conference on Intelligent Virtual Agents, Virtual Event, Scotland, UK, October 20-22, 2020*. ACM, 1–8. https://doi.org/10.1145/3383652.3423873

[16] 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.

[17] Johan Hensman. 2023. Perceptions of Artificial Social Agents: The cultural similarities and differences between Dutch and Chinese speakers in their perception of artificial social agents. *Bachelor's thesis* (2023). http://resolver.tudelft.nl/uuid:49a3edc5-1ec1-4518-befc-56ab742625c2

[18] Johan Hensman, Kriss Tesink, Nele Albers, and Willem-Paul Brinkman. 2023. Dutch ASA Questionnaire Translation - Translation and Formative Assessment: Rounds 1 and 2. https://doi.org/10.4121/9DB602BE-F748-46D7-B75D-874A1A0C244F

[19] Boleslav Khodakov, Emma Bokel, Nele Albers, Andrea Bönsch, Jonathan Ehret, and Willem-Paul Brinkman. 2023. German ASA Questionnaire Translation - Part 1: Translation and Formative Assessment. https://doi.org/10.4121/1975AF9A-9B58-4DDE-AE58-58E1001EF553

[20] Boleslav Khodakov, Emma Bokel, Kriss Tesink, Johan Hensman, Nele Albers, and Willem-Paul Brinkman. 2023. German and Dutch ASA Questionnaire Translations - Part 1: Translation and Formative Assessment. https://doi.org/10.17605/OSF.IO/ADKNW

[21] Boleslav A Khodakov. 2023. Differences and similarities in perceptions of interactions with Artificial Social Agents between German and English speakers. *Bachelor's thesis* (2023). http://resolver.tudelft.nl/uuid:065c0c92-37eb-4282-9fbb-073c3b5512bf

[22] Fengxiang Li, Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, Fu Guo, and Willem-Paul Brinkman. 2023. Mandarin Chinese translation of the Artificial-Social-Agent questionnaire instrument for evaluating human-agent interaction. *Frontiers in Computer Science* 5 (2023). https://doi.org/10.3389/FCOMP.2023.1149305

[23] Richard McElreath. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

[24] Mohammad Obaid, Maha Salem, Micheline Ziadee, Halim Boukaram, Elena Moltchanova, and Majd F. Sakr. 2016. Investigating Effects of Professional Status and Ethnicity in Human-Agent Interaction. In *Proceedings of the Fourth International Conference on Human Agent Interaction, HAI 2016, Biopolis, Singapore, October 4-7, 2016*. ACM, 179–186. https://doi.org/10.1145/2974804.2974813

[25] David Obremski, Helena Babette Hering, Paula Friedrich, and Birgit Lugrin. 2022. Exploratory Study on the Perception of Intelligent Virtual Agents With Non-Native Accents Using Synthetic and Natural Speech in German. In *International Conference on Multimodal Interaction, ICMI 2022, Bengaluru, India, November 7-11, 2022*. ACM, 15–24. https://doi.org/10.1145/3536221.3556608

[26] Maha Salem, Micheline Ziadee, and Majd F. Sakr. 2014. Marhaba, how may i help you?: effects of politeness and culture on robot acceptance and anthropomorphization. In *ACM/IEEE International Conference on Human-Robot Interaction, HRI'14, Bielefeld, Germany, March 3-6, 2014*. ACM, 74–81. https://doi.org/10.1145/2559636.2559683

[27] Ligat Shalev, Christian D. Helfrich, Moriah Ellen, Keren Avirame, Renana Eitan, and Adam J. Rose. 2023. Bridging Language Barriers in Developing Valid Health Policy Research Tools: Insights from the Translation and Validation Process of the SHEMESH Questionnaire. *Israel Journal of Health Policy Research* 12, 1 (2023). https://doi.org/10.1186/s13584-023-00583-8

[28] Ami D Sperber. 2004. Translation and Validation of Study Instruments for Cross-Cultural Research. *Gastroenterology* 126 (2004), S124–S128. https://doi.org/10.1053/j.gastro.2003.10.016

[29] Özge Nilay Yalçin. 2023. How (not) to Evaluate Computational Empathy: Testing the Assumptions of the Evaluation Methods in a Use-Case. In *11th International Conference on Affective Computing and Intelligent Interaction, ACII 2023 - Workshops and Demos, Cambridge, MA, USA, September 10-13, 2023*. IEEE, 1–7. https://doi.org/10.1109/ACIIW59127.2023.10388150