Rethinking Trust Scales for Embodied Conversational Agents: Integrating Explainability, Anthropomorphism and Self-Similarity

Nathalia Sautchuk-Patricio, Matthias Wölfel, Peter A. Henning
Institute for Intelligent Interaction and Immersive Experience
Karlsruhe University of Applied Sciences
Karlsruhe, Germany
0000-0001-9468-3479, 0000-0003-1601-5146, 0009-0003-3150-9950

2026 IEEE Smart World Congress (SWC) | 979-8-3315-5021-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/SWC69865.2026.11764063

***Abstract*—Embodied Conversational Agents (ECAs) offer strong potential to enhance interactive experiences. However, trust remains a critical role in influencing their successful adoption, engagement, and effectiveness. Existing trust measurement scales often fail to capture the multidimensional nature of trust, particularly overlooking ECA-specific factors such as explainability, anthropomorphism and self-similarity. Furthermore, these scales have primarily been validated with adults, leaving a gap in understanding their applicability to school-aged students. To address this gap, the study was conducted with 111 students who interacted with an ECA during a career planning activity. The confirmatory factor analyses showed good model fit with strong factor loadings, reliability, and validity, indicating an overall satisfactory representation of the data. Descriptive results revealed generally positive ECA evaluations with a moderate overall trust level, highlighting the importance of incorporating ECA-specific dimensions into trust models.**

***Index Terms*—education, trust, chatbot, embodied conversational agent, artificial intelligence**

I. Introduction

The integration of technology into educational support systems has become a focus in addressing challenges related to career decision-making and learner engagement. Career counseling, a cornerstone of educational guidance, faces persistent challenges in aligning students' aspirations with viable pathways, while Embodied Conversational Agents (ECAs) offer considerable potential to enrich interactive learning experiences in educational settings by combining verbal and non-verbal communication modalities, such as speech, facial expressions, and gestures. Through these multimodal capabilities, ECAs can create more engaging and socially meaningful interactions. However, their successful implementation depends not only on technological performance but also on learners' trust. Trust plays a critical role in shaping user acceptance, engagement and, ultimately, the effectiveness of ECAs in education [1].

Within human–computer interaction, *trust* is commonly defined as the willingness to accept vulnerability based on the expectation that a system will act in the user's best interest, even in situations where its behavior cannot be fully monitored or controlled [2]. Trust is a multi-faceted concept that cannot be fully captured by a single construct [3]. Existing trust measurement scales often include *ability* (perceived competence and skills), *benevolence* (perceived intention to act in the user's interest), and *integrity* (perceived adherence to acceptable principles), but overlook factors particularly relevant to ECAs, such as *anthropomorphization*, *self-similarity*, and other factors such as *explainability*. These factors may play a decisive role in how learners perceive and evaluate embodied agents. Furthermore, most established trust scales have been developed and validated primarily with adult populations, revealing a lack of empirical testing with teenagers and school-aged students. These limitations likely affect how accurately such instruments capture learners' perceptions and interactions with virtual agents. This highlights the need for more comprehensive trust models tailored to embodied educational systems, as well as for their validation across diverse populations beyond adult samples.

To address these limitations, this study models trust in ECAs by integrating established trust dimensions with ECA-specific factors, including explainability, anthropomorphism and self-similarity. Using data from students who interacted with an ECA in a career planning activity, it examines the multidimensional nature of trust and hypothesizes that benevolence, competence, integrity, explainability, anthropomorphism and self-similarity significantly predict overall trust in ECAs within educational settings.

II. Related Work: Trust in Embodied Conversational Agents

User interfaces are increasingly recognized not merely as functional conduits for information, but as relational spaces where trust is cultivated through embodiment. Building on traditional conversational agents, ECAs transcend text-based interaction by integrating non-verbal communication cues, such as gestures, facial expressions, and vocal prosody, into a cohesive, human-like presence [4]. This embodiment is not decorative; it works as a foundational mechanism for establishing credibility and reliability. By approximating the synchronous, context-sensitive signals of face-to-face commu-

979-8-3315-5021-9/26/$31.00 ©2026 IEEE

Authorized licensed use limited to: University of Applied Sciences Karlsruhe c/o KIT Library. Downloaded on November 11,2026 at 14:56:22 UTC from IEEE Xplore. Restrictions apply.

nication, ECAs generate a sense of authenticity that text-only interfaces cannot replicate [5].

Establishing trust in AI-driven systems remains challenging due to their complexity and opacity, often characterized as "black boxes," which can lead to misunderstandings, negative perceptions, and reduced collaboration if not properly addressed [6], [7]. Enhancing transparency and interpretability is therefore essential, particularly in sensitive domains such as career counseling [8]. Thus, trust is a multidimensional construct that cannot be captured by a single measure [3]. A widely adopted framework by Mayer et. al. [2] conceptualizes trust through three core dimensions: ability (perceived competence), benevolence (intent to act in the user's best interest), and integrity (adherence to acceptable principles). Building on this model, prior research on ECAs has explored additional factors influencing trust, including voice quality, facial similarity, and agent behavior in different interaction contexts [9]–[12]. Several trust models have been developed to capture trust-related constructs, including the Münster Epistemic Trustworthiness Inventory (METI) and the Human-Computer Trust Scale (HCTS) [13]. However, existing models often overlook ECA-specific aspects such as anthropomorphism, self-similarity, and explainability, highlighting the need for more comprehensive frameworks that better capture the unique characteristics of embodied educational agents.

III. Method

This section outlines the study's methodology, including participants, conversational interface, testing procedure, and questionnaire.

A. Participants

To align with the critical timing of career decision-making, we recruited students in their final years of secondary school from partner schools with established prior collaborations. Of the 117 participants who began the study, 111 completed it and were included in the analyses. The mean age of participants was 17.2 years ($SD = 6.5$). Within the sample, 38.1% identified as female, 55.9% as male, 1.8% as non-binary, and 3.6% did not disclose their gender.

B. Conversational User Interface

In order to select a task that is relevant for students and high stake, we decided to focus on career counseling. This domain represents one of the most consequential decision-making processes in young adulthood, with long-term implications. It also demands a rare combination of cognitive clarity and affective support: students require accurate labor market data, skill-mapping, and pathway planning alongside motivation, confidence building, and values clarification. By simulating sustained, context-sensitive dialogue, ECAs can lower psychological barriers to help-seeking while delivering structured guidance. Crucially, the high-stakes nature of career decisions amplifies the consequences of mistrust: students who doubt an agent's reliability, empathy, or competence are likely to disengage, abandon exploration, or make poorly informed choices. Methodologically, career counseling provides a rigorous testbed for evaluating multidimensional trust in ECAs.

The assistant's architecture includes a front-end (see Fig. 1) running on the client side that provides the user interface for interacting with the ECA. Developed in Unity and deployed as a WebGL application, it presents the ECA as a half-body avatar capable of displaying facial expressions and gestures. The interface also integrates speech-to-text and text-to-speech functionalities, along with a panel showing occupation recommendations and suggested educational pathways.

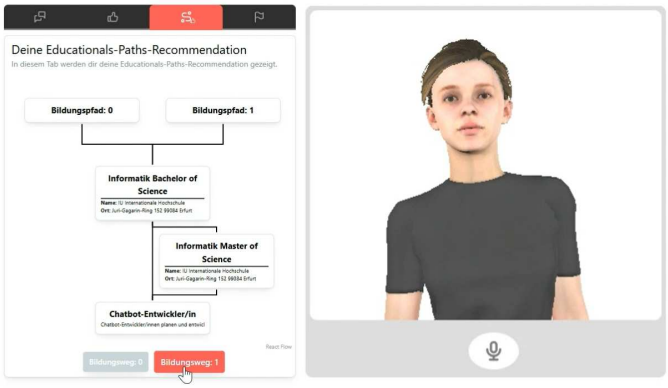


Fig. 1. Assistant frontend

Within the architecture, the back-end manages dialogue, processes user input, and generates context-aware responses using a Large Language Model (LLM) and prediction mechanisms, while leveraging Retrieval-Augmented Generation (RAG) to mitigate hallucinations [14]. LLaMA3.2 was selected as the LLM and was hosted on a GPU-enabled local server to provide enhanced data privacy and reduced latency.

Finally, the data layer handles all aspects of data storage, retrieval, and processing. The dataset powering the assistant was curated, validated, and structured to link occupations with corresponding educational programs in Germany explicitly.

C. Testing Procedure

Testing sessions were conducted in university and school computer labs between February 2025 and February 2026, with up to 20 students participating simultaneously using individual computers and headsets with built-in microphones. The testing session began with an onboarding process, during which researchers explained the testing procedure and obtained participation consent. Following this, the participants continued to interact with the ECA through a web browser. All interactions and instructions during the testing session were conducted in German. Each participant engaged with the ECA at their own pace, exchanging messages through voice and receiving personalized audio responses. The average duration of participant interaction with the ECA was 11 minutes and 57 seconds (SD = 5 minutes and 11 seconds). Each session concluded with participants completing a post-interaction questionnaire.

Authorized licensed use limited to: University of Applied Sciences Karlsruhe c/o KIT Library. Downloaded on November 11,2026 at 14:56:22 UTC from IEEE Xplore. Restrictions apply.

D. Questionnaire

The questionnaire used in this study assessed six key factors of trust in educational ECAs: benevolence, competence, integrity, explainability, anthropomorphism, and self-similarity, as shown in Table I. Benevolence (4 items), competence (6 items), and integrity (4 items) were assessed using the METI [15]. Explainability was measured with an adapted version of the Explanation Satisfaction Scale (ESS), comprising six items [16], while anthropomorphism was assessed using five items from the Godspeed Questionnaire [17]. Self-similarity was measured using items specifically developed for this study.

The questionnaire used a semantic differential scale for five of the six latent constructs, consisting of bipolar adjective pairs representing opposite characteristics. This format was selected for its simplicity and ease of use, as it only requires participants to indicate where their perception lies on a continuum [18], which is particularly suitable for teenage respondents. The Godspeed and METI scales already followed this structure, while items from the ESS were adapted into adjective pairs. Self-similarity, however, was measured using a Likert-type scale assessing perceived similarity to the agent. All measures used a 1–5 scale, with semantic differential responses interpreted as reflecting positive or negative attitudes toward each construct [19].

Moreover, participants provided sociodemographic information, including gender, age, and school year.

TABLE I
QUESTIONNAIRE GROUPED BY LATENT CONSTRUCT
*Note. * indicator removed in the final results*

Latent Construct	Indicator	Observed variable
Benevolence (BE)	BE1	immoral-moral
	BE2	unethical-ethical
	BE3	irresponsible-responsible
	BE4	inconsiderate-considerate
Competence (CO)	CO1	incompetent-competent
	CO2	unintelligent-intelligent
	CO3	poorly educated-well educated
	CO4*	unprofessional-professional
	CO5	inexperienced-experienced
	CO6	unqualified-qualified
Integrity (IN)	IN1	insincere-sincere
	IN2	dishonest-honest
	IN3	unjust-just
	IN4	unfair-fair
Explainability (EX)	EX1	incomprehensible-understandable
	EX2	frustrating-satisfying
	EX3	undetailed-detailed
	EX4	incomplete-complete
	EX5	useless-useful
	EX6	inaccurate-accurate
Anthropomorphism (AN)	AN1	fake–natural
	AN2	machinelike–humanlike
	AN3	unconscious–conscious
	AN4	artificial–lifelike
	AN5	moving rigidly-moving elegantly
Self-Similarity (SS)	SS1	Body (height, weight, fitness)
	SS2	Age
	SS3	Ethnic group
	SS4	Gender

IV. RESULTS

The user study data were analyzed using both descriptive and inferential statistical methods implemented in R with the `lavaan` package.

A. Inferential Statistical Analysis

First, the factorial validity was assessed by conducting a Confirmatory Factor Analysis (CFA) using the Robust Maximum Likelihood method. CFA was chosen because it is a quantitative data analysis method within structural equation modeling that tests how well observed data fit a theoretically specified model of relationships between latent factors and their observed variables [20]. In the first step, a CFA was performed to validate the six-factor structure of benevolence, competence, integrity, explainability, anthropomorphism, and self-similarity (first-order CFA).

The factor loadings of the observed variables on their respective latent constructs are reported in Table II. All loadings were statistically significant at $p < 0.001$ and exceeded 0.50, except for CO4 and SS2, which showed a negative and a low loading, respectively. These results indicate strong associations between most indicators and their corresponding constructs, while suggesting that CO4 and SS2 do not adequately represent their intended dimensions. This is consistent with established CFA guidelines, which recommend removing items with standardized loadings below .40 due to their weak relationship with the latent construct [21], [22].

In addition, Cronbach's alpha was calculated to assess the internal consistency of the scales used in this study. The values of > 0.7 indicate acceptable reliability and > 0.8 indicate very good reliability [23]. In general, the results indicate good to excellent internal consistency for most scales, with Cronbach's alpha values exceeding the commonly accepted threshold of 0.8. The only exception was the self-similarity scale, which demonstrated acceptable but comparatively lower reliability.

Given the low factor loadings of CO4 and SS2, Cronbach's alpha was recalculated after removing each item individually to assess their impact on scale reliability. Excluding CO4 led to a substantial improvement in the internal consistency of the competence scale, increasing Cronbach's alpha from 0.76 to 0.88, thereby supporting its removal. In contrast, removing SS2 reduced the reliability of the self-similarity scale, with Cronbach's alpha decreasing from 0.64 to 0.61, which does not justify its exclusion. Accordingly, CO4 was removed from all subsequent analyses presented in this paper.

Table III summarizes the revised Cronbach's alpha values for each latent construct, obtained after item removal when applicable.

Correlations among the latent constructs were examined using standardized estimates from the CFA model. Values were interpreted as moderate (0.40–0.59), strong (0.60–0.79), and very strong (≥ 0.80). Correlations among the six latent

Authorized licensed use limited to: University of Applied Sciences Karlsruhe c/o KIT Library. Downloaded on November 11,2026 at 14:56:22 UTC from IEEE Xplore. Restrictions apply.

TABLE II
INDICATOR LOADINGS FOR OBSERVED VARIABLES
*Note. * indicator removed in the final results*

Benevolence	Competence	Integrity	Explainability	Anthropomorphism	Self-Similarity
BE1: 0.82	CO1: 0.85	IN1: 0.84	EX1: 0.66	AN1: 0.88	SS1: 0.77
BE2: 0.81	CO2: 0.74	IN2: 0.74	EX2: 0.82	AN2: 0.77	SS2: 0.43
BE3: 0.73	CO3: 0.78	IN3: 0.89	EX3: 0.80	AN3: 0.58	SS3: 0.54
BE4: 0.65	CO4*: -0.23	IN4: 0.84	EX4: 0.81	AN4: 0.80	SS4: 0.50
	CO5: 0.72		EX5: 0.50	AN5: 0.77	
	CO6: 0.79		EX6: 0.80		

TABLE III
CRONBACH'S ALPHA, CONVERGENT VALIDITY AND COMPOSITE RELIABILITY FOR LATENT CONSTRUCTS

Latent construct	Final # Items	Final α	AVE	CR	$\sqrt{\text{AVE}}$
Benevolence	4	0.83	0.57	0.84	0.75
Competence	5	0.88	0.61	0.88	0.78
Integrity	4	0.90	0.69	0.90	0.83
Explainability	6	0.87	0.55	0.88	0.74
Anthropomorphism	5	0.86	0.58	0.87	0.76
Self-Similarity	4	0.64	0.33	0.65	0.58

constructs (see Table IV) revealed positive and statistically significant relationships, with effect sizes ranging from weak ($r = 0.34$) to very strong ($r = 0.85$). The strongest association was observed between competence and explainability ($r = 0.85,\ p < 0.001$). Benevolence and integrity were also strongly correlated ($r = 0.77$, $\mathrm{p} < 0.01$). Conversely, anthropomorphism and self-similarity exhibited moderate correlations with the remaining constructs.

As part of the model evaluation, it is also necessary to assess whether the latent constructs adequately capture the variance of their indicators and demonstrate internal consistency. Convergent validity and composite reliability were therefore assessed using Average Variance Extracted (AVE) and Composite Reliability (CR). Table III presents the results for all latent constructs.

An AVE value below 0.50 indicates that the construct explains less than 50% of the variance in its indicators, suggesting adequate convergent validity [24]. As reported in Table III, this threshold was met for all constructs except self-similarity. This may indicate that the items do not adequately represent the self-similarity construct, possibly due to weaker factor loadings, and suggests that refinement of the measurement scale could be considered.

Regarding composite reliability, values above 0.70 indicate satisfactory internal consistency [24]. In this study, all constructs met this criterion, with the exception of self-similarity, which fell slightly below the recommended threshold. This result points to somewhat limited internal consistency for this construct, consistent with the Cronbach's alpha findings. Nevertheless, since the value was close to the recommended threshold and the construct is theoretically relevant, it was retained for subsequent analyses.

Discriminant validity was assessed using the Fornell–Larcker criterion to ensure that the latent constructs are empirically distinct. This requires that the square root of the Average Variance Extracted ($\sqrt{\text{AVE}}$) for each construct exceeds its correlations with other constructs [21]. As shown in Table III, the $\sqrt{\text{AVE}}$ values were compared with the inter-construct correlations in Table IV. The results indicate discriminant validity issues between competence and explainability, as well as between benevolence and integrity, since in both cases the correlations exceeded the corresponding $\sqrt{\text{AVE}}$ values. This suggests limited empirical distinctiveness between these construct pairs according to the Fornell–Larcker criterion.

Given the substantial correlations among some first-order CFA dimensions, a second-order trust model was specified, integrating all observed indicators across the six latent constructs [25]. Convergent validity and composite reliability were assessed at the higher-order level. The second-order trust construct demonstrated satisfactory reliability ($\mathrm{CR} = 0.88$) and convergent validity ($\mathrm{AVE} = 0.55$), exceeding recommended thresholds ($\mathrm{CR} \geq 0.70$; $\mathrm{AVE} \geq 0.50$). Figure 2 shows the first-order CFA model, while Figure 3 presents the second-order CFA model.

To evaluate the models fit, key statistics commonly reported in CFA were used, including the chi-square statistic (χ^2), degrees of freedom (df), the relative chi-square (χ^2/df), Standardized Root Mean Square Residual (SRMR), Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), p-value, and the Root Mean Square Error of Approximation (RMSEA) with its 90% confidence interval. These indices indicate how well the models represent the observed data according to standard criteria. Table V presents the fit indices for both the first- and second-order CFA.

As shown in Table V, both the first- and second-order CFA demonstrate an overall acceptable to good fit to the data. The χ^2/df ratios are 1.20 for the first-order CFA and 1.27 for the second-order CFA, both well below the recommended threshold of 2, indicating a good model fit [26].

The RMSEA further supports this conclusion, with values of 0.043 (90% CI [0.024, 0.057]) for the first-order CFA and 0.050 (90% CI [0.034, 0.063]) for the second-order CFA. These values indicate a good fit according to commonly accepted criteria, where values below 0.05 suggest a close fit and values below 0.08 indicate an acceptable fit [27].

Similarly, the SRMR values of 0.062 for the first-order CFA and 0.074 for the second-order CFA are below the recom-

Authorized licensed use limited to: University of Applied Sciences Karlsruhe c/o KIT Library. Downloaded on November 11,2026 at 14:56:22 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
CORRELATION BETWEEN THE SIX LATENT CONSTRUCTS FROM FIRST-ORDER CONFIRMATORY FACTOR ANALYSIS.
Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Latent construct	Benevolence	Competence	Integrity	Explainability	Anthropomorphism	Self-Similarity
Benevolence	1	0.61***	0.77***	0.55***	0.35**	0.41**
Competence		1	0.63***	0.85***	0.59***	0.54***
Integrity			1	0.59***	0.34**	0.45**
Explainability				1	0.44***	0.43**
Anthropomorphism					1	0.52***
Self-Similarity						1

TABLE V
SUMMARY OF MODEL FIT INDEXES FROM FIRST- AND SECOND-ORDER CONFIRMATORY FACTOR ANALYSIS

Model	χ^2	df	χ^2/df	SRMR	TLI	CFI	p-value	RMSEA	90% CI
First-order CFA	402.68	335	1.20	0.062	0.95	0.96	0.0065	0.043	0.024 – 0.057
Second-order CFA	438.45	344	1.27	0.074	0.94	0.94	0.0004	0.050	0.034 – 0.063

mended cutoff of 0.10, indicating a good fit [27]. Other fit indices also support model adequacy: the CFI reaches 0.96 for the first-order CFA and 0.94 for the second-order CFA, while the TLI shows values of 0.95 and 0.94, respectively. These values exceed the commonly suggested minimum of 0.90, with the first-order CFA approaching excellent fit according to stricter criteria [28].

Overall, the fit indices indicate that both models provide an adequate representation of the data, with the first-order CFA exhibiting slightly better fit than the second-order CFA.

B. Descriptive Statistical Analysis

Descriptive analyses were conducted to summarize participants' responses and examine the distribution of the latent constructs. Latent factor scores were computed for each participant as weighted averages of their observed indicators, using standardized factor loadings as weights. This approach assigns greater influence to indicators more strongly associated with their underlying constructs, resulting in more representative composite measures. It is consistent with CFA guidelines, which interpret factor loadings as indicators of the strength of the relationship between observed variables and latent constructs, justifying their use as weights in score construction [21], [22], [27].

The resulting latent factor scores were used to compute descriptive statistics for each construct, as shown in Table VI. These results provide an overview of participants' evaluations across the different dimensions, including benevolence, competence, integrity, explainability, anthropomorphism, self-similarity, and overall trust. A repeated-measures ANOVA was also conducted to examine differences across latent constructs, revealing a significant main effect of construct ($F(4.13, 453.96) = 84.44, p < .001, ges = 0.268$), indicating that participants' evaluations varied meaningfully across dimensions.

Finally, an overall trust score was derived by combining the first-order latent constructs into a single composite measure. This aggregation was based on the standardized structural relationships obtained from the second-order CFA, ensuring that constructs contributing more strongly to trust had greater influence on the final score. The resulting trust measure therefore reflects the relative importance of each latent factor in explaining overall trust in the ECA.

TABLE VI
MEAN VALUES AND STANDARD DEVIATIONS OF LATENT CONSTRUCTS AND TRUST CONSTRUCT

Construct	Mean	SD
Benevolence	3.66	0.89
Competence	3.07	0.92
Integrity	3.78	0.90
Explainability	2.99	1.08
Anthropomorphism	2.12	0.83
Self-Similarity	2.75	0.92
Trust	3.10	0.69

V. DISCUSSION

The correlation analysis highlights the interrelated nature of trust dimensions in ECAs. Competence and explainability showed the strongest association, suggesting that perceived ability is closely tied to the clarity and transparency of its explanations. Benevolence and integrity were also strongly correlated, reflecting their conceptual proximity in trust theory. Anthropomorphism and self-similarity were positively associated with most trust dimensions, indicating that human-like characteristics and perceived resemblance may enhance evaluations of competence, benevolence, and explainability. Weaker associations between integrity and self-similarity suggest that shared identity cues do not necessarily imply reliability. Overall, these results support the multidimensional yet interdependent structure of trust, justifying a higher-order trust factor.

The findings of this study provide preliminary support for the scale as a reliable and psychometrically robust instrument for measuring users' trust in ECAs within educational contexts. The application of a second-order CFA proved appropri-

Authorized licensed use limited to: University of Applied Sciences Karlsruhe c/o KIT Library. Downloaded on November 11,2026 at 14:56:22 UTC from IEEE Xplore. Restrictions apply.

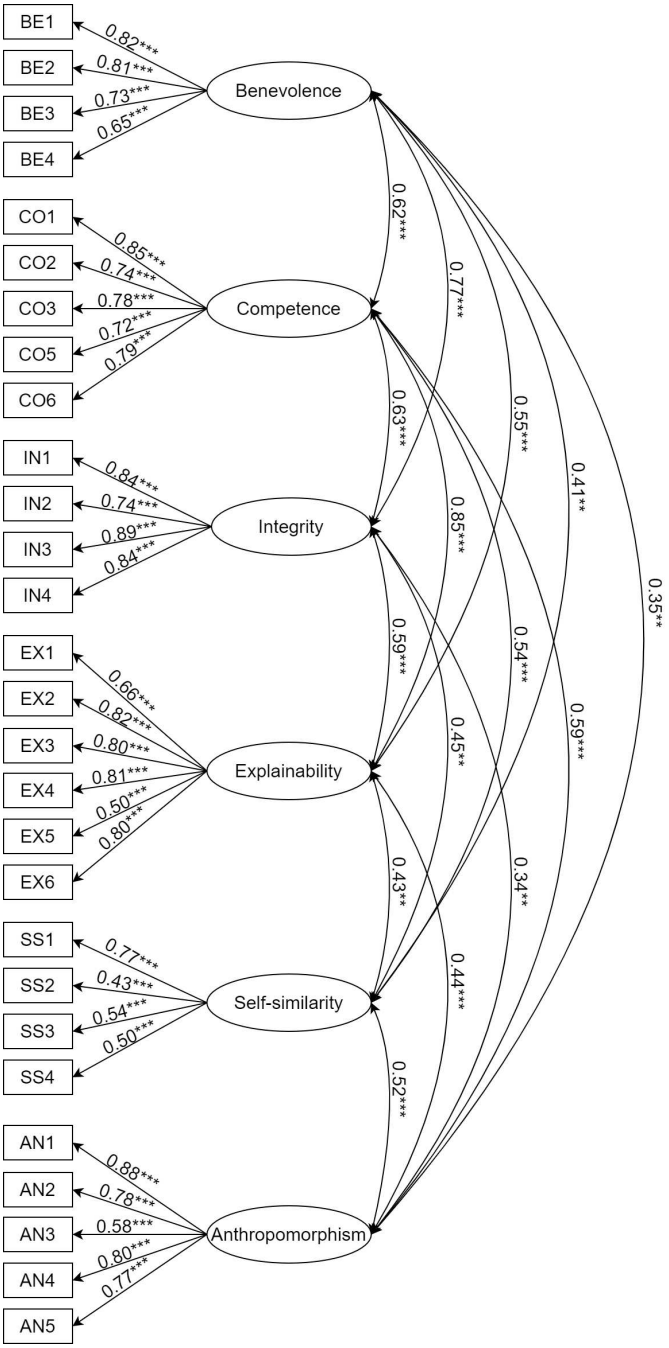


Fig. 2. First-order confirmatory factor analysis

ate for capturing the multidimensional nature of the construct by specifying a higher-order trust factor that accounts for the shared variance among its underlying first-order dimensions. While first-order CFA leaves the meaning of the correlations between the latent factors unexplained, second-order CFA considers theoretically these correlations, providing a better theoretical explanation of the data than first-order CFA [29]. This results in a more concise and conceptually coherent representation of the trust construct, thereby enhancing the theoretical interpretability of the measurement model.

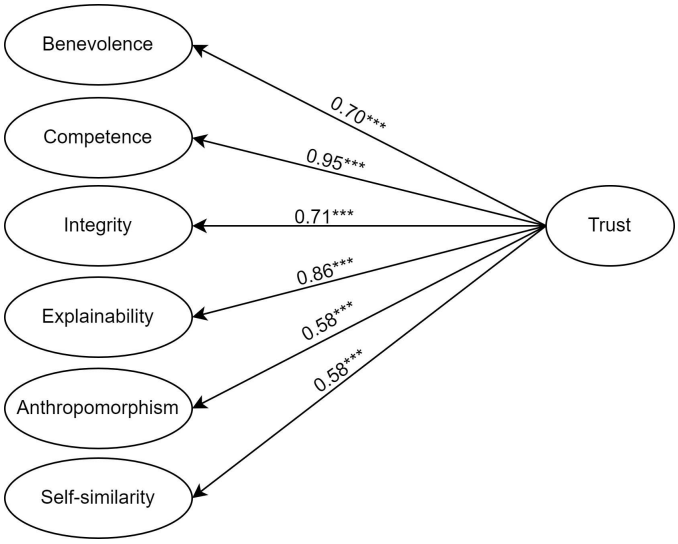


Fig. 3. Second-order confirmatory factor analysis

Analysis of the descriptive statistics revealed that anthropomorphism and self-similarity were the lowest-rated constructs, highlighting that participants generally perceived the ECA as less human-like and less personally similar. Benevolence and integrity were the highest-rated constructs overall. Competence and explainability received moderate ratings, showing that participants recognized the functional and cognitive capabilities of the ECAs, but with less distinction than for benevolence and integrity.

VI. Conclusion

This study examined trust in ECAs in the career counseling context, emphasizing its multidimensional nature and its role in the adoption of ECAs. The results highlight the importance of moving beyond single-dimensional trust measures toward comprehensive models that include constructs such as explainability, anthropomorphism, and self-similarity. These factors were found to predict overall trust significantly, providing initial evidence for the validity and reliability of the proposed trust measurement approach in an educational ECA setting.

Although ECAs elicited generally positive trust evaluations, relational and human-like qualities were comparatively underdeveloped. In particular, lower ratings for anthropomorphism and self-similarity suggest the need for deliberate design strategies to strengthen users' social connection with ECAs. Potential approaches include enhancing anthropomorphic cues through more naturalistic and adaptive behaviors, as well as increasing perceived self-similarity by tailoring appearance and communication to users' characteristics and preferences. Personalization features may further support this alignment, although such enhancements should be implemented carefully to avoid unintended negative effects on trust.

Some limitations of this study should be acknowledged. The relatively homogeneous sample, primarily consisting of a specific student population, limits the generalizability of the findings. In particular, the lack of variation in age, educational

Authorized licensed use limited to: University of Applied Sciences Karlsruhe c/o KIT Library. Downloaded on November 11,2026 at 14:56:22 UTC from IEEE Xplore. Restrictions apply.

background, and cultural diversity constrains the extent to which the results can be transferred to broader populations. Moreover, the use of an immediate post-interaction questionnaire restricts insights into long-term trust formation. No longitudinal or repeated-interaction design was implemented, although such an approach would be valuable for examining how trust evolves over time with continued exposure to the ECA. Additionally, the visual appearance of the ECAs may have influenced participant perceptions. Characteristics such as hairstyle or perceived age may have been interpreted differently, potentially shaping users' impressions of the agents.

Further limitations relate to the technical and experimental setup. Occasional response delays and latency, partly due to network instability, may have affected the quality of interactions and the perceived naturalness of the experience, consequently affecting participants' evaluations. The shared study environment may have introduced background noise and minor interference in speech recognition. Moreover, the use of a smaller, self-hosted LLM may have limited response richness, contextual sensitivity, or linguistic variability.

Future research should address these limitations by including more diverse samples, validating the scale across languages and ECA embodiments, and employing longitudinal and more controlled designs. The integration of more advanced LLMs would further support a deeper understanding of trust formation in ECAs.

REFERENCES

[1] Y. Zhang and W. Pan, "A scoping review of embodied conversational agents in education: trends and innovations from 2014 to 2024," *Interactive Learning Environments*, Feb. 2025.

[2] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," *The Academy of Management Review*, vol. 20, no. 3, Jul. 1995.

[3] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective," *International Journal of Human–Computer Interaction*, vol. 40, no. 5, Mar. 2024.

[4] H. M. Aljaroodi, M. T. P. Adam, R. Chiong, and T. Teubner, "Avatars and Embodied Agents in Experimental Information Systems Research: A Systematic Review and Conceptual Framework," *Australasian Journal of Information Systems*, vol. 23, Oct. 2019.

[5] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani, "Explainable Embodied Agents Through Social Cues: A Review," *ACM Transactions on Human-Robot Interaction*, vol. 10, no. 3, Jul. 2021.

[6] S. Gulati, S. Sousa, and D. Lamas, "Design, development and evaluation of a human-computer trust scale," *Behaviour & Information Technology*, vol. 38, no. 10, Oct. 2019.

[7] P. Guleria and M. Sood, "Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling," *Education and Information Technologies*, vol. 28, no. 1, Jan. 2023.

[8] F. Trujillo, M. Pozo, and G. Suntaxi, "Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction," *Journal of Technology and Science Education*, vol. 15, no. 1, Mar. 2025.

[9] E. K. Chiou, N. L. Schroeder, and S. D. Craig, "How we trust, perceive, and learn from virtual humans: The influence of voice quality," *Computers & Education*, vol. 146, Mar. 2020.

[10] A. C. Elkins, D. C. Derrick, J. K. Burgoon, and J. F. Nunamaker Jr., "Predicting Users' Perceived Trust in Embodied Conversational Agents Using Vocal Dynamics," in *2012 45th Hawaii International Conference on System Sciences*, Jan. 2012, iSSN: 1530-1605.

[11] L. Tang and M. Bashir, "Do We Trust Embodied Agents who Look Like us?" in *2023 International Conference on Intelligent Metaverse Technologies & Applications (iMETA)*. Tartu, Estonia: IEEE, Sep. 2023.

[12] R. Moradinezhad and E. T. Solovey, "Investigating Trust in Interaction with Inconsistent Embodied Virtual Agents," *International Journal of Social Robotics*, vol. 13, no. 8, Dec. 2021.

[13] N. Sautchuk-Patricio and M. Wölfel, "Trust in embodied conversational agents: A cross-scale evaluation of hcts and meti," in *2026 2nd International Conference on Human-AI Interaction and Experience Design (HAXD)*, 2026, pp. 206–211.

[14] O. Ayala and P. Bechard, "Reducing hallucination in structured outputs via retrieval-augmented generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, Y. Yang, A. Davani, A. Sil, and A. Kumar, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024.

[15] F. Hendriks, D. Kienhues, and R. Bromme, "Measuring Laypeople's Trust in Experts in a Digital Age: The Muenster Epistemic Trustworthiness Inventory (METI)," *PLOS ONE*, vol. 10, no. 10, Oct. 2015.

[16] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance," *Frontiers in Computer Science*, vol. 5, Feb. 2023.

[17] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics*, vol. 1, no. 1, Jan. 2009.

[18] J. Al-Hindawe, "Considerations when constructing a semantic differential scale," 7 1996.

[19] J. R. Lindner and N. Lindner, "Interpreting Likert type, summated, unidimensional, and attitudinal scales: I neither agree nor disagree, Likert or not," *Advancements in Agricultural Development*, vol. 5, no. 2, Jan. 2024.

[20] R. O. Mueller and G. R. Hancock, "Factor analysis and latent structure analysis: Confirmatory factor analysis," in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, second edition ed., J. D. Wright, Ed. Oxford: Elsevier, 2015.

[21] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 8th ed. Cengage, 2019.

[22] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, 2nd ed. Guilford Press, 2015.

[23] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Research in Science Education*, vol. 48, no. 6, Dec. 2018.

[24] G. W. Cheung, H. D. Cooper-Thomas, R. S. Lau, and L. C. Wang, "Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations," *Asia Pacific Journal of Management*, vol. 41, no. 2, 2024.

[25] M. Tavakol and A. Wetzel, "Factor analysis: A means for theory and instrument development in support of construct validity," *International Journal of Medical Education*, vol. 11, 2020.

[26] K. A. Bollen, *Structural Equations with Latent Variables*, 1st ed. Wiley, Apr. 1989.

[27] R. B. Kline, *Principles and practice of structural equation modeling*, 3rd ed., ser. Methodology in the social sciences. New York: Guilford Press, 2011, oCLC: ocn636917444.

[28] L. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 6, no. 1, Jan. 1999.

[29] B. M. Byrne, *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming*, 2nd ed. Routledge, 2010.

Authorized licensed use limited to: University of Applied Sciences Karlsruhe c/o KIT Library. Downloaded on November 11,2026 at 14:56:22 UTC from IEEE Xplore. Restrictions apply.