Development and Validation of the Conversational Agents Scale (CAS)

Ina Wechsung[1], Benjamin Weiss[1], Christine Kühnel[2], Patrick Ehrenbrink[1], Sebastian Möller[1]

[1] Quality & Usability Lab, TU Berlin, Deutsche Telekom Innovation Laboratories, Berlin, Germany
[2] Potsdam Institute for Climate Impact Research, Potsdam, Germany

{Ina.Wechsung, BWeiss, Patrick.Ehrenbrink, Sebastian.Moeller}@telekom.de,
christine.kuehnel@pik-potsdam.de

Abstract

The gulf between user and system can be minimized by adapting the system to the user‘s natural characteristics. So-called anthropomorphic interfaces represent one strategy of such an adaption as they are assumed to provide a more human-like and therefore more natural interaction. However, regarding the evaluation of anthropomorphic interfaces, the well-known and empirically tested instruments are limited to educational contexts. Hence, this paper describes the first steps towards the development of an evaluation instrument applicable to a wide range of such interfaces.

Index Terms: evaluation, anthropomorphic interfaces, user-centered design.

1. Introduction

Due to the possibilities to use rich natural languages and new flexible ways of interaction, many of today’s interfaces realize the system-as-an-agent metaphor proposed by Jokinen [1]. Jokinen describes such agents as interaction partners mediating between the user and the application, rather than as a tool that is used to perform certain tasks. In the course of this development “anthropomorphic interfaces” became a popular research topic, since such interfaces are assumed to provide a more human-like and therefore more natural interaction. Hence, by adapting the system to users’ natural characteristics, such interfaces represent one strategy to reduce the gulf between the user and the system assumed by Norman [2].

Virtual anthropomorphic interfaces have entered the mass market years ago and have been an active research topic for a long time (for an overview cf. [3]). Currently, the use of social robots, the “next generation” of human-like systems, is increasing. But despite the popularity of such interfaces, a standardized, valid and reliable instrument to measure and evaluate their quality is to our best knowledge not yet available. Consequently, the constructs measured are quite diverse and the results are hardly comparable. Moreover, without a proper validation, it is uncertain if a questionnaire actually measures the constructs, which were intended to measure [4]. Two relatively well-known and psychometrically tested questionnaires are the Agent Persona Instrument (API) [5] and the Attitude Towards Tutoring Agent Scale (ATTAS) [6]. However, both are limited to the evaluation of pedagogical agents. Therefore, we developed a questionnaire applicable to a wide range of anthropomorphic interfaces.

As a theoretical framework we used the dimensions proposed from Ruttkay and colleagues [7]. The original dimensions are: satisfaction, engagement, helpfulness, naturalness and believability, trust, perceived task difficulty, likeability and entertainment. According to Ruttkay et al., satisfaction and engagement are high-level constructs, which comprise several of the other dimensions. For example engagement may include likeability and trust [7]; hence these meta-aspects can be said to derive from the sub-aspects. Consequently, we decided that the questionnaire should comprise the sub-aspects only, which are helpfulness, naturalness and believability, trust, perceived task difficulty, likeability and entertainment.

2. Questionnaire construction

2.1. Item generation

In a first step, 51 items in the form of a semantic differential, i.e. each item consists of a bipolar pair of adjectives, were generated covering all the above dimensions. A seven-point answer format was chosen, because, as indicated by prior research (for an overview see [8]), reliability is highest for seven, and validity highest for five to nine response categories. This initial item set was used in a first study with 50 participants.

2.2. Data collection - Study 1

The study was conducted in a smart home environment. Fifty subjects took part in the study; one participant was excluded as he was a non-native speaker. Each participant rated four different system configurations. Two different virtual agents, Thinking Head [9] and Massy [10], were combined with two different speech synthesis systems, MARY [11] and Mbrola [12]. Participants had to carry out seven different tasks, after task completion the ratings were collected. This was repeated for each system configuration. This way 192 completely filled in questionnaires were gathered. Details on the study are described in Weiss et al. [13].

2.3. Construction of the dimensions

First the item difficulty indices were calculated in accordance with Moosbrugger and Kelava [14]. Item difficulty indices varied between 23 and 76 and were thus all within the recommended range between 20 and 80.

As the questionnaire was based on the theoretical assumptions of Ruttkay et al. [7] a mixed method approach, employing exploratory and confirmatory methods, was chosen to identify items matching the intended dimensions best. Confirmatory approaches are designed to test pre-defined factor structures whereas exploratory analyses do not require any hypothesized factors structures [15]. The development process as described in Homburg and Giering [16] was largely followed. As recommended by Homburg and Giering [16], Cronbach's α, the internal consistency, was calculated for each dimension. Values for Cronbach‘s α were all higher than the .7 (cf. Table 1), which is defined as the minimum [16].

Furthermore, item discrimination indices were computed as the corrected item-scale correlation for each dimension [15]. All items showed sufficient discrimination indices (>.30); hence no item was excluded.

Table 1. *Cronbach's α for all dimensions of initial item set.*

Scale	Initial N of items	Cronbach's α
Likeability	9	.903
Entertainment	8	.827
Helpfulness	12	.947
Naturalness	7	.921
Trust	9	.856
Perceived Task Difficulty	6	.900

As the next step, Homburg and Giering [16] advise calculating an exploratory factor analysis for each dimension, to ensure that the items for each dimension actually form only one factor. Furthermore, each of the single factors is required to explain at least 50% of the variance. Accordingly, exploratory Maximum-Likelihood (ML) factor analyses were calculated for each dimension. The ML methods allows for testing the fit between the hypothesized factor structure (defined number of factors) and the data structure via the χ^2-goodness-of-fit test. If the test shows a non-significant result (p>.05) the null hypothesis, which assumes that the hypothesized factor structure fits the data structure, is kept. Hence, if all items of one dimension are loading on one factor, the χ^2-goodness-of-fit should be non-significant. For all dimensions the χ^2-goodness-of-fit test was significant, indicating a poor fit for the single factor solution. To identify the items not loading on the same factor, a ML factor analyses with two factors was calculated. If also the two-factor solution did not show a good fit, a three-factor solution was investigated. Items loading highest on the second or third factor and/or lowest on the first factor were excluded in order to obtain only one factor for each assumed dimension. After the respective items were removed, the fit of another single-factor solution was investigated for the remaining items to confirm that they compose a single factor structure. Cronbach's α was calculated again for all dimensions. After this step, 32 items remained in the item pool (Table 2).

Table 2. *Cronbach's α and results of the ML factor analyses for all dimensions of the intermediate item set.*

Dimension	N of items	Cronbach's α	Explained Variance	χ^2 (*df*)	*p*
Likeability	5	.838	61.20	4.19 (5)	.523
Entertainment	5	.824	59.71	4.26 (5)	.513
Helpfulness	7	.924	69.95	19.65 (14)	.142
Naturalness	6	.916	71.0	12.41 (9)	.191
Trust	5	.859	64.20	3.20 (5)	.670
Perceived Task Difficulty	4	.830	66.36	.90 (2)	.638

Next, all dimensions were modeled using AMOS, a software package for confirmatory methods like structural equation modeling and confirmatory factor analysis. For confirmatory factor analyses models are specified based on previous theoretical assumptions. Then, it is tested if the hypothesized structure of the model fits the empirical data. For modeling the dimensions (i.e. questionnaire factors), the following criteria, suggested by Homburg and Giering [16], were employed:

Indicator reliability (IR) ≥ 0.4. The indicator reliability is the squared factor loading (regression weight) which equals the squared multiple correlation of the item with the factor. It is the part of the item's variance explained by the factor [15]. Indicator reliability ranges from 0 to 1. Higher values indicate higher reliability.

Composite reliability (CR) ≥ 0.6. The composite reliability (also known as factor reliability) describes how well the factor can be measured via the items and is accordingly a measure of consistency of the factor [18]. Composite reliability ranges from 0 to 1. Higher values indicate higher reliability.

Quotient between χ^2 and the degrees of freedom (*df*) $\left(\frac{\chi^2}{df}\right) \leq 3$. The χ^2-test checks the null hypothesis ("the theoretical model fits the data structure') against the alternative hypothesis ("the theoretical model does not fit the data structure'). Accordingly, a not significant χ^2-test and a high p-value are desired [15]. As p-values are decreasing with increasing sample sizes, the quotient between the χ^2 and the degrees of freedom is used [16].

Average variance extracted (AVE) ≥ 0.5. The average variance extracted indicates the part of the variance explained by the factor compared to the part of the variance resulting from the measurement error [19]. Like the CR, the AVE is a measure of the internal consistency of the factor. The AVE ranges between 0 and 1. Higher values indicate higher consistency.

Additionally the following criteria recommended by Bühner [15] will be used:

Comparative fit index (CFI) ≈ .95. The CFI compares the proposed model with a null model, where all variables are uncorrelated [15]. If the proposed model is better than the null model the CFI increases, i.e. the fit gets better. The CFI ranges between 0 and 1, with 1 indicating a perfect fit.

Root mean square error of approximation (RMSEA) ≤ 0.08. The RMSEA is a so-called badness-of-fit- measure, reflecting the deviation of the observed variance from the hypothesized variance [15].

If several of the above criteria were violated, items showing low indicator reliability were excluded until the criteria were met. For the dimensions Entertainment, Likeability and Trust, one item was excluded each, due to low indicator reliability, leaving 29 items in the item pool. In Table 3 the criteria explained above are presented for each dimension.

2.4. Construction of the whole questionnaire

In the following step the whole questionnaire was investigated using exploratory factor analysis. The aim was to investigate if the proposed factor structure matches the structure of the data. ML factor analyses with Promax rotation were carried out. Promax rotation was chosen over Varimax because Promax rotation allows the factors to correlate [15]. Correlations between the dimensions were expected, for example a higher degree of Entertainment may lead to a higher Likeability.

Table 3. *Fit indices for all dimensions of the intermediate item set.*

Scale	IR Min./ Max.	CR	$\frac{\chi^2}{df}$ (p)	AVE	CFI	RSMEA	N of Items
Likeability	.41/.74	.75	.496 (.609)	.56	1	<.001	4
Entertainment	.42/.65	.72	1.48 (.228)	.54	.97	.050	4
Helpfulness	.54/.71	.89	1.43 (.129)	.64	.99	.048	7
Naturalness	.47./84	.88	1.41 (.179)	.66	1	.046	6
Trust	.49/.75	.82	1.30 (.272)	.64	1	.040	4
Perceived Task Difficulty	.48/.60	.73	.457 (.633)	.55	1	<.001	4

The analysis showed a “clean” pattern matrix. Except for one Naturalness item, all items were loading highest on the assumed factors; cross-loadings were generally very low. The item, which was supposed to measure Naturalness, showed its highest loading on the Entertainment factor. Thus this item was removed. Another factor analysis with the remaining 28 item was carried out. The resulting solution was satisfactory, all items were loading on the expected factors and cross-loadings above 0.3 were not observed (cf. Table 4).

Next the whole questionnaire was modeled in Amos. Again, the criteria explained above were checked. Additionally, the discriminant validity was investigated with the Fornell-Larcker-criterion:

Fornell-Larcker-criterion: AVE(Factor$_x$) > R²(Factor$_x$, Factor$_y$). According to the Fornell-Larcker-criterion the average variance extracted (AVE, see above) of a factors has to be higher than each squared correlation (R²) of this factor with another factor [16]. With the discriminant validity, it is tested if the dimensions (e.g. Likeability and Trust), which are assumed to be different, actually measure different constructs.

The Fornell-Larcker-criterion was met for each dimension. However, the CFI was slightly too low; $\frac{\chi^2}{df}$=1.79, RMSEA=.064, CFI=.92, IR$_{min}$=.41, IR$_{max}$=.83, CR$_{min}$=.72, CR$_{max}$=.89, AVE$_{min}$=.54, AVE$_{max}$=.69. Hence, the modification indices, which are automatically computed by AMOS, were investigated. Modification indices describe which additional parameters (i.e. paths, loadings or regression weights and correlations or covariances) to specify, in order to improve the model fit [15]. High modification indices were observed for one of the Helpfulness items. The item shared a high amount of covariance with the factor Perceived Task Difficulty. Consequently, this item was excluded.

For the next model the CFI increased only marginally. Hence modification indices were inspected another time. This was repeated until the fit was sufficient. In the course of these analyses two further items, one for Helpfulness and one for Naturalness, were removed. The resulting model showed an acceptable fit; $\frac{\chi^2}{df}$=1.68, RMSEA=.060, CFI=.94, IR$_{min}$=.41, IR$_{max}$=.84, CR$_{min}$=.72, CR$_{max}$=.87, AVE$_{min}$=.54, AVE$_{max}$=.70. Also the Fornell-Larcker criterion was met for each factor as the highest observed squared correlation was smaller than the lowest value for the average variance explained; R²(Factor$_x$, Factor$_y$)$_{max}$=.51, AVE$_{min}$=.54. Accordingly, this solution was accepted for the final questionnaire. It consists of 25 items on six scales (cf. Table 4).

3. Validation

Next, the model was validated with the data from another study using anthropomorphic robots. The aim was to ensure that the data structure is not sample dependent and is also valid for other samples and anthropomorphic interfaces, respectively.

3.1. Data collection - Study 2

The study was conducted with different anthropomorphic robotic heads, which served as artificial experimenters. In the baseline condition the robotic head was replaced by a box. For the instructions during the experiment the MARY text-to-speech system was used. The participants had to carry out arithmetic tasks of different complexity levels. The instructions were presented by the robotic heads. Ratings for the robotic heads were collected after all trials. Apart from the CAS, the AttrakDiffMini [20] was employed.

The AttrakDiff comprises four scales; these are Hedonic Qualities-Stimulation (HQS), Hedonic Qualities-Identity (HQI), Pragmatic Qualities (PQ) and Attractiveness (ATT). Hedonic qualities refer to the non-instrumental attributes of a system: The ability to evoke pleasure and emphasize the psychological well-being of the user [20]. The scale HQS measures the extent to which a product can provide stimulation, e.g. by providing novel and exciting features. The scale HQI measures a products ability to express the owner’s self [20]. The scale PQ covers the classical usability attributes: the functionality and the access to the functionality. ATT is the global scale measuring both, hedonic and pragmatic qualities [20]. Twenty-four participants completed both questionnaires. Details on the study are described in Wechsung et al. [21].

3.2. Validation of the theoretical structure and the convergent validation

As the number of cases (N=24) is too small to employ factor analyses, Cronbach’s α was calculated for all dimensions. The results showed satisfactory internal consistency for all dimensions; $\alpha_{Likeability}$=.841, $\alpha_{Entertainment}$=.740, $\alpha_{Helpfulness}$=.944, $\alpha_{Naturalness}$=.816, α_{Trust} =.823, $\alpha_{PerceivedTaskDifficulty}$=.801.

To check the convergent validity (i.e. the correlation between measures of similar constructs) the six CAS scales were correlated with the AttrakDiff sub scales. We expected highest correlations between Likeability and the AttrakDiff scales ATT and HQI as well as between Entertainment and the AttrakDiff scale HQS. Furthermore, Likeability and Entertainment were supposed to be uncorrelated with the AttrakDiff PQ scale. Instead the PQ scale was expected to show highest correlations with Helpfulness and Perceived Task Difficulty. For Trust no correlations were expected. Regarding Naturalness we had no specific hypothesis.

Table 4. *Factor loadings and communalities based on a maximum likelihood analysis with Promax rotation (KMO=.874, Bartlett's test:* χ^2*(378, N=192) =3470.98, p=.000).*

		Helpfulness	Naturalness	Trust	Perceived Task Difficulty	Likeability	Entertainment
The agent is …	diversified vs. monotonic						.698
	lame vs. enthralling						.546
	unconvential vs. convential						.670
	boring vs. entertaining						.785
	pleasant vs. unpleasant					.674	
	friendly vs. unfriendly					.622	
	likable vs. not likable					.684	
	agreeable vs. disagreeable					.922	
	real vs. not real		.670				
	human like vs. artificial		.843				
	unnatural vs. natural		.936				
	unrealistic vs. realistic		.855				
	far from reality vs. close to reality		.754				
The prompts of the agents are …	impractical vs. practical	.719					
	supportive vs. obstructive	.713					
	helpful vs. not helpful	.714					
	destructive vs. constructive	.852					
	useless vs. useful	.820					
	meaningful vs. meaningless	.799					
	worthless vs. valuable	.874					
	sincere vs. insincere			.867			
	honest vs. dishonest			.644			
	non credible vs. credible			.841			
	trustworthy vs. untrustworthy			.734			
The interaction with the agent is …	taxing vs. untaxing				.674		
	undemanding vs. demanding				.783		
	effortful effortless				.652		
	easy vs. difficult				.669		

Note. Factor loadings < .3 are suppressed; items in italic are not part of the final questionnaire. The original items are in German.

The results are largely in line with our assumptions (cf. Table 5). The hedonic scales of the AttrakDiff, which measure non-instrumental aspects, showed high correlations with Likeability and Entertainment, the CAS scales which also refer to emotional, non-task related attributes.

Table 5. *Pearson's correlation between AttrakDiff and CAS scales (N=24).*

	HQS	HQI	ATT	PQ
Likeability	$.800^{**}$	$.860^{**}$	$.863^{**}$	.215
Entertainment	$.842^{**}$	$.728^{**}$	$.712^{**}$	-.095
Helpfulness	.083	.313	.231	$.549^{**}$
Naturalness	$.678^{**}$	$.782^{**}$	$.713^{**}$	.094
Trust	.207	.251	$.460^{*}$	$.450^{*}$
Perceived Task Difficulty	-.050	.152	.199	$.609^{**}$

Note. $^{**}p<.01$, $^{*}p<.05$

The scales Helpfulness and Perceived Task Difficulty, which cover instrumental, goal-oriented aspects, were related to the scale PQ, the AttrakDiff's scale assessing such instrumental characteristics. For Naturalness the highest correlation was observed for HQI. An explanation may be that the more natural the anthropomorphic interface is, the more the user can identify with the agent. Trust showed the lowest correlation with the AttrakDiff. However, the AttrakDiff does not aim to assess such a concept. Hence, the rather low correlations are in line with the theoretical assumptions of the AttrakDiff and the CAS.

4. Discussion and Future Work

The current paper describes the first steps in the development of the CAS (Conversational Agent Scale), which is theoretically grounded in the framework by Ruttkay et al. [7]. The current results indicate that for the evaluation of anthropomorphic virtual agents as well as for anthropomorphic robots the CAS provides reliable assessments of the dimensions proposed by Ruttkay et al. [7]. Thus, the CAS enables researcher to benchmark and compare their interfaces and it can be seen as step towards a unified evaluation framework for anthropomorphic interfaces. However, the CAS still needs to be considered as "work in progress"; further data is needed for its final validation. The data set from Study 2 is rather small and results should be interpreted with caution. Hence, in order to be able to complete the development of the CAS, we hope that many researchers will employ the CAS. Moreover, the CAS is currently available in German only. While we have a translated version, this is not yet validated. Again we need more and larger data sets and hope that the research community is interested in using our current tentative version.

5. References

[1] Jokinen, K.,"Natural Language and Dialogue Interfaces", in C. Stephanidis [Ed], The Universal Access Handbook, 495-506, Taylor & Francis, 2009.

[2] Norman, D.A., "The psychology of everyday things", Basic Books, 1988.

[3] Ruttkay, Z. and Pelachaud, C. [Eds], "From Brows to Trust: Evaluating Embodied Conversational Agents", Kluwer, 2004.

[4] Larsen, L.B., "Assessment of spoken dialogue system usability - what are we really measuring?", in Proccedings of EUROSPEECH 2003, 1945–1948, 2003.

[5] Baylor, A. and Ryu, J., "The API (Agent Persona Instrument) for Assessing Pedagogical Agent Persona", in D. Lassner and C. McNaught [Eds.], Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 448-451, AACE, 2003.

[6] Adcock, A.B. and Van Eck, R., "Reliability and factor structure of the Attitude Toward Tutoring Agent Scale (ATTAS), Journal of Interactive Learning Research, 16(2):195-217, 2005.

[7] Ruttkay, Z., Dormann C. and Noot, H., "Embodied Conversational Agents on a Common Ground. A Framework for Design and Evaluation", in: Z. Ruttkay and C. Pelachaud [Eds], From Brows to Trust: Evaluating Embodied Conversational Agents, 27-66, Kluwer, 2004.

[8] Krosnick, J.A., and Fabrigar, L.R., "Designing rating scales for effective measurement in surveys." in L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D.Trewin [Eds], Survey Measurement and Process Quality, 141-164, Wiley, 1997.

[9] Burnham, D., Abrahamyan, A., Cavedon, L., Davis, C. Hodgins, A., Kim, J., Kroos, C., Kuratate, T., Lewis, T., Luerssen, M., Paine, G., Powers, D. Riley, M., Stelarc, and Stevens, K., "From Talking to Thinking Heads: Report 2008", in R. Göcke, P. Lucey, and S. Lucey [Eds], Proceedings of the International Conference on Auditory-Visual Speech Processing 2008 (AVSP), 127-130, 2008.

[10] Fagel, S., Elisei, F., and Bailly, G., "From 3-d speaker cloning to text-to-audiovisual-speech", in Proceedings of INTERSPEECH 2008, 2325, 2008.

[11] Schroeder, M., and Trouvain; J., "The German text-to-speech synthesis system MARY: A tool for research, development and teaching", International Journal of Speech Technology 6: 365–377, 2003.

[12] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and Van der Vreken, O., "The MBROLA Project: Towards s set of high-quality speech synthesizers free of use for non-commercial purposes", in Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 1996), 1393–1396, 1996.

[13] Weiss, B., Kühnel, C., Wechsung, I., Fagel, S., Möller, S., "Quality of talking heads in different interaction and media contexts", Speech Communication 52(6): 481-492, 2010.

[14] Moosbrugger, H. and Kelava, A., „Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]", Springer, 2008.

[15] Bühner, M., „Einführung in die Test- und Fragebogenkonstruktion [Introduction to test and questionnaire construction]", Pearson, 2011.

[16] Homburg, C., and Giering, A., „Konzeptualisierung und Operationalisierung komplexer Konstrukte - Ein Leitfaden für die Marketingforschung [Conceptualisation and operationalisation of complex constructs - a guideline for market research.]", Marketing – Zeitschrift für Forschung und Praxis 18(1):5-24, 1996.

[17] Amelang, M., Bartussek, D., Stemmler, G. and Hagemann, D., „Differentielle Psychologie und Persönlichkeitsforschung. [Differential psychology and personality research]", Kohlhammer, 2006.

[18] Ruge, M., „Stimmungen und Erwartungen im System der Märkte: eine Analyse mit DPLS-Modellen [Sentiments and Expectations in the systems of the markets: an analysis with DPLS-models]" Universitätsverlag Potsdam, 2011.

[19] Fornell, C., and Larcker, D.F., "Evaluating structural equation models with unobservable variables and measurement error.", Journal of Marketing Research 18(1): 39-50, 1981.

[20] Hassenzahl, M., and Monk, A., "The inference of perceived usability from beauty", Human-Computer Interaction 25(3): 235-260, 2010.

[21] Wechsung, I., Ehrenbrink, P., Schleicher, R. and Möller, S. „Investigating the social facilitation effect in human-robot-interaction.", Proceedings of IWSDS 2012: International Workshop on Spoken Dialog Systems: Towards a Natural Interaction with Robots, Knowbots and Smartphones. Springer, 125-134.