A Systematic and Validated Translation of the Perceived Empathy of Technology Scale from English to German

Matthias Schmidmaier
LMU Munich
Munich, Bavaria, Germany
matt@schmidmaier.org

Lukas Schöberl
LMU Munich
Munich, Bavaria, Germany
lukas.schbrl@gmail.com

Jonathan Rupp
University of Innsbruck
Innsbruck, Austria
jonathan.rupp@uibk.ac.at

Sven Mayer
LMU Munich
Munich, Germany
TU Dortmund University
Dortmund, Germany
info@sven-mayer.com

Abstract

Empathic interaction is becoming increasingly important in human-AI interaction, particularly for applications in emotional and mental health support. As these technologies expand globally, culturally and linguistically adapted evaluation tools become essential, as research shows that emotional processing and empathic responses are stronger in one's native language. We present a systematic translation and validation of the Perceived Empathy of Technology Scale (PETS) from English to German, following a comprehensive back-translation methodology. Our process included multiple independent translations, expert group discussions, and validation with $N = 400$ participants across both languages. Through confirmatory factor analysis and measurement invariance testing, we demonstrate that the German PETS maintains the two-factor structure of the original scale with excellent reliability and achieves configural, metric, and scalar invariance across languages. This validated German PETS enables researchers and developers to accurately assess how German-speaking users perceive the empathic behavior of technological systems, supporting the development of culturally appropriate empathic technologies while further establishing a methodological foundation for future scale translations in HCI.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **HCI theory, concepts and models**.

Keywords

back-translation, empathy, cross-cultural research, scale translation

ACM Reference Format:
Matthias Schmidmaier, Lukas Schöberl, Jonathan Rupp, and Sven Mayer. 2025. A Systematic and Validated Translation of the Perceived Empathy of Technology Scale from English to German. In *Mensch und Computer 2025 (MuC '25), August 31–September 03, 2025, Chemnitz, Germany*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3743049.3743082

MuC '25, Chemnitz, Germany

ACM ISBN 979-8-4007-1582-2/25/08
https://doi.org/10.1145/3743049.3743082

Table 1: The translated and validated German PETS. To be used with randomized 101-point sliders ranging from *"stimme überhaupt nicht zu"* to *"stimme voll und ganz zu"*. For the original English PETS version see Table 2.

PETS-ER	Emotionale Reaktionsfähigkeit
E1	Das System berücksichtigte meine mentale Verfassung.
E2	Das System wirkte emotional intelligent.
E3	Das System drückte Emotionen aus.
E4	Das System zeigte Sympathie mir gegenüber.
E5	Das System zeigte Interesse an mir.
E6	Das System unterstützte mich dabei, mit einer emotionalen Situation umzugehen.
PETS-UT	**Verständnis und Vertrauen**
U1	Das System verstand meine Ziele.
U2	Das System verstand meine Bedürfnisse.
U3	Ich vertraute dem System.
U4	Das System verstand meine Absichten.

1 Introduction

Empathic interaction has emerged as a foundational concept in human-AI interaction, potentially transforming how users engage with current and future digital systems such as chatbots or social robots. Recent technological advances have significantly expanded the capabilities of such systems, particularly in regard to the processing, understanding, and expression of emotional and empathic context [36, 37]. This capability is especially valuable in emotional support scenarios or mental health applications, where empathic behavior has been shown to increase engagement, trust, and help users cope with emotional challenges [2, 5, 19]. The growing availability and use of such digital mental health applications reflect this trend, especially in Germany, where the digital healthcare act of 2019 has created a regulatory framework supporting their integration into standard care [24]. However, research indicates that emotional perception and empathic reactions are subject to significant cross-cultural effects and native-language influences [31, 38, 46], with studies showing, for example, that emotional responses are typically stronger in one's native language [20, 44]. This underscores the need for validated, localized evaluation tools for assessing empathic systems in that context. With our translation of the Perceived Empathy of Technology Scale (PETS) [43], we provide researchers and developers with a validated instrument to evaluate the perceived empathic behavior of systems among German-speaking users (see Table 1). Our overall

objective is to foster future language-adapted human-AI interaction with empathic systems such as context-specific chatbots, personal assistants, and social or care robots. Building on the work of Brislin [7], Klotz et al. [32] and Jones et al. [29], our study demonstrates the successful application of a validated, systematic back-translation process for HCI scales. This provides a methodological foundation for future translations of similar measurement instruments across different cultural and linguistic contexts. Additional information on how to apply the PETS can be found at https://perceived-empathy-of-technology-scale.com.

2 Related Work

Next, we introduce the overall context of empathic systems, the PETS, cross-cultural aspects of empathy in HCI, and the back-translation process as the foundation of our approach.

2.1 Empathic Systems

Empathy is a core construct in human interaction and in modern HCI, especially with the increasing conversational capabilities of artificial systems such as chatbots, voice agents, or social robots [36, 37]. Implementing empathic behavior in such systems was found to increase engagement and trust [2, 5] and could help users cope with emotional challenges such as anxiety [2] or social exclusion [19]. These benefits are especially pronounced in the context of medical or mental health support. For example, in medical contexts, research found that empathic chatbots are preferred over non-empathic ones [16, 33]. A famous early example is "Woebot", a rule-based chatbot that generates empathic responses designed to reduce symptoms of depression [22]. More recent research particularly focuses on LLM-based applications, showing that systems such as *ChatGPT* were able to outperform human empathic response generation, for example, in conversations between patients and healthcare providers [1, 21, 34, 47].

In the context of mental health, Yonatan-Leus and Brukner [49] explored AI-based chatbots, finding that artificial responses scored higher in perspective-taking, empathic concern, and supportive interventions compared to human psychotherapists. Similarly, Seitz [45] explored empathic communication in healthcare chatbots in more detail and found that empathy enhances perceived warmth, trust, and intention to use. However, they also highlight that empathic and sympathetic responses can decrease the perceived authenticity of a system. Although for the scope of this work we will not go deeper into empathic system design, we do highlight that applications in this context involve not only benefits, but also risks and ethical concerns, including issues related to privacy, liability, and social implications such as emotional attachment or technological dependency [8, 14, 40]. Despite such risks, the increasing number of available applications shows the potential of employing empathic agents for mental health support [18, 26, 39].

2.2 Evaluating Empathic Systems: PETS

The growing trend toward empathic agents is also reflected in research on how to evaluate such systems, for example, in terms of effectiveness, therapeutic alliance [4, 35], or underlying concepts such as empathic interaction [12, 43]. While related research often applied adapted, unvalidated scales in this context, the Perceived Empathy of Technology Scale (PETS) offers a validated tool to measure how users perceive the empathic behavior of systems [43]. PETS consists of two factors, Emotional Responsiveness (PETS-ER) and Understanding and Trust (PETS-UT) with in total 10 items (see Table 2). The items in PETS-ER relate to a system's emotional understanding, expressions, and support, while the items in PETS-UT reflect a system's understanding of the user's goals, needs, and intentions, as well as its trustworthiness. The scale items are intended to be used with 101-point sliders ranging from "strongly disagree" to "strongly agree", and to be applied in randomized order [43]. Although PETS was developed using a bottom-up approach, the authors provide an interpretation of how items potentially relate to the cognitive and affective dimensions of empathy. This multidimensional definition is commonly used in empathy research, with cognitive empathy referring to the understanding and perception of another person's situation and affective empathy referring to the emotional reactions derived from that understanding [3, 15, 17]. The PETS was developed in English, and validated with 300 participants mainly from the European Economic Area [43]. As many of the empathic applications described in Section 2.1 are conversational interfaces aimed at providing emotional support, we argue that the evaluation should take place in the native language of the users, taking into account potential cross-cultural aspects of such interaction. Therefore, we see the importance of translating the original English PETS scale into different target languages.

Table 2: The original English version of PETS [43].

PETS-ER	Emotional Responsiveness
E1	The system considered my mental state.
E2	The system seemed emotionally intelligent.
E3	The system expressed emotions.
E4	The system sympathized with me.
E5	The system showed interest in me.
E6	The system supported me in coping with an emotional situation.
PETS-UT	**Understanding and Trust**
U1	The system understood my goals.
U2	The system understood my needs.
U3	I trusted the system.
U4	The system understood my intentions.

2.3 Cross-Cultural Aspects of Empathy

Various research indicates that the experience of emotions [20], emotion recognition [44] or concepts related to empathy such as self-compassion [6], personal distress and empathy concerns [9] may vary between cultures and languages. For example, Cassels et al. [9] found that East Asian participants reported greater personal distress and less empathic concern than Western subjects. Pavlenko [38] reviewed research on the effects of second languages on cognitive and affective processing and suggested that emotional responses are often stronger in one's native language. For example, Keysar et al. [31] suggest that communication in a foreign language creates a greater cognitive and emotional distance. Similarly, Ward and Ragosko [46] found that processing information in one's first

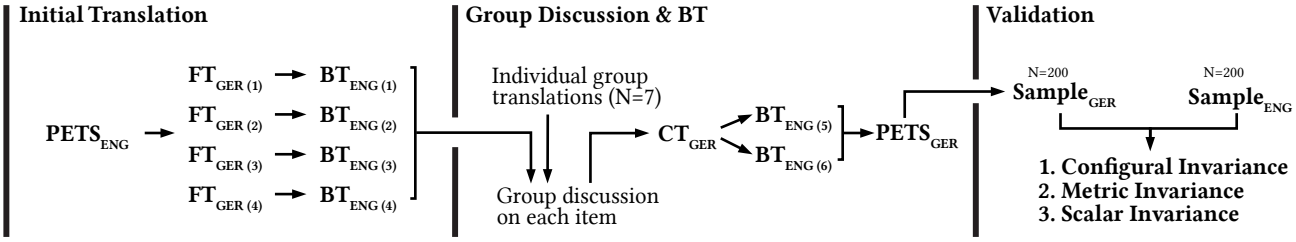


Figure 1: The process we applied to translate PETS from English (ENG) to German (GER) with multiple forward (FT) and back translation (BT) steps, and a group discussion to generate a consolidated translation (CT). The final German PETS was validated with Confirmatory Factor Analyses (CFA) for invariance testing, based on a German and English study sample (each N=200).

or native language results in higher scores on measures of empathy and emotional intelligence. With regard to the increasing of potential empathic systems in the area of mental health (Section 2.1), we also explored related research on the role of language in mental health support and therapy. Griner and Smith [25] published a widely-cited meta-analysis, describing that cultural and language adaptations positively affect the outcome of mental health interventions. For example, code-switching in multilingual therapy may help patients to express themselves more fully, encouraging therapists to provide such practices or ideally offer therapy in the client's primary language [13, 41]. As the research described above focuses primarily on cross-cultural aspects of emotional and empathic behavior in human-to-human interaction, we argue that the evaluation of artificial empathic systems, such as those used in digital mental health support, requires the use of validated measurement tools adapted to different languages and cultural contexts.

2.4 Translation for Cross-Cultural Research

In 1970, Brislin [7] introduced their back-translation process for cross-cultural translation. It involved four key steps: (1) translation from source to target language, (2) blind back-translation to source language, (3) comparison of versions to identify discrepancies, and (4) revision and iteration as needed. In their work, Brislin [7] established five criteria for translation quality and demonstrated that content type, difficulty level, and language similarity significantly affect translation outcomes. Thirty years later, Jones et al. [29] published a widely cited publication that suggested significant refinements to Brislin's back-translation model. Rather than the original sequential approach, they recommend simultaneous independent translations by multiple bilingual experts, followed by collaborative discussions to resolve discrepancies. Their six-step process includes concurrent translations, blind back-translations, group consensus meetings, and cross-lingual validation testing with bilingual participants, which helps uncover subtle differences in meaning between seemingly equivalent items.

Klotz et al. [32] provided a recent review of over 300 articles related to back-translation methodology and revealed significant shortcomings. They found that while back-translation was the dominant procedure for translating scales in organizational research, only 15.6% of the publications reported pretesting translated scales, with only 3.9% reporting quantitative evidence of equivalence. To address that lack of validation, the authors recommend conducting qualitative analysis (committee reviews, random-probe techniques) to identify problematic items, or to perform confirmatory factor analysis for invariance testing to statistically demonstrate that individuals would respond similarly to items regardless of the language.

3 Methodology

Our three-phased translation process (Figure 1) is based on the work of Brislin [7], Jones et al. [29] and Klotz et al. [32]. The actual scale translation consisted of two phases: the initial translation and a consolidating group discussion followed by a final back-translation. The third phase then covered the validation of the translated scale.

Initial Translation. For the initial translation phase, Jones et al. [29] recommend having two or more independent translators simultaneously creating target versions of the scale. We followed this recommendation and produced four independent forward translations and four independent back-translations from a total of eight professional English-German translators (Section 4.1).

Group Discussion & BT. In the second phase, experts with backgrounds in HCI and psychology, all fluent in English and German, discussed the initial translations (Section 4.2). The goal of this group discussion was to create a consolidated German scale, focusing on contextual meaning and consistent grammar. We then had a new group of independent translators back-translate this consolidated version for comparison with the original English scale. At this point, if consensus had not been reached in the group discussion or if the back-translations were not satisfactory, we had the option of repeating these steps, although this proved unnecessary (Section 4.3).

Validation. The third phase of our process involved statistical validation as recommended by Klotz et al. [32]. For that, we analyzed study data from two samples, one conducted in English and one in German (Section 5). Based on established guidelines for confirmatory factor analysis (CFA) and variance testing [10, 48], we selected a sample size of $N = 200$ for each language group to ensure adequate statistical power. For the two studies, we used the publicly available test scenarios from the original PETS development [43] and translated them to German for the new target sample. We then conducted CFA for invariance testing to determine whether the basic factor structure holds across both language versions (configural invariance) and to test the similarity of factor loadings (metric invariance). In addition, we followed the recommendation of Klotz et al. [32] and tested scalar invariance to see if the two samples reflect similar response styles.

Table 3: Overview of the experts' demographic background and their self-assessed language skills and expertise in related fields. The black squares represent the individual ratings on 5-point scales. Average values show a numerical representation (1-5).

	P1	P2	P3	P4	P5	P6	P7	Avg	SD
Age	28	28	30	29	30	30	29	29.14	0.83
Gender	female	female	female	female	male	male	male		
Degree	Master's	Master's	Doctoral	Master's	Doctoral	Master's	Master's		
Expertise assessment[1]									
Psychology	■■□□□	■■□□□	■■□□□	■■■■□	■■■■□	■■■■■	■■■□□	3.14	1.12
HCI	■■■■■	■■■□□	■■■■■	■■■■■	■■■■■	■■■■□	■■■■□	4.43	0.73
Affective systems	■■■■□	■■■□□	■□□□□	■■■■□	■■■■□	■■■■□	■■■■■	3.57	1.18
Emotion / Social / Behavioral theories	■■■■□	■■■□□	■■□□□	■■■■□	■■■■■	■■■■■	■■■■□	3.86	0.99
Empathy measurement or theories	■■□□□	■■■□□	■□□□□	■■■□□	■■□□□	■■■■□	■■■□□	2.57	0.90
Scientific scales application	■■■■■	■■■■□	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■	4.86	0.35
Scientific scales development	■■■■□	■■□□□	■■□□□	■■■■□	■■□□□	■■■■□	■■■■■	3.29	1.16
Language skill assessment[2]									
English reading	■■■■■	■■■■□	■■■■□	■■■■□	■■■■□	■■■■□	■■■■□	4.14	0.35
English writing	■■■■■	■■■■■	■■■■□	■■■■□	■■■■□	■■■■□	■■■■□	4.29	0.45
German reading	■■■■■	■■■■□	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■	4.86	0.35
German writing	■■■■■	■■■■□	■■■■■	■■■■■	■■■■■	■■■■■	■■■■■	4.86	0.35

[1] expertise rating scales: strongly disagree | disagree | neutral | agree | strongly agree
[2] language rating scales: Basic (A1-A2) | Intermediate (B1) | Upper Intermediate (B2) | Advanced (C1) | Mastery (C2)

4 Translation

For scale translation, we conducted forward and back-translations of the original scale, an expert group discussion for consolidation, and a final back-translation, as described in Section 3.

4.1 Initial Translation

To create the four forward translations (English to German) and the four back-translations (German to English), we recruited in total eight individual bilingual translators through the online platform *Fiverr*. The translators were required to be fluent in both languages and have at least a platform rating level of 2. Each translator was compensated for their service with between 8 € and 14 €, depending on individual experience and asking price. We instructed four translators to translate all scale items including the titles of the underlying factors to German, and not to use AI translation services. Subsequently, we let four different translators translate the German versions back to English, following the same process. The resulting forward (F1-F4) and back-translations (B1-B4) for each item and the factor titles are displayed in the appendix (Table 9 and Table 10).

4.2 Group Discussion

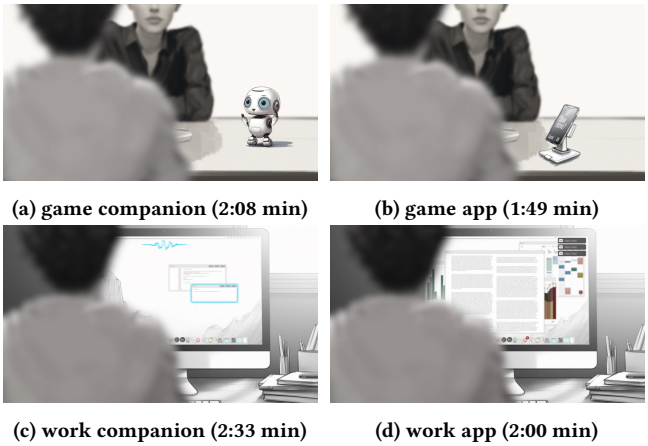
Based on the initial translations, we conducted an in-person group discussion with seven experts in HCI and psychology, facilitated by three of the authoring researchers. In total, the session lasted approximately 90 minutes.

Participants. Table 3 provides demographic information and self-assessed ratings of participants' background, expertise, and language skills. For language skill assessment, participants had to rate their English and German reading and writing skills on a 5-point scale reflecting CEFR language levels, ranging from Basic (A1-A2) to Mastery (C2). As with the development of the original PETS [43], we assessed the experts' backgrounds in affective systems, emotional, social, and behavioral theories, empathy-related constructs and measurement, as well as their experience with scientific scale development and application. For these assessments, we used 5-point rating scales (see Table 3). We deliberately selected participants based on these criteria, as contextual understanding was critical during the consolidation phase. In addition, participation required advanced proficiency in both English and German language.

Individual Translations. First, after an introduction to the overall goals and the English PETS, each expert was asked to independently translate the original scale, using prepared handouts. The aim of this step was to identify potentially new perspectives compared to the original translations of Section 4.1 and to further introduce experts to the scale. This step took approximately 15 minutes.

Group Discussion. The group then systematically examined each scale item separately, reviewing the forward and backward translations we had generated with external translators as described in Section 4.1. At this stage, each expert also shared their individual translation of the item. The primary goal of this collaborative discussion was to agree on a final translation of each item that accurately captured the meaning of the original. An additional goal was to maintain a consistent style across all items.

Results. During the discussion, several key language choices emerged. The group debated the use of simple past versus past perfect tense in German, ultimately choosing simple past. Word choice discussions included, for example, translating "intentions" (item U4) as either "Absicht" or "Intention". The group chose "Absicht" because it is more commonly used and easier to understand. A major discussion centered on how to translate "mental state" (item E1). The literal translation "mentaler Zustand" suggests more of a medical condition, so the group agreed on a more generic translation ("mentale Verfassung"). Another example was the translation



(a) game companion (2:08 min) (b) game app (1:49 min)

(c) work companion (2:33 min) (d) work app (2:00 min)

Figure 2: Screenshots of each of the four animated test scenarios: (a) an empathic robot companion that provides support during a board game, (b) a smartphone application for logging and view game information, (c) an empathic voice-based work assistant and (d) an notification and popup based office application. All scenarios included narrative audio tracks.

of "sympathized" (item E4). While three of four forward translations translated "sympathized" to showing "Mitgefühl", the group determined that this term would rather suggest compassion and chose to stay with the German term "Sympathie" in that item. Despite several other discussions on specific wording, a consensus was reached quickly, though sometimes against the more literal forward translations, highlighting the importance of this phase.

4.3 Final Back-Translation

Appendix Table 9 and Table 10 contain the consolidated translation (CT) resulting from the group discussion. Two independent translators (B5, B6) translated each of these German items back to English. As in the initial translation (Section 4.1), we recruited professional translators for this step through *Fiverr*. We then reviewed these back-translations and found that all of the items reflected the original meaning and, in most cases, also the literal wording perfectly well. Based on this evaluation, we entered the validation phase with the consolidated German PETS as depicted in Table 1.

5 Validation

We followed the design of the validation study from the original PETS development [43], using four different video scenarios of empathic and non-empathic systems, and created two different samples, one in English and one with the translated German scale.

5.1 Participants

To obtain two equally sized samples (for German and English, each $N = 200$), we recruited 300 new participants through Prolific and re-used data from a $N = 100$ sample from the original PETS validation [43]. For the newly recruited participants, we inferred language fluency through Prolific's pre-screening and qualitative analysis of the task summaries.

German Sample. We recruited $N = 200$ German-speaking participants from 28 different countries. The majority (171) resided in the European Economic Area (106 of them in Germany), 19 in North America, four in the Asia-Pacific region, three in South America, two in East Africa, and one from Southern Africa. The mean age was 39.3 years ($SD = 11.1$), with 96 participants identifying as female and 104 as male.

English Sample. For the English-speaking sample ($N = 200$), we combined data ($N = 100$) from the original, PETS validation run, as provided by Schmidmaier et al. [43], with data from another 100 newly recruited participants. This combined English sample consisted of participants from 20 different countries, with the majority of 167 participants residing in the European Economic Area, 29 in North America, two in the Asia-Pacific region, and one each from the Middle East and South America. The mean age was 35.2 years ($SD = 11.8$), with 95 participants identifying as female, 104 as male, and one as non-binary.

5.2 Material

For scale validation, we used the four task videos from the original PETS publication [43], depicting interactions with two empathic systems, a game companion robot (a) and a work companion application (c), and two non-empathic systems, a functional game support application (b) and an office work application (d) (see screenshots in Figure 2). For the German validation study, we translated the voice lines of the videos into German, using manual text translation and *OpenAI*'s text-to-speech generation. However, we made sure to stay close to the original voice styles, for example by choosing a robot voice in scenario (a). The newly generated video scenarios can be found in the *Supplementary Material*.

5.3 Procedure

The study procedure closely followed the original validation study [43], with all instructions being written in the corresponding target language (English for the original scale and German for the translated scale sample). Participation via Prolific was voluntary and could be terminated at any time. After explaining the study objectives and how the data would be processed, we obtained the consent of the participants and collected their demographic information. For the main task, each participant was randomly assigned to one of the four test scenarios (Section 5.2) and instructed to watch the corresponding video at least once. To ensure comprehension and the quality of responses, we asked participants to write a brief summary of the scenario after they finished watching. We then asked participants to rate their perception of the depicted system using the 10-item PETS, with items presented in random order. The sliders had an internal range from 0 to 100 and were labeled from *strongly disagree* to *strongly agree* in the English language group, and from *stimme überhaupt nicht zu* to *stimme voll und ganz zu* in the German version. On average, the study took 7.68 minutes ($SD = 4.12$) to complete in the combined English sample and 8.32 minutes ($SD = 4.88$) in the German sample. Participants were compensated with 1.10£.

Table 4: Median PETS scores across the four scenarios for the original and the translated scale (each N=200). The scenarios depicted two empathic companions (a, c) and two purely functional systems (b, d). PETS ratings ranged from 0..100.

	English (Original)					
	PETS		PETS-ER		PETS-UT	
Scenario	MD	IQR	MD	IQR	MD	IQR
(a) game comp.	80.60	25.80	81.25	24.04	77.50	29.44
(b) game app	34.50	24.25	17.92	29.79	58.75	22.50
(c) work comp.	69.65	35.02	69.25	35.17	71.50	30.06
(d) work app	14.10	18.90	7.42	22.58	23.25	28.25
	German (Translated)					
	PETS		PETS-ER		PETS-UT	
Scenario	MD	IQR	MD	IQR	MD	IQR
(a) game comp.	74.65	23.50	76.75	29.25	70.75	23.19
(b) game app	33.25	19.10	18.33	23.33	54.75	28.75
(c) work comp.	71.55	20.98	73.17	26.62	71.12	27.06
(d) work app	22.20	25.23	12.33	18.83	33.12	36.81

5.4 Results

Following the validation process described by Klotz et al. [32], we assessed the equivalence between the original and the translated PETS using a series of increasingly restrictive tests for measurement invariance. In addition, we tested internal consistency and examined descriptive statistics for each study scenario. The data can be found in the *Supplemental Material*. Table 4 shows the perceived empathy ratings for each scenario in both languages.

PETS within Samples. To examine differences between scenarios, we performed nonparametric analyses due to the non-normal distributions of the data (Shapiro-Wilk tests, $p < .05$). Kruskal-Wallis tests revealed significant differences in PETS ratings between scenarios for both language samples. The empathic scenarios (a) and (c) consistently received significantly higher PETS ratings than the non-empathic scenarios (b) and (c). Post hoc analyses with Dunn's test (Bonferroni-adjusted) revealed significant differences ($p < .05$) between empathic (a, c) and non-empathic (b, d) scenarios in almost all comparisons, confirming the intended effect of different levels of empathy between the scenarios. For the PETS-UT subscale in the English sample, the work companion was not rated significantly differently from the game app ($p = .122$). We discuss this effect further in Section 6.2. The detailed results can be found in the Appendix Table 8 and Table 7.

PETS between Samples. We further conducted Kruskal-Wallis tests on the combined sample to examine potential language group effects. The results did show no significant differences between language groups for PETS, PETS-ER, and PETS-UT (all $p >= .799$), while differences between scenarios were still significant across all measures (all $p < .001$). In addition, we performed Bayesian Mann-Whitney U tests (data augmentation, 5 chains of 1000 iterations) with JASP [28] to examine language group effects for each scenario. For scenarios (a), (b) and (c), the results (see Table 6) showed moderate evidence ($BF_{10} = 0.209 - 0.325$), for scenario (d) anecdotal evidence ($BF_{10} = 0.422 - 0.458$), that the results are the same, i.e. they come from the same population, regarding PETS and both subscales.

Invariance. We computed CFA with the *lavaan* package in R [42] and followed established guidelines [10, 27] for results interpretation. As shown in Table 5, the *configural invariance* model demonstrated a good fit ($CFI = .961$, $TLI = .949$, $RMSEA = .107$, $SRMR = .038$), indicating that the basic factor structure holds for both language versions. When constraining the factor loadings to be equal across groups, the model maintained a good fit ($CFI = .960$, $TLI = .952$, $RMSEA = .103$, $SRMR = .056$). The change in fit indices was minimal ($\Delta CFI = -.001$, $\Delta RMSEA = -.004$), well within the recommended thresholds ($\Delta CFI < .01$, $\Delta RMSEA < .015$), supporting *metric invariance*. This suggests that items function similarly across language versions and that relationships between latent constructs can be meaningfully compared.

Figure 3 shows the factor structure and standardized loadings for both scale versions. While the factor loadings were largely comparable, item U3 showed a substantially lower loading in the German version (0.47) compared to the English version (0.63). We suggest that this localized difference did not compromise the overall metric invariance, as evidenced by the acceptable changes in fit indices when constraints were imposed, and the strong psychometric properties maintained by the overall scale and subscales in both languages (Section 6.2). The *scalar invariance* model, with both loadings and intercepts constrained to be equal, also demonstrated good fit ($CFI = .960$, $TLI = .957$, $RMSEA = .098$, $SRMR = .056$). Comparison with the metric model showed negligible changes in fit indices ($\Delta CFI = .000$, $\Delta RMSEA = -.005$), providing strong evidence for scalar invariance. This indicates that not only the factor loadings are equivalent across language versions, but also the item intercepts are comparable, allowing valid comparisons of latent means across scales. These results provide robust evidence of measurement equivalence between the original and translated PETS, supporting its use for cross-cultural comparisons. Although the RMSEA values across all models were slightly above the conventional threshold of .08 [27], the excellent CFI and TLI values, together with the low SRMR values, provide substantial evidence of good model fit.

Internal Consistency. To assess internal consistency, we calculated Cronbach's alpha for both the total scale and the subscales in each language. The original English scale demonstrated excellent internal consistency ($\alpha = .956$), which was closely matched by the translated German version ($\alpha = .950$). The PETS-ER subscale showed excellent reliability in both the original ($\alpha = .960$) and translated versions ($\alpha = .955$). The PETS-UT subscale demonstrated good reliability in both the original ($\alpha = .877$) and translated versions ($\alpha = .853$). These results indicate that the translated version maintains comparable internal consistency to the original scale, providing evidence of the psychometric quality of the translation. The slight decrease in reliability for the PETS-UT subscale in the translated version ($\Delta\alpha = .024$) is minimal and still within the range considered good for research purposes [23].

Table 5: Measurement invariance testing results for the PETS across two samples with original (N=200) and translated (N=200) versions. Model comparisons show the difference between metric and configural as well as scalar and metric model results.

	Model Fit Indices						Model Comparisons			
Model	χ^2	df	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$	Δdf	ΔCFI	ΔRMSEA
1. Configural Invariance	222.56	68	0.961	0.949	0.107	0.038	—	—	—	—
2. Metric Invariance	237.65	76	0.960	0.952	0.103	0.056	15.09	8	-0.001	-0.004
3. Scalar Invariance	244.29	84	0.960	0.957	0.098	0.056	6.63	8	0.000	-0.005

6 Discussion

In this work, we introduced and applied a combined methodological process (see Section 3) to translate and validate the Perceived Empathy of Technology Scale (PETS) to German. In the following, we discuss validation results, discrepancies, cross-cultural implications and limitations, and reflect on the overall methodology.

6.1 Translation Equivalence and Validity

Our comprehensive validation approach addressed the methodological concerns raised by Klotz et al. [32], regarding the frequent lack of quantitative validation of translated scales. Measurement invariance testing provided robust evidence of equivalence between the original English and translated German versions of the PETS. As described in Section 5.4, the configural invariance model demonstrated a good fit, confirming that the factor structure of the PETS holds in both language versions. This finding is particularly important because it confirms that German-speaking users conceptualize perceived empathy in technological systems similarly to English-speaking users. Metric invariance testing further strengthened our validation by showing that factor loadings were equivalent across versions, with minimal changes in fit indices (see Table 5). This suggests that the items function similarly across the two languages and that relationships between latent constructs can be meaningfully compared between the two versions of the scale. The scalar invariance results provide evidence that not only factor loadings but also item intercepts are comparable, allowing for valid comparisons of latent means across language groups.

As noted in Section 5.4, the RMSEA values for all models (ranging from .098 to .107) were slightly above the conventional threshold of .08 suggested by Hu and Bentler [27]. Although this may initially raise concerns, the following methodological considerations contextualize these results. First, RMSEA tends to penalize simpler models with fewer degrees of freedom, which is relevant for our two-factor model with 10 items, and 68 to 84 degrees of freedom [30]. Second, the excellent values for CFI ($\geq$ .960) and TLI ($\geq$ .949), along with the consistently low SRMR values ($\leq$ .056), provide substantial counter-evidence for a good model fit. This consistent pattern across all invariance levels, with minimal changes in fit indices between models, further supports the validity of our findings despite the slightly elevated RMSEA. Finally, the excellent internal consistency values for both the overall PETS and its subscales provide additional support for its reliability. Therefore, we argue that the German PETS is a valid and reliable instrument for assessing perceived empathy in systems among German-speaking users, allowing researchers and practitioners to conduct cross-cultural studies.

6.2 PETS-UT Subscale Discrepancy

While the emotional responsiveness subscale (PETS-ER) consistently discriminated between empathic and non-empathic systems in both languages, the understanding and trust subscale (PETS-UT) showed a discrepancy in the English sample. Specifically, participants did not rate the work companion significantly differently from the game app on this subscale (p = .122), even though they represented empathic and non-empathic systems, respectively. Furthermore, as shown in Figure 3, item U3 ("Ich vertraute dem System") showed a significantly lower loading in the German version (0.47) compared to the original English version (0.63).

We suggest that this finding may reflect cultural and individual nuances in how trust in a system is defined and perceived. This interpretation can be supported by qualitative insights from participants' task summaries: in the work companion scenario (c), some participants described the assistant as overbearing, manipulative, or even intrusive - despite its helpfulness - while in the game app scenario (b), participants expressed skepticism or irritation about technology in leisure contexts, possibly indicating a preference for autonomy and low-tech interaction in that context. Also, the lack of interactivity in the third-person test scenarios is a limitation already described by Schmidmaier et al. [43], which we believe potentially weakens the required experience of trust.

While we still decided to follow the same approach for the sake of comparability of the validation, we envision further validation with interactive systems in the future, with an additional focus on the PETS-UT subscale, especially on item U3. Researchers using the German PETS should be aware of these discrepancies when interpreting results, especially when they are focusing on the understanding and trust dimension.

6.3 Cross-Cultural Considerations

Our translation process revealed several cross-cultural and linguistic considerations. German and English differ substantially in their grammatical structures, vocabulary nuances, and emotional expressions, all of which influenced the translation decisions during the group discussion and the evaluation of the final back-translation. A primary consideration was the choice of verb tense. German offers several forms of the past tense, and our experts debated between simple past and past perfect (Section 4.2). Although the group found both forms to be valid, simple past was chosen for the final consolidated version to keep the scale concise. During the translation process *grammatical consistency* across all items also emerged as a key feature recommended for scale translation in general. The

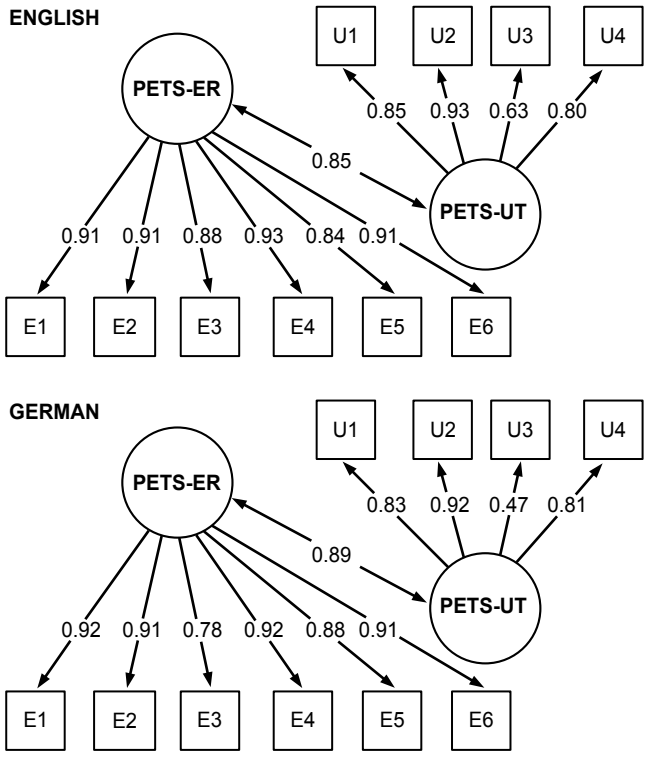


Figure 3: Factor structure and standardized loadings of the English PETS (left) and the translated German version (right).

most substantive discussions centered on the translation of emotional and empathic terminology. As described in Section 4.2, the translation of "mental state" in item E1 caused considerable debate, as the literal translation might carry a more clinical or medical connotation in German. The group agreed on "mentale Verfassung" as a more neutral and appropriate alternative that better preserves the intended meaning. As another example, for the factor title "Emotional Responsiveness", experts considered several German terms, including "Reaktion", "Empfänglichkeit", and "Reaktivität" before settling on "Reaktionsfähigkeit" as best capturing the outward expression of a system's emotional capacity. Perhaps the most controversial was the discussion around item E4's "sympathized". While three of the four forward translations rendered this as "Mitgefühl zeigen" (showing compassion), the expert group recognized that this implied a deeper emotional connection than the English term and ultimately chose "Sympathie". These examples emphasize the importance of the group discussion phase, and the focus on contextual and meaningful rather than literal translations.

6.4 Methodological Reflections and Use of AI

Based on our validation results, we suggest that our application of a systematic back-translation process proved highly effective in developing an equivalent German version of PETS. Especially the expert discussion phase allowed us to identify and resolve potential issues that might have been missed with a simpler translation approach. This confirms Jones et al. [29]'s claim that collaborative expert discussion can reveal differences in meaning that might otherwise go unnoticed. Furthermore, as also explored by Chung and Kim [11], we experimented with AI-based forward and back-translation, using *ChatGPT*, *Claude*, and *DeepL*. However, as some of the translations were too literal, we argue that human expertise is still essential for consolidation and evaluation, especially for instruments measuring complex psychological constructs such as empathy. In future projects, a hybrid approach could combine the efficiency of AI translation in the initial phase with human expert review and empirical validation to ensure quality and validity. For such an approach, we recommend elaborate contextual prompting, which in turn would require prior consideration of terminology or language effects. Contextual AI translation could therefore take place after the group discussion, for example.

6.5 Limitations and Future Directions

Despite our rigorous approach, we acknowledge several limitations. As with the original PETS validation, we used third-person video scenarios, although this approach may not fully capture how users would perceive empathy in direct interactions with a system. Future research should further validate the German PETS in more interactive settings. Furthermore, our samples, while adequately sized for psychometric analyses, were recruited through Prolific, which may introduce self-selection bias. In addition, the diverse geographical distribution of the participants harbors the risk that the results are influenced not only by language, but also by regional cultural differences. Finally, the slightly elevated RMSEA values (Section 6.1), while offset by excellent CFI and TLI scores, suggest that further refinement of the scale may be beneficial. In particular, the lower loading of item U3 in the German version indicates an area for potential improvement. Future validation might also include further tests, for example, regarding convergent validity [43].

Looking ahead, the German PETS opens up new research opportunities in German-speaking areas, especially in the evaluation of empathic technologies for mental health support. Furthermore, we plan to apply our translation approach to adapt PETS to other languages, such as Japanese or Chinese, to further enable cross-cultural research on empathic technology perception, for example, regarding social robots.

7 Conclusion

In this paper, we present a systematic, empirically validated translation of the Perceived Empathy of Technology Scale (PETS) from English to German. Our comprehensive back-translation process, involving multiple independent translations, expert group discussions, and rigorous psychometric validation based on a German (N = 200) and an English sample (N = 200), resulted in a German version of the PETS that maintains equivalence to the original scale across configural, metric, and scalar levels of invariance. This implies two important contributions to HCI research. First, we provide researchers and practitioners with a validated instrument for assessing empathic systems with German-speaking users. Second, our translation process offers methodological insights for cross-cultural scale adaptation in HCI. By combining qualitative expertise with quantitative validation, we demonstrate how to develop translations that preserve both semantic meaning and measurement properties.

Author Contributions

Matthias Schmidmaier: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Lukas Schöberl**: Data curation, Investigation, Methodology, Software, Writing – original draft; **Jonathan Rupp**: Formal Analysis, Validation, Writing – original draft, Writing – review & editing; **Sven Mayer**: Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing – original draft, Writing – review & editing

Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

References

[1] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and Davey M Smith. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183, 6 (jun 2023), 589–596. doi:10.1001/jamainternmed.2023.1838

[2] Petter Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 257, 13 pages. doi:10.1145/3411764.3445318

[3] C. Daniel Batson. 2009. These things called empathy: Eight related but distinct phenomena. *The social neuroscience of empathy.* 255 (2009), 3–15. doi:10.7551/mitpress/9780262012973.003.0002

[4] Clare Beatty, Tanya Malik, Saha Meheli, and Chaitali Sinha. 2022. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): A mixed-methods study. *Front. Digit. Health* 4 (April 2022), 847991. doi:10.3389/fdgth.2022.847991

[5] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. doi:10.1145/1067860.1067867

[6] M Birkett. 2014. Self-compassion and empathy across cultures: Comparison of young adults in China and the United States. *International Journal of Research Studies in Psychology* 3, 3 (2014), 25–34. doi:10.5861/ijrsp.2013.551

[7] Richard W Brislin. 1970. Back-translation for cross-cultural research. *J. Cross. Cult. Psychol.* 1, 3 (Sept. 1970), 185–216. doi:10.1177/135910457000100301

[8] Johana Cabrera, M Soledad Loyola, Irene Magaña, and Rodrigo Rojas. 2023. Ethical dilemmas, mental health, artificial intelligence, and LLM-based chatbots. In *Lecture Notes in Computer Science*. Springer Nature Switzerland, Cham, 313–326. doi:10.1007/978-3-031-34960-7_22

[9] Tracy G Cassels, Sherilynn F Chan, and Winnie W Chung. 2010. The role of culture in affective empathy: Cultural and bicultural differences. *Journal of Cognition and Culture* 10 (2010), 309–326. doi:10.1163/156853710x531203

[10] Fang Fang Chen. 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 14, 3 (July 2007), 464–504. doi:10.1080/10705510701301834

[11] Ji-Bum Chung and Taehyun Kim. 2025. Leveraging Large Language Models for Enhanced Back-Translation: Techniques and Applications. *IEEE Access* 13 (2025), 61322–61328. doi:10.1109/ACCESS.2025.3557014

[12] Shauna Concannon and Marcus Tomalin. 2023. Measuring perceived empathy in dialogue systems. *AI Soc.* 39, 5 (July 2023), 2233–2247. doi:10.1007/s00146-023-01715-z

[13] Beverley Costa. 2010. Mother tongue or non-native language? Learning from conversations with bilingual/multilingual therapists about working with clients who do not share their native language. *Ethn. Inequal. Health Soc. Care* 3, 1 (March 2010), 15–24. doi:10.5042/eihsc.2010.0144

[14] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F. Jung, Nicola Dell, Deborah Estrin, and James A. Landay. 2024. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 446, 18 pages. doi:10.1145/3613904.3642336

[15] Benjamin M. P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. 2016. Empathy: A Review of the Concept. *Emot. Rev.* 8, 2 (apr 2016), 144–153. doi:10.1177/1754073914558466

[16] Karl Daher, Jacky Casas, Omar Abou Khaled, and Elena Mugellini. 2020. Empathic Chatbot Response for Medical Assistance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3383652.3423864

[17] Mark H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology* 44, 1 (1983), 113. doi:10.1037/0022-3514.44.1.113

[18] Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. doi:10.48550/arXiv.2311.14693

[19] Mauro de Gennaro, Eva G Krumhuber, and Gale Lucas. 2019. Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. *Front. Psychol.* 10 (2019), 3061. doi:10.3389/fpsyg.2019.03061

[20] Jean-Marc Dewaele. 2010. *Emotions in Multiple Languages* (1 ed.). Palgrave Macmillan, Basingstoke, England. doi:10.1002/9781405198431.wbeal0795

[21] Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. ChatGPT outperforms humans in emotional awareness evaluations. *Front. Psychol.* 14 (may 2023), 1199058. doi:10.3389/fpsyg.2023.1199058

[22] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment. Health* 4, 2 (jun 2017), e19. doi:10.2196/mental.7785

[23] Darren George and Paul Mallery. 2024. *IBM SPSS statistics 29 step by step: A simple guide and reference.* Routledge, New York. doi:10.4324/9781032622156

[24] S Gerke, A Stern, and T Minssen. 2020. Germany's digital health reforms in the COVID-19 era: lessons and opportunities for other countries. *NPJ Digit. Med.* 3 (July 2020), 94. doi:10.1038/s41746-020-0306-7

[25] Derek Griner and Timothy B Smith. 2006. Culturally adapted mental health intervention: A meta-analytic review. *Psychotherapy* 43, 4 (2006), 531–548.

[26] M D Romael Haque and Sabirat Rubya. 2023. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. *JMIR Mhealth Uhealth* 11 (22 May 2023), e44838. doi:10.2196/44838

[27] Li-Tze Hu and Peter M Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Modeling* 6, 1 (Jan. 1999), 1–55. doi:10.1080/10705519909540118

[28] JASP Team. 2025. JASP (Version 0.19.3)[Computer software]. https://jasp-stats.org/

[29] Patricia S Jones, J Lee, L Phillips, X Zhang, and K Jaceldo. 2001. An adaptation of Brislin's translation model for cross-cultural research. *Nurs. Res.* 50, 5 (Sept. 2001), 300–304. doi:10.1097/00006199-200109000-00008

[30] David A Kenny and D Betsy McCoach. 2003. Effect of the number of variables on measures of fit in structural equation modeling. *Struct. Equ. Modeling* 10, 3 (July 2003), 333–351. doi:10.1207/S15328007SEM1003_1

[31] Boaz Keysar, Sayuri L Hayakawa, and Sun Gyu An. 2012. The foreign-language effect: thinking in a foreign tongue reduces decision biases: Thinking in a foreign tongue reduces decision biases. *Psychol. Sci.* 23, 6 (June 2012), 661–668. doi:10.1177/0956797611432178

[32] Anthony C Klotz, Brian W Swider, and Seo Hyun Kwon. 2023. Back-translation practices in organizational research: Avoiding loss in translation. *J. Appl. Psychol.* 108, 5 (May 2023), 699–727. doi:10.1037/apl0001050

[33] Bingjie Liu and S Shyam Sundar. 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychol. Behav. Soc. Netw.* 21, 10 (oct 2018), 625–636. doi:10.1089/cyber.2018.0110

[34] Siru Liu, Allison B McCoy, Aileen P Wright, Babatunde Carew, Julian Z Genkins, Sean S Huang, Josh F Peterson, Bryan Steitz, and Adam Wright. 2024. Leveraging large language models for generating responses to patient messages—a subjective analysis. *Journal of the American Medical Informatics Association* 31, 6 (may 2024), 1367–1379. doi:10.1101/2023.07.14.23292669

[35] Amylie Malouin-Lachance, Julien Capolupo, Chloé Laplante, and Alexandre Hudon. 2025. Does the digital therapeutic alliance exist? Integrative review. *JMIR Ment. Health* 12, 1 (Feb. 2025), e69294. doi:10.2196/69294

[36] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in Virtual Agents and Robots: A Survey. *ACM Trans. Interact. Intell. Syst.* 7, 3 (Sept. 2017), 1–40. doi:10.1145/2912150

[37] Sung Park and Mincheol Whang. 2022. Empathy in Human-Robot Interaction: Designing for Social Robots. *Int. J. Environ. Res. Public Health* 19, 3 (Feb. 2022), 1889. doi:10.3390/ijerph19031889

[38] Aneta Pavlenko. 2012. Affective processing in bilingual speakers: disembodied cognition? *Int. J. Psychol.* 47, 6 (2012), 405–428. doi:10.1080/00207594.2012.743665

[39] Kay T Pham, Amir Nabizadeh, and Salih Selek. 2022. Artificial intelligence and chatbots in psychiatry. *Psychiatr. Q.* 93, 1 (mar 2022), 249–253. doi:10.1007/s11126-022-09973-8

[40] Julie Prescott and Terry Hanley. 2023. Therapists' attitudes towards the use of AI in therapeutic practice: considering the therapeutic alliance. *Ment. Health Soc. Incl.* 27, 2 (may 2023), 177–185. doi:10.1108/MHSI-02-2023-0020

[41] Louise Rolland, Jean-Marc Dewaele, and Beverley Costa. 2017. Multilingualism and psychotherapy: exploring multilingual clients' experiences of language practices in psychotherapy. *Int. J. Multiling.* 14, 1 (Jan. 2017), 69–85. doi:10.1080/14790718.2017.1259009

[42] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. doi:10.18637/jss.v048.i02

[43] Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. 2024. Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 456, 18 pages. doi:10.1145/3613904.3642035

[44] Helena S Schmitt, Cornelia Sindermann, Mei Li, Yina Ma, Keith M Kendrick, Benjamin Becker, and Christian Montag. 2020. The dark side of emotion recognition - evidence from cross-cultural research in Germany and China. *Front. Psychol.* 11 (July 2020), 1132. doi:10.3389/fpsyg.2020.01132

[45] Lennart Seitz. 2024. Artificial empathy in healthcare chatbots: Does it feel authentic? *Computers in Human Behavior: Artificial Humans* 2, 1 (jan 2024), 100067. doi:10.1016/j.chbah.2024.100067

[46] Rebecca Ward and Malgozata Ragosko. 2025. Does language experience and bilingualism shape empathy and emotional intelligence? *Int. J. Billing.* 0, 0 (Jan. 2025), 13670069241308078. doi:10.1177/13670069241308078

[47] Jeremy J Webb. 2023. Proof of concept: Using ChatGPT to teach emergency physicians how to break bad news. *Cureus* 15, 5 (may 2023), e38755. doi:10.7759/cureus.38755

[48] Erika J Wolf, Kelly M Harrington, Shaunna L Clark, and Mark W Miller. 2013. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety: An evaluation of power, bias, and solution propriety. *Educ. Psychol. Meas.* 76, 6 (Dec. 2013), 913–934. doi:10.1177/0013164413495237

[49] Refael Yonatan-Leus and Hadas Brukner. 2025. Comparing perceived empathy and intervention strategies of an AI chatbot and human psychotherapists in online mental health support. *Couns. Psychother. Res.* 25, 1 (mar 2025), 0. doi:10.1002/capr.12832

A Appendix

The appendix provides tables with detailed information on PETS score statistics for the validation study (Table 7), including scenario-wise post-hoc tests (Table 8) and Bayesian Mann-Whitney U test over both samples (Table 6), as well as the individual forward / back-translations for the initial translation and the group discussion (Table 9 and Table 10).

Table 6: Results of Bayesian Mann-Whitney U tests (data augmentation, 5 chains of 1000 iterations) over combined data from both samples, examining language group effects for each scenario.

Measure	Scenario	BF_{10}	W	Rhat
PETS	(a) game companion	0.277	1410.500	1.001
PETS-ER	(a) game companion	0.325	1417.000	1.001
PETS-UT	(a) game companion	0.234	1364.500	1.001
PETS	(b) game app	0.213	1315.500	1.000
PETS-ER	(b) game app	0.225	1226.000	1.000
PETS-UT	(b) game app	0.241	1373.500	1.003
PETS	(c) work companion	0.224	1238.000	1.000
PETS-ER	(c) work companion	0.225	1224.000	1.001
PETS-UT	(c) work companion	0.209	1247.500	1.001
PETS	(d) work app	0.440	1023.500	1.000
PETS-ER	(d) work app	0.422	1054.000	1.001
PETS-UT	(d) work app	0.458	1025.500	1.002

Table 7: Kruskal-Wallis test results for differences across scenarios within language samples.

Measure	Language	Test Statistic	p-value
PETS	original	109.71	<.001
PETS-ER	original	121.18	<.001
PETS-UT	original	71.38	<.001
PETS	translated	120.03	<.001
PETS-ER	translated	127.71	<.001
PETS-UT	translated	72.01	<.001

Table 8: Post-hoc comparison results between scenarios within each language sample (Dunn's Test with Bonferroni Correction) with (a) game companion, (b) game app, (c) work companion, and (d) work app.

Comparison	Measure	Language	Z	p-value	
(b) vs. (a)	PETS	original	−6.46	<.001	***
(b) vs. (a)	PETS	translated	−7.21	<.001	***
(b) vs. (a)	PETS-ER	original	−7.75	<.001	***
(b) vs. (a)	PETS-ER	translated	−8.10	<.001	***
(b) vs. (a)	PETS-UT	original	−3.37	.002	**
(b) vs. (a)	PETS-UT	translated	−4.25	<.001	***
(b) vs. (d)	PETS	original	2.59	.029	*
(b) vs. (d)	PETS	translated	1.74	.244	
(b) vs. (d)	PETS-ER	original	1.12	.786	
(b) vs. (d)	PETS-ER	translated	.60	1.000	
(b) vs. (d)	PETS-UT	original	4.51	<.001	***
(b) vs. (d)	PETS-UT	translated	3.42	.002	**
(b) vs. (c)	PETS	original	−5.19	<.001	***
(b) vs. (c)	PETS	translated	−6.29	<.001	***
(b) vs. (c)	PETS-ER	original	−6.52	<.001	***
(b) vs. (c)	PETS-ER	translated	−7.21	<.001	***
(b) vs. (c)	PETS-UT	original	−2.05	.122	
(b) vs. (c)	PETS-UT	translated	−3.22	.004	**
(a) vs. (d)	PETS	original	9.05	<.001	***
(a) vs. (d)	PETS	translated	8.95	<.001	***
(a) vs. (d)	PETS-ER	original	8.87	<.001	***
(a) vs. (d)	PETS-ER	translated	8.70	<.001	***
(a) vs. (d)	PETS-UT	original	7.88	<.001	***
(a) vs. (d)	PETS-UT	translated	7.67	<.001	***
(a) vs. (c)	PETS	original	1.27	.613	
(a) vs. (c)	PETS	translated	.93	1.000	
(a) vs. (c)	PETS-ER	original	1.23	.658	
(a) vs. (c)	PETS-ER	translated	.89	1.000	
(a) vs. (c)	PETS-UT	original	1.33	.554	
(a) vs. (c)	PETS-UT	translated	1.02	.918	
(d) vs. (c)	PETS	original	−7.78	<.001	***
(d) vs. (c)	PETS	translated	−8.03	<.001	***
(d) vs. (c)	PETS-ER	original	−7.64	<.001	***
(d) vs. (c)	PETS-ER	translated	−7.81	<.001	***
(d) vs. (c)	PETS-UT	original	−6.56	<.001	***
(d) vs. (c)	PETS-UT	translated	−6.64	<.001	***

Table 9: The results for PETS-ER based on four individual forward translations (F1-F4), four individual backward translations (B1-B4), the consolidated group translation (CT), and its back-translations (B5, B6). The final selection is printed in bold.

	Forward-Translations		Back-Translations
	PETS-ER: "Emotional Responsiveness"		
F1	Emotionale Reaktionsfähigkeit	B1	Emotional reactivity
F2	Emotionale Reaktionsfähigkeit	B2	Emotional Responsiveness
F3	Emotionale Reaktionsfähigkeit	B3	Emotional Responsiveness
F4	Emotionale Reaktionsfähigkeit	B4	Emotional Responsiveness
CT	**Emotionale Reaktionsfähigkeit**	B5	Emotional Responsiveness
		B6	Emotional Reactivity
	PETS-E1: "The system considered my mental state."		
F1	Das System berücksichtigte meinen mentalen Zustand.	B1	The system considered my mental state.
F2	Das System berücksichtigte meinen mentalen Zustand.	B2	The system took my mental state into account.
F3	Das System berücksichtigte mein mentales Befinden.	B3	The system took my mental state into account.
F4	Das System berücksichtigte meinen mentalen Zustand.	B4	The system took into consideration my mental state.
CT	**Das System berücksichtigte meine mentale Verfassung.**	B5	The system considered my mental state.
		B6	The system took my mental state into account.
	PETS-E2: "The system seemed emotionally intelligent."		
F1	Das System wirkte emotional intelligent.	B1	The system appeared emotionally intelligent.
F2	Das System schien emotional intelligent zu sein.	B2	The system appeared to be emotionally intelligent.
F3	Das System zeigte emotionale Intelligenz.	B3	The system showed emotional intelligence.
F4	Das System schien emotional intelligent zu sein.	B4	The system seemed to be emotionally intelligent.
CT	**Das System wirkte emotional intelligent.**	B5	The system appeared emotionally intelligent.
		B6	The system appeared emotionally intelligent.
	PETS-E3: "The system expressed emotions."		
F1	Das System hat Emotionen ausgedrückt.	B1	The system expressed emotions.
F2	Das System drückte Emotionen aus.	B2	The system expressed emotions.
F3	Das System hatte Gefühle zum Ausdruck gebracht.	B3	The system had expressed feelings.
F4	Das System drückte Emotionen aus.	B4	The system expressed emotions.
CT	**Das System drückte Emotionen aus.**	B5	The system expressed emotions.
		B6	The system expressed emotions.
	PETS-E4: "The system sympathized with me."		
F1	Das System empfand Mitgefühl mit mir.	B1	The system empathized with me.
F2	Das System zeigte Mitgefühl mit mir.	B2	The system showed compassion towards me.
F3	Das System zeigte Mitgefühl.	B3	The system showed compassion.
F4	Das System sympathisierte mit mir.	B4	The system sympathised with me.
CT	**Das System zeigte Sympathie mir gegenüber.**	B5	The system showed sympathy towards me.
		B6	The system showed sympathy towards me.
	PETS-E5: "The system showed interest in me."		
F1	Das System zeigte Interesse an meiner Person.	B1	The system showed interest in my person.
F2	Das System zeigte Interesse an mir.	B2	The system showed interest in me.
F3	Das System hatte sich für mich interessiert.	B3	The system had taken an interest in me.
F4	Das System zeigte Interesse an mir.	B4	The system showed interest in me.
CT	**Das System zeigte Interesse an mir.**	B5	The system showed interest in me.
		B6	The system showed interest in me.
	PETS-E6: "The system supported me in coping with an emotional situation."		
F1	Das System hat mich dabei unterstützt, mit einer emotionalen Situation umzugehen.	B1	The system supported me in dealing with an emotional situation.
F2	Das System unterstützte mich beim Bewältigen einer emotionalen Situation.	B2	The system supported me in dealing with an emotional situation.
F3	Das System half mir, eine emotionale Situation zu bewältigen.	B3	The system helped me to deal with an emotional situation.
F4	Das System unterstützte mich bei der Bewältigung einer emotionalen Situation.	B4	The system supported me in managing an emotional situation.
CT	**Das System unterstützte mich dabei, mit einer emotionalen Situation umzugehen**	B5	The system supported me in dealing with an emotional situation
		B6	The system supported me in dealing with an emotional situation

Table 10: The results for PETS-UT based on four individual forward translations (F1-F4), four individual backward translations (B1-B4), the consolidated group translation (CT), and its back-translations (B5, B6). The final selection is printed in bold.

	Forward-Translations		Back-Translations
	PETS-UT: "Understanding and Trust"		
F1	Verständnis und Vertrauen	B1	Understanding and trust
F2	Verständnis und Vertrauen	B2	Understanding and Trust
F3	Verständnis und Vertrauen	B3	Understanding and trust
F4	Verständnis und Vertrauen	B4	Understanding and Trust
CT	**Verständnis und Vertrauen**	B5	Understanding and trust
		B6	Understanding and trust
	PETS-U1: "The system understood my goals."		
F1	Das System hat meine Ziele verstanden.	B1	The system understood my goals.
F2	Das System verstand meine Ziele.	B2	The system understood my goals.
F3	Das System verstand meine Ziele.	B3	The system understood my goals.
F4	Das System verstand meine Ziele.	B4	The system understood my goals.
CT	**Das System verstand meine Ziele**	B5	The system understood my goals
		B6	The system understood my goals
	PETS-U2: "The system understood my needs."		
F1	Das System hat meine Bedürfnisse verstanden.	B1	The system understood my needs.
F2	Das System verstand meine Bedürfnisse.	B2	The system understood my needs.
F3	Das System hatte Verständnis für meine Bedürfnisse.	B3	The system understood my needs.
F4	Das System verstand meine Bedürfnisse.	B4	The system understood my needs.
CT	**Das System verstand meine Bedürfnisse**	B5	The system understood my needs
		B6	The system understood my needs
	PETS-U3: "I trusted the system."		
F1	Ich konnte dem System vertrauen.	B1	I was able to trust the system.
F2	Ich vertraute dem System.	B2	I trusted the system.
F3	Dem System vertraute ich.	B3	I trusted the system.
F4	Ich vertraute dem System.	B4	I trusted the system.
CT	**Ich vertraute dem System**	B5	I trusted the system
		B6	I trusted the system
	PETS-U4: "The system understood my intentions."		
F1	Das System hat meine Absichten verstanden.	B1	The system understood my intentions.
F2	Das System verstand meine Absichten.	B2	The system understood my intentions.
F3	Das System konnte meine Intentionen nachvollziehen.	B3	The system was able to understand my intentions.
F4	Das System verstand meine Absichten.	B4	The system understood my intentions.
CT	**Das System verstand meine Absichten**	B5	The system understood my intentions
		B6	The system understood my intentions