Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User

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

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

Darina Cvetanova
LMU Munich
Munich, Germany
darinatsvet@gmail.com

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

ABSTRACT

Affective computing improves rapidly, allowing systems to process human emotions. This enables systems such as conversational agents or social robots to show empathy toward users. While there are various established methods to measure the empathy of humans, there is no reliable and validated instrument to quantify the perceived empathy of interactive systems. Thus, we developed the Perceived Empathy of Technology Scale (PETS) to assess and compare how empathic users perceive technology. We followed a standardized multi-phase process of developing and validating scales. In total, we invited 30 experts for item generation, 324 participants for item selection, and 396 additional participants for scale validation. We developed our scale using 22 scenarios with opposing empathy levels, ensuring the scale is universally applicable. This resulted in the PETS, a 10-item, 2-factor scale. The PETS allows designers and researchers to evaluate and compare the perceived empathy of interactive systems rapidly.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

human-computer interaction, empathy, technology, scale

ACM Reference Format:
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 CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642035

1 INTRODUCTION

Recently, we have seen a rapid development in affective computing, with machines enhancing interaction through human-like emotional expression. This allows systems such as conversational agents or social robots to empathize with users. In particular, with recent advances in artificial intelligence, namely the development of large-scale language models, the ability of conversational agents to conduct elegant conversations has dramatically improved. This is already helping to advance the field of affective computing in the expression and communication of simulated human emotions, allowing systems to appear empathic to the user and provide emotional support [16, 31, 54, 80]. While there are several established methods for measuring human empathy [30, 32, 81], there is no reliable and validated instrument for quantifying the perceived empathy of interactive systems.

Therefore, developers and researchers of empathic systems have mainly used scales initially designed to measure human empathy or related concepts, adapting the item selection, wording, or user perspective [21, 22, 25, 45, 57, 88, 89]. As these scales were not designed to measure the perceived empathy of intelligent systems, it is questionable whether they are valid in this context. A larger set of previous work has modified existing scales to make them suitable for assessing systems [22, 45, 45, 57, 88, 89]. Specifically, the authors selected or modified individual scale items in order to use them in specific contexts and for measuring systems' empathy. Concannon and Tomalin [25] took the first steps in creating a new scale focused on conversational agents by adapting the TES framework [32]. However, with respect to established scale development approaches cf. Boateng et al. [13], their scale has yet to be validated, as acknowledged by the authors.

Our work explores how users perceive the empathy expressed by technologies such as social robots and conversational agents. We developed and validated the Perceived Empathy of Technology Scale (PETS) to measure how well users perceive the expressed empathy of systems. To do so, we followed a structured approach based on Boateng et al. [13]'s guidelines for developing and evaluating

Table 1: The 10-item, two factor Perceived Empathy of Technology Scale (PETS). To be used with randomized 101-point sliders ranging from *strongly disagree* to *strongly agree*.

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.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642035

empirical scales. After consolidating the literature, we invited 30 experts to generate items. With these items, we started a multi-step process involving 324 participants to reduce the initial 100 items. After an exploratory factor analysis, we administered the items to a total of 396 additional participants for scale validation. The final confirmatory factor analysis left us with ten final items. We used 22 scenarios with opposing levels of empathy for development and evaluation, ensuring that the scale is universally applicable.

The multi-stage scale development process resulted in PETS, illustrated in Table 1, a 10-item scale for assessing and comparing how empathic users perceive technology. PETS consists of two factors, the first assessing the user's *Emotional Responsiveness* (PETS-ER) and the second assessing the user's *Understanding and Trust* (PETS-UT) of the technology. The scale allows designers and researchers to measure and compare interactive systems' perceived empathy quickly.

2 RELATED WORK

The initial phase of our scale development process involved gaining a broad understanding of the concept of empathy, with a specific focus on empathic systems. We also examined existing approaches to measuring empathy in technology to determine the need for and requirements of a standardized scale.

2.1 Empathy

Empathy has been studied for over a century, yet it lacks a single, agreed-upon definition. For example, Hoffman [44] defines empathy as an emotional state "triggered by another's emotional state or situation, in which one feels what the other feels or would usually be expected to feel in this situation" [44]. In general, research often describes empathy as a multidimensional construct that can be divided into *cognitive* and *emotional* components [3, 6, 27, 30, 44]. Cuff et al. [27] examined 43 definitions of empathy and analyzed aspects like the relationship between empathy and related concepts such as compassion or sympathy [27, 43]. They conclude that empathy has an *affective* component that elicits emotional responses and a *cognitive* component for understanding and perceiving a subject through various concepts such as perspective taking, interpretation of nonverbal cues, or projection. Further, Cuff et al. [27] conclude that an observer's empathic reactions may be similar but not necessarily identical to the subject's emotion, that empathy can occur without direct stimulation, and that it requires awareness of its cause.

The related concept of *emotional contagion* can be described as experiencing emotions based on the emotions of another person [7, 42]. However, Cuff et al. [27] emphasize that, unlike emotional contagion, empathy requires the observer to be aware of their emotional response and its cause. *Empathic accuracy* is defined as the ability to correctly perceive someone's internal state [7, 49]. Other terms often used in the context of empathy research are *empathic concern* and *personal distress* [30]. In his empathy-altruism hypothesis, Batson [7] describes empathic concern as an other-oriented emotional response, that is, feeling *for* someone else and serving as a motivation for altruistic behavior [7, 15]. Personal distress, in turn, refers to the experience of self-oriented feelings such as anxiety or discomfort as a result of observing another person's distress [7, 15, 33, 44, 49]. Hoffman [44] describes five underlying patterns that can cause empathic distress: mimicry, conditioning, direct association, verbally mediated association, and perspective-taking.

2.2 Empathy in HCI

Empathy is typically discussed in the context of human interaction as a motivator for pro-social behavior that can enhance the quality of social relationships [5, 44, 62]. However, there is also growing interest in the role of empathy in human-computer interaction. We distinguish between technology that *mediates* empathy between agents, empathy *toward* a system, and systems that *act empathic*. Given the context of our paper, we will focus primarily on the third category, empathic systems.

Empathy Mediating Systems. Although it often limits nonverbal modalities, computer-mediated communication (CMC) also offers the potential to foster empathy between users by allowing them to share experiences and emotions over distance and in new ways [70]. For example, Hassib et al. [41] developed an application that extends nonverbal communication in CMC by visualizing heart rate in text messaging, which increased contextual understanding and empathy between users. Similarly, Frey et al. [37] describe a device that captures and potentially shares breathing patterns to increase connectedness and empathy. Curran et al. [28] studied narrative text and bio-signal visualization in virtual reality. They found that narrative text positively affected empathic accuracy, while visualization of electrodermal activity had a negative effect.

Empathy Toward Systems. Empathy toward a system can improve the user-system relationship, but it can also have other positive social effects for the user [23]. Chin et al. [23] examined how different response styles affected users' empathy toward a voice-based agent after it was verbally abused. The empathic response style increased feelings of guilt, reduced anger, and improved the perceived capabilities of the system. Lee et al. [53] created a chatbot that expressed vulnerability and sought advice from the user to elicit empathy, increasing users' self-compassion. In addition, social robots such as *Paro* [39, 48], which are designed to evoke empathy and feelings of care, can help users cope with loneliness or the perception of pain [39].

Empathic Systems. We find that most systems that are intended to behave empathically and are perceived as such by the user are in the area of conversational agents (CA) and human-robot interaction (HRI) [66, 67]. Such empathic systems are typically designed to recognize user emotions and respond in two ways: emotional expressions and functional adaptations like supportive behavior.

A simple approach to simulate empathy is to mimic the detected affective state. Hu et al. [45] described a voice-based CA that reflects the user emotion detected in their speech by responding with vocal utterances ("ha-ha", "wow", "um..."). Besides these emotional expressions, the CA also adapted the conversation flow through praising, distraction, and reappraising strategies, increasing the empathic agent's perceived emotional intelligence [45]. Yang et al. [89] also presented a virtual agent that simulates empathy by reflecting the user's emotions verbally and through the avatar's body language. Bickmore and Picard [11] described another embodied

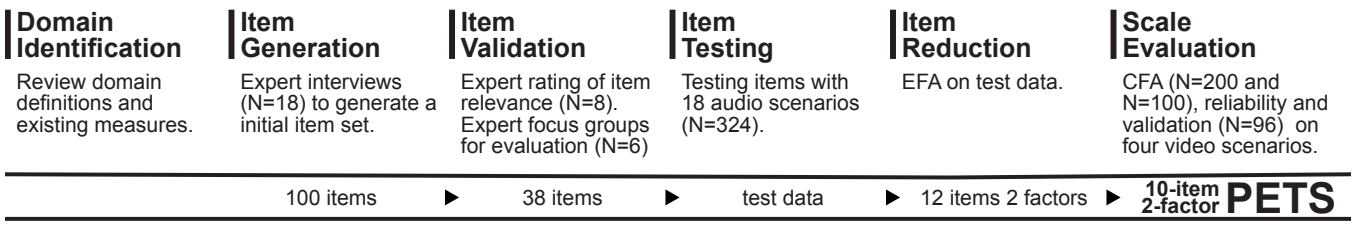

Figure 1: PETS creation based on scale development process by Boateng et al. [13].

virtual agent that expresses emotional responses through nonverbal cues such as gestures, gaze, and proximity. They found that empathic behavior increased respect, liking, and trust toward the system during long-term use.

Burmester et al. [21] introduced the concept of a digital companion that provides an empathic experience in an office work context through support, security, collaboration, and positive feedback. In contrast to most other approaches, their system does not focus on recognizing or expressing emotions. In another context, Xu et al. [87] and Hu et al. [47] presented two chatbots that respond to emotional requests in social media customer service. Again, they did not establish empathic behavior through emotional expressions but through helpfulness, politeness, apologetic behavior, and expressions of understanding [47, 87].

Empathic behavior is particularly important for systems that offer psychological or medical support. Brandtzæg et al. [16] explored how interacting with a chatbot can foster social and emotional support by expressing empathy, trust, and care, for example, by showing interest in the user's emotional state. They also point out that interacting with an artificial system can provide "a feeling of anonymity and freedom" [16] that encourages self-disclosure and trust. As de Gennaro et al. [31] explored, chatbots can also help users cope with social exclusion by showing engagement, engaging in small talk about themselves, and responding empathically when users report negative feelings. Liu and Sundar [56] and Daher et al. [29] studied empathic chatbots that provided personal health advice. They found that empathic behavior was preferred, especially by users who were skeptical of intelligent devices. The chatbots examined expressed sympathy and understanding and showed that they recognized and acknowledged the user's situation to show empathy [29, 56].

Besides conversational agents, a second popular application area for empathic systems is social robots [26, 66, 67]. Buono et al. [20] presented the concept of an empathic care robot that responds to the user's emotions to increase trust and confidence in elderly care. It analyzes the user's speech and facial expressions to determine empathic behavior with the goal of consoling, encouraging, motivating, or calming the user. Their model also includes the determination of an internal affective state for the robot [20]. Ullrich et al. [83] describe an empathic robot that acts as a "companion in misfortune" for children in a doctor's waiting room by reacting to their emotions, providing comfort, and narrating relatable experiences. In addition, the robot uses positive coping strategies and generates attention by addressing the children by name and making eye contact [83]. The social robot introduced by Leite et al. [55] uses similar strategies to act as a companion in a real-world board game scenario. It tries to build a personal relationship and provides motivational and empathic comments and gestures.

The systems described above follow different strategies to generate empathic behavior. A common but not necessary component is the analysis of context and nonverbal cues in speech, text, facial expression, or gestures. Typical reactions include mirroring emotional expressions, providing support and positive feedback, acting engaged and polite and showing interest, sympathy, understanding, and perspective-taking. We used the empathic systems presented in this paragraph to design our test scenarios as described in Section 5.2.

2.3 Measuring Empathy

Several established methods and scales exist for assessing empathy in person-to-person interactions [34, 81]. Some of the most popular measures for assessing empathy are the *Therapist Empathy Scale* (TES) by Decker et al. [32], the *Interpersonal Reactivity Index* (IRI) by Davis [30], and the *Toronto Empathy Questionnaire* (TEQ) [81].

They developed the 9-item TES scale to assess "cognitive, affective, attitudinal, and attunement aspects of therapist empathy" [32], to be used by third-person raters analyzing speech and vocal cues. The IRI provides 28 items and four subscales covering self-assessed perspective-taking, fantasy, empathic concern, and personal distress, with fantasy referring to the tendency to identify with fictional characters [30]. The TEQ consists of 16 items for self-assessment, covering empathic responding, emotion comprehension, sympathetic physiological arousal, altruism, emotional contagion, and the perception and assessment of emotional states in others [81]. There are also older scales, such as Hogan's *The Empathy Scale* from 1969 or the *Questionnaire Measure of Emotional Empathy* (QMEE) from 1972. However, according to Spreng et al. [81], they are no longer recommended for reliable assessment of empathy. Besides these established, generalized scales, there are also methods designed for specific use, such as measuring autism symptoms or nursing empathy in a medical context [81].

Powell and Roberts [70] constructed the 9-item, 3-factor *Measure of State Empathy* (MSE) scale to assess users' cognitive, affective, and compassionate empathy in digital interaction. They also introduced two additional 7-point Likert items that asked participants directly to what degree they experience empathy in CMC and face-to-face interaction. In a more perspective-taking approach, Curran et al. [28] had participants rate another subject's emotions on a continuous valence scale and compared this to the subject's self-rating to calculate empathic accuracy.

Measuring System Empathy. To measure perceived system empathy, most researchers adopted existing scales for assessing human empathy and modified the wording and perspective for the system context. For example, Yalçın and DiPaola [88] modified the *Toronto Empathy Questionnaire* (TEQ) to measure system empathy in a user-avatar interaction from a third-person perspective by watching videos of interaction scenarios. However, Yalçın and DiPaola [88] did not provide details on how they modified the 16-item TEQ. Concannon and Tomalin [25] developed the 10-item *Empathy Scale for Human-Computer Communication* (ESHCC) based on the *Therapist Empathy Scale* (TES). They rephrased the nine original TES items to refer to a system instead of a therapist and to work in text- or voice-based scenarios. While these nine items address empathic concern, expressiveness, acknowledgment, warmth, attunement, understanding, acceptance, and responsiveness, Concannon and Tomalin [25] also extended the scale by an item to assess fallacy avoidance. As TES, the scale was designed to measure empathy from a third-person observer perspective. Charrier et al. [22] developed the *Robot's Perceived Empathy* (RoPE) scale based on an exploratory workshop approach. The scale consists of 18 items covering empathic understanding and empathic response to determine a system's perceived empathy.

Pelau et al. [68] explored the relationship between perceived empathy, anthropomorphism, interaction quality and acceptance and trust of human-AI interaction in a restaurant service context. To measure perceived empathy, they introduced a subscale consisting of 13 items taken from various related studies. Some researchers have also modified questionnaires designed to measure human emotional intelligence rather than empathy. Yang et al. [89] and Ma et al. [57] used 20 items based on a modified MSCEIT questionnaire [60] to create a *Perceived Emotional Intelligence Questionnaire* (PEI), that measures how well an agent is able to perceive, use, understand, and manage emotions. Hu et al. [45] selected four items of this PEI questionnaire [57] to assess the empathic behavior of their voice-based CA. Finally, Burmester et al. [21] suggested using a submodule of the MeCUE questionnaire [63] that captures positive and negative emotions in interaction to evaluate emotionally expressive agents.

As shown above, a growing number of systems are designed to act empathically. Assessing the perceived empathy of these systems is crucial for evaluating their effectiveness and comparing their performance in different contexts and applications. However, there is currently no established approach, especially from a direct user perspective. In the following, we describe the development of a comprehensive and standardized scale, the *Perceived Empathy in Technology Scale* (PETS).

3 ITEM GENERATION

Figure 1 illustrates the process we followed to develop PETS based on the standardized scale development approach of Boateng et al. [13]. We decided to follow a bottom-up approach, as empathy, when perceived from a system, does not necessarily follow the same mental model as when perceived from a human. For instance, an artificial system might not be able - or be believed - to experience emotions on its own and, therefore, might not cover affective components such as empathic concern or personal distress, as defined in established empathy definitions (see Section 2.1). Concannon and Tomalin [25] outline that simulating the experience of such affective states might lead to credibility fallacy. They argue that such distinctive aspects prevent using unadapted human empathy metrics to assess system empathy. We also argue that artificial systems offer advantages by design that could influence the user experience and be considered when modeling the perceived empathy of a system. For example, as suggested by Brandtzæg et al. [16], an unbiased, anonymous system can increase trust and encourage users to open up. Artificial systems also offer advanced cognitive capabilities when analyzing human behavior. In addition to humanly perceptible cues such as gestures, they could, for example, recognize heart rate or, in the case of personal assistants, have insights into a large amount of personal data. Such specific capabilities could lead to a different view of the cognitive dimension of empathy.

In Section 2.3, we described the state-of-the-art approaches for measuring empathy, which are typically limited to certain scenarios or modalities. Thus, we aimed to develop a questionnaire to measure a wide range of systems. Moreover, prior approaches mostly modified an existing empathy scale to adapt it to the specifics of a human-system use case and, as such, often lacked validation. Furthermore, research by Elliott et al. [34] suggests that third-person observer perspectives of empathy tend to be less effective. Therefore, our scale items focus on assessing the perceived empathy of a system from the user's perspective, ideally after interacting with the system. The potential differences and requirements defined above required an unrestricted exploration of system empathy. For this reason, we did not set any modeling requirements in the first interviews, in order not to influence the experts' mental models of empathy in the human-system context.

This bottom-up approach resulted in 18 expert interviews to generate the initial item set. The 100 generated initial items are listed in Table 3.

3.1 Procedure

We conducted the interviews remotely using *Zoom* and *Miro AI*. After welcoming the participants to the video call, we explained the interview procedure and asked them to provide written consent to participate in the interview and to the audio and video recordings. We followed a semi-structured approach with 25 questions and collaborative tasks covering the following topics: empathy in human interaction, empathy in HCI, empathic systems, empathy toward systems, and methods to measure empathic systems. The complete interview guide is available at https://perceived-empathy-of-technology-scale.com and in the Supplementary Material. In addition, the participants had to self-assess their level of expertise in related areas using 5-point scales (see Table 2).

3.2 Participants

We interviewed 18 experts between the ages of 25 and 38 years (M = 30.9, SD = 3.7), of which seven identified as female and eleven as male. Table 2 provides an overview of background and self-assessed expertise in affective systems (M = 3.4, SD = 1.3), emotional, social or behavioral theories (M = 3.8, SD = 0.9), empathy measurement or theories (M = 2.4, SD = 1.2) and application (M = 4.7, SD = 0.5)

Table 2: Overview of the backgrounds and expertise of the 18 experts who participated in the initial item generation. Including scientific degree and domain as well as self-assessed expertise in psychology (Psych.), HCI, affective systems (AS), emotional, social or behavioral theories (ESB), empathy measurement or theories (Emp.), and application (App.) and development (Dev.) of scientific rating scales. Domains: affective computing, human-robot interaction (HRI), human-AI interaction (HAI), user research, cognitive neuroscience, digital health, HCI and psychology.

	Degree	Psych.	HCI	Domain	AS	ESB	Emp.	App.	Dev.	Age	Gender
1	Doctoral	■■□□□	■■■□□	Aff. Comp.	■■■■■	■■■■□	■■■■□	■■■■□	■■■■□	38 y	male
2	Master's	■■□□□	■■■■□	HRI	■■■■□	■■■■■	■■■■□	■■■■■	■■■□□	25 y	female
3	Doctoral	■■■■■	■■□□□	Psych.	■■□□□	■■■■■	■■■□□	■■■■■	■■■■■	32 y	female
4	Master's	■■□□□	■■■■□	HCI	■■■■■	■■■■□	■■■■□	■■■■□	■■■□□	28 y	female
5	Master's	■■■■■	■■■■□	Psych.	■■■■■	■■■■■	■■■■□	■■■■■	■■■■□	28 y	male
6	Doctoral	■■□□□	■■■■■	HAI	■■■■■	■■■■□	■■■■□	■■■■■	■■■■■	32 y	male
7	Master's	■■■□□	■■■■■	Dig. Health	■■□□□	■■■■□	■□□□□	■■■■■	■■■■□	32 y	female
8	Master's	■■□□□	■■■■□	HRI	■■■■□	■■■□□	■■□□□	■■■■□	■■■□□	27 y	male
9	Doctoral	■■■□□	■■■■□	Aff. Comp.	■■■■■	■■■■□	■■□□□	■■■■□	■■■□□	37 y	male
10	Doctoral	■■□□□	■■■■□	User Res.	■■■■□	■■■■□	■■□□□	■■■■■	■■■■■	33 y	female
11	Master's	■■□□□	■■■■□	HCI	■■□□□	■■■□□	■□□□□	■■■■□	■■■■□	31 y	male
12	Master's	■■□□□	■■■■□	HCI	■■■■□	■■■■□	■■■□□	■■■■■	■■■■□	31 y	male
13	Master's	■■□□□	■■■■□	HRI	■□□□□	■■□□□	■□□□□	■■■■■	■■□□□	35 y	male
14	Master's	■■■■□	■■■■□	Cog. Neurosc.	■■■□□	■■■■□	■■□□□	■■■■■	■■□□□	29 y	male
15	Doctoral	■■□□□	■■■■■	HCI	■■□□□	■■□□□	■□□□□	■■■■■	■■■□□	35 y	male
16	Master's	■■■■□	■■■■■	Dig. Health	■■■■□	■■■■□	■■■□□	■■■■■	■■■■■	27 y	female
17	Master's	■■■□□	■■■■■	HAI	■■■□□	■■■■□	■□□□□	■■■■■	■■■■■	30 y	male
18	Master's	■■■□□	■■■■□	HCI	■■□□□	■■■■□	■■□□□	■■■■□	■■□□□	27 y	female

and development ($M = 3.7$, $SD = 1.1$) of scientific rating scales. All participants had expertise in HCI ($M = 4.1$, $SD = 0.8$) or psychology ($M = 2.8$, $SD = 1.1$), 14 were actively engaged in academic research, and four were employed in industry, working in user research, and affective computing. As visualized in Table 2, we covered a broad spectrum and high level of expertise regarding both related theories and applications. The participants rated their expertise as high particularly in the areas of HCI, emotional, social, or behavioral theories, and the application of scientific rating scales. Further, their self-assessment indicated good expertise in the development of scientific scales.

3.3 Qualitative Results

The interview sessions lasted between 32 and 68 minutes ($M = 45.4$, $SD = 8.8$). The 18 interviews were transcribed verbatim. We then followed the Blandford et al. [12] procedure for thematic analysis. First, one researcher open-coded all interviews using an inductive approach with *ATLAS.ti*. Then, over several hour-long sessions, three researchers extracted themes from the interviews for item generation. From an initial set of 1809 identified codes, we finally formulated 100 statements representing the perspective of the interviewed experts on what is essential to quantify empathy. These statements formed the basis of our item set, as listed in Table 3. Furthermore, the qualitative data served as additional input for generating our test scenarios in the following steps.

4 ITEM VALIDATION

As illustrated in Figure 1, item validation was the next step after generating the initial set of items. According to Boateng et al. [13], at this point in development, content validity is required to assess whether the generated items measure the target domain using the Content Validity Index (CVI). We followed a two-step process to filter our initial set of 100 items. As suggested by Boateng et al. [13], we had domain experts ($N = 8$) rate all items for content relevance to ensure representativeness and technical quality. In a second step, we conducted an expert focus group ($N = 6$) to refine the remaining relevant items from a potential PETS end-user perspective.

4.1 Expert Ratings: Procedure

We conducted an online expert rating survey to calculate the CVI. We asked eight experts to rate each of the 100 items on a 4-point scale regarding its appropriateness to the topic.

Procedure. Before item rating, we informed the participants of the purpose of our study, emphasizing that the items were potential candidates for a scale that was intended to measure the perceived empathy of technology from the user's perspective. We instructed them to rate each of the 100 initial items for its relevance to such a questionnaire. We then presented a randomized list of our item texts, with each text followed by the question of how relevant the item is for measuring empathy. Experts had to choose the answer for each item on a 4-point scale: *not relevant - somewhat relevant - quite relevant - very relevant*. We derived this rating procedure from research on content validation by Polit and Beck [69] and Yusoff [91], and decided to go with the suggested 4-point scale to avoid having an ambivalent midpoint [69]. In addition, the experts could add comments to each item in a free text field. Finally, we asked participants to self-assess their level of expertise in related areas via 5-point scales. A study session lasted approximately 30 minutes.

Table 3: Initial item generation resulted in 100 items. Expert ratings removed 55 items due to lack of content validity. The subsequent focus group removed another nine items and added two new modified item variants (*), resulting in a validated set of 38 items. Finally, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) reduced the set to 10 items, resulting in the final PETS with two factors (E1-6, U1-4). All negatively scored items (-) have been removed throughout the process.

	Final PETS items
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.
U1	The system understood my goals.
U2	The system understood my needs.
U3	I trusted the system.
U4	The system understood my intentions.

	Items removed in CFA iteration
11	The system reacted to my emotions.
12	The interaction with the system felt very social.

	Items removed in EFA
13	The system considered my emotions.
14	The system perceived my concerns.
15	I felt an emotional connection with the system.
16	The system understood my thoughts.
17	The system understood my perspective.
18	The system understood my goals, intentions, and needs.
19	The system showed kindness toward me.
20	The system recognized my non-verbal cues.
21	The system did show no interest in me. (-) *
22	The system cared about me.
23	The system was respectful toward me.
24	The system shared my feelings.
25	The system understood what I was communicating.
26	The system reacted to my behavior.
27	The system helped me better cope with difficult moments.
28	The system considered my past experiences.
29	The system considered my physical state.
30	The system was egocentric. (-)
31	The system helped me to open up.
32	The system considered my personal preferences.
33	Over time, my emotional connection with the system increased.
34	The system seemed intelligent.
35	The system cared for my well-being.
36	The system made me feel comfortable.
37	The system acted selfish. (-)
38	The system was purely functional. (-)

	Items removed through validation by focus group
39	The system reacted to my stress level.
40	The system spent time to engage with me.
41	The system would be a great companion in my everyday life.
42	The system was tolerant.
43	The system made me feel calm.
44	The system actively engaged.
45	The system was polite.
46	The system acted responsibly.
47	The system provided a positive user experience.

	Items removed through validation by expert ratings
48	The system reacted dynamically.
49	The system showed empathy in a different form than a human would.
50	The system make me feel better.
51	The system considered the context.
52	The system reacted at the right time.
53	The system was judgemental. (-)
54	The system appeared to be human-like.
55	The system provided assistance.
56	The system was patronizing. (-)
57	The system seemed to have a personality.
58	The system mitigated my sadness.
59	The system reacted to my explicit input.
60	The system gave advice and solutions.
61	The system was personalized to me.
62	The system acted unexpectable. (-)
63	My level of empathy influenced my perception of the system.
64	The system reacted appropriately.
65	The system mimicked me.
66	The system increased my wellbeing.
67	The system was reliable.
68	The system reduced my frustration.
69	The system tried to build me up.
70	The system tried to manipulate me. (-)
71	The system considered my workload.
72	The system actively started conversations.
73	The system was interruptive. (-)
74	The system was unobtrusive. (-)
75	I cared for the system.
76	The system offered me hedonic benefits.
77	The system provided a good user experience.
78	The system considered personal information about me.
79	The system acted natural.
80	The system was easy to understand.
81	Communicating with the system helped me improve my social skills.
82	The system was straightforward.
83	The system used non-verbal cues to express itself.
84	The system helped me combat loneliness.
85	The system does not require empathy from the user.
86	The system acted correctly.
87	The system was annoying.
88	The system was not biased.
89	The system acted on its own.
90	The system acted rule-based. (-)
91	The system acted unnatural. (-)
92	My level of tech affinity influenced my perception of the system.
93	I thanked the system.
94	I apologized to the system.
95	The system acted repetitive. (-)
96	The system gave advice.
97	The system helped me to think about myself.
98	The system was competent.
99	The system portrayed specific gender stereotypes.
100	The system had some sort of embodiment.
101	The system gave solutions.
102	The system increased my efficiency in completing tasks.

Participants. We invited eight experts with a psychology and/or HCI research background who had not participated in the previous item generation. The anonymous survey included information about the study and data processing, to which all participants consented. Participation was voluntary and could be terminated at any time. The average age of the participants was 34.6 years ($SD = 4.9$), with three participants identifying as female and five as male. Five participants had a doctoral degree, and three had a master's degree. Table 4 provides an overview of the expert group. It shows self-assessed expertise in affective systems ($M = 3.0$, $SD = 0.9$), emotional, social or behavioral theories ($M = 4.0$, $SD = 1.3$), empathy measurement or theories ($M = 3.1$, $SD = 1.3$) and application ($M = 4.6$, $SD = 0.5$) and development ($M = 3.6$, $SD = 1.2$)

of scientific rating scales. All participants had expertise in HCI ($M = 4.5$, $SD = 0.8$) or psychology ($M = 3.5$, $SD = 1.2$)

4.2 Expert Ratings: Results

As one of the possible methods suggested by Boateng et al. [13], we measured proportional agreement by calculating the content validity index (CVI) per item [69, 86, 91]. Polit and Beck [69] suggest using a minimum CVI of < 0.78 with ratings from six to ten experts for each item to achieve excellent content validity and to remove, revise, or improve items with values below this threshold. Therefore, we removed all 55 items with a CVI of < 0.6 without revision. The remaining set consisted of 13 items with a $CVI >= 0.88$, 20 items with a CVI of 0.75, and 12 items with a CVI of 0.63. For the remaining items, we calculated an overall scale content validity index (S-CVI) of 0.76 using the average approach suggested by Polit and Beck [69]. Therefore, we decided to take all remaining items, including those with a CVI of 0.63, to the next step for revision or removal. For the lower-scoring items of this set, we decided that they should be given special consideration due to their low scores and were more likely to be removed. In total, we selected 45 of the original 100 items to be reviewed by the focus group in the next step.

4.3 Focus Group: Item Evaluation

In our second step, two authors conducted open, in-depth, face-to-face focus group discussions with four additional researchers from psychology and HCI. Thus, six participants contributed to the open discussions. The average age of the participants was 30.1 years ($SD = 3.9$), with one participant identifying as female and five as male. After briefing the participants on the general topic and the specific phase of our process, we asked everyone to go through the remaining 45 items independently and write suggestions and comments for improvement or deletion. We then discussed each item with the entire group and voted on each suggestion and comment, considering CVI values, wording, relevance, and understandability. This process resulted in some minor rewording and the removal of nine items, six of which were from the low CVI subset. We also added two new items that resulted from slightly modifying two existing items. For example, as item 34 ("The system seemed intelligent") proved to be too generic for the empathy context, we added the rephrased item E2 ("The system seemed emotionally intelligent"). Similarly, in item 27, the originally ambiguous term "certain moments" was changed to "difficult moments." Other items have been reworded or removed to increase applicability to different systems and modalities. For example, we have changed item 25 from "The system understood what I was saying" to "The system understood what I was communicating" in order not to limit user input to voice input and removed item 39 due to its focus on the user's stress level. Additionally, we changed items slightly to emphasize the user as the target of the interaction. For example, we changed item E5 from "The system showed interest" to "The system showed interest in me." In the last part of the meeting, we presented existing approaches to empathic systems from the literature and had a brainstorming session on potential test scenarios for the item testing phase. In total, both steps of the item validation led to a reduction from 100 to 38 potential scale items. At this stage, we also decided to keep negatively worded items in the set to examine their performance in the exploratory factor analysis, as they were explicit results of the expert feedback in the different phases. In general, we aimed for a positively worded scale, as the inclusion of reversed items might have required further changes to other items to obtain a balanced construct, thus also contradicting our intention of a bottom-up approach [14]. Table 3 lists all items in their final wording and shows the resulting sets after each development step.

5 ITEM TESTING

The next step in our scale development process was to test the validated item set and collect data for the subsequent item reduction. For this step, Boateng et al. [13] recommend a sample size of 200 to 300 participants to ensure the availability of sufficient data. We followed this recommendation and conducted a study with 324 participants who rated 18 scenarios with empathic and non-empathic systems using the validated 38 items listed in Table 3.

5.1 Procedure

We conducted an online survey with 338 initial participants recruited through *Prolific*. Participation in our study was voluntary and could be terminated at any time. After providing information about the study and data processing, we asked participants for their consent and demographic information. Next, we assigned each participant a scenario and asked them to read the scenario description and/or listen to the audio playback. In the absence of real-world system implementations, we chose an imaginative approach to simulate the scenario experience from a first-person perspective. To ensure understanding and quality of responses, we required the subject to write a three-sentence summary of the scenario. We then asked participants to rate the system in the scenario perception using our 38-item set and to express their agreement using 101-point sliders displayed in random order. The sliders offered an internal range from 0 to 100 and were labeled *strongly disagree* on the left and *strongly agree* on the right, supporting online participation [38] and more statistical testing [72] while reducing visual bias [59]. As a final step, we asked participants to complete the 9-item *Affinity for Technology Interaction* scale (ATI) [36] to gain insight into general attitudes toward technology. On average, participants took 10.0 minutes ($SD = 4.2$) to complete a procedure and received 1.5£ in compensation.

5.2 Scenario Design

We designed our 18 scenarios (see Table 5) based on applications from prior research on empathic systems (see Section 2.2) and input from the interviews and focus group (see Section 4.3). We designed nine basic scenarios and created two versions of each: one describing interaction with an empathic system, and one describing interaction with a non-empathic, task-oriented application.

To cover a broad range of technology and, therefore, a broad range of use for our scale, we varied the level of system embodiment, interaction modalities, and context in the scenarios. We created scenarios featuring robots and conversational agents like personal assistants and chatbots, kiosks, and smartphone applications. The systems interacted using speech, text, audio and graphical cues, gestures, facial expressions, and other nonverbal cues. Based on the definitions of empathy and the descriptions of empathic systems

Table 4: Overview of the background and expertise of the eight domain experts who rated item relevance. Including scientific degree and domain as well as self-assessed expertise in psychology (Psych.), HCI, affective systems (AS), emotional, social or behavioral theories (ESB), empathy measurement or theories (Emp.) and application (App.) and development (Dev.) of scientific rating scales. Domains: affective computing, user research, HCI and psychology.

	Degree	Psych.	HCI	Domain	AS	ESB	Emp.	App.	Dev.	Age	Gender
1	Doctoral	■■■■■	■■■■■	Psych.	■■■■□	■■■■■	■■■■■	■■■■■	■■■■■	30 y	male
2	Master	■■■■□	■■■■■	HCI	■■■□□	■■■■■	■■□□□	■■■■□	■■■□□	29 y	male
3	Master	■■■■■	■■■■■	HCI	■■□□□	■■■■■	■■■□□	■■■■■	■■■■□	28 y	female
4	Doctoral	■■□□□	■■■■■	HCI	■■□□□	■■■■□	■■■■□	■■■■□	■■□□□	35 y	female
5	Doctoral	■■□□□	■■■■□	HCI	■■□□□	■□□□□	■□□□□	■■■■□	■■□□□	40 y	male
6	Master	■■■□□	■■■□□	User Res.	■■■□□	■■■■□	■■■■□	■■■■■	■■■■■	38 y	female
7	Doctoral	■■■■□	■■■■■	Aff. Comp.	■■■■□	■■■■□	■■■□□	■■■■■	■■■■□	38 y	male
8	Doctoral	■■■□□	■■■■□	Aff. Comp.	■■■■□	■■■■□	■■■□□	■■■■■	■■■■□	39 y	male

in Section 2, we focused on the following key points to portray empathic behavior.

Cognitive Understanding. Our empathic systems should show a cognitive understanding of the user's situation and emotions to address the cognitive component of empathy as defined in Section 2.1. We designed the empathic systems to reflect and acknowledge the user's emotional expressions, their verbal input, and the scenario context to indicate understanding and perspective-taking. For that, the systems provided nonverbal feedback such as nodding and verbal responses such as "I can understand...," "You seem to be...," or "I know how it feels to..."

Affective Expressions. To address the affective dimensions of empathy, we designed the systems in the empathic scenarios to show emotional reactions. Similar to the applications presented in Section 2.2, our systems reflected the detected user states, for example, through vocal utterances, laughing, blushing, facial expressions, or through verbal responses, e.g., "I'm getting angry too..."

Supportive Behavior. As the third design principle, we wanted the empathic systems to provide active support, appear helpful, and try to improve the user's situation, reflecting a compassionate behavior dimension as defined by Powell and Roberts [70]. To achieve that, our systems acted proactive, offered suggestions, and tried to motivate and console the user based on the context and the detected affective states. In addition, the systems also offered support beyond the actual task in order to promote relationship skills and trust (e.g. "I'm here whenever you need support or someone to talk to.", "Let me know if I can help you, I'm a good listener ...").

We designed the non-empathic systems to be purely task-oriented, without affective expressions and less proactive behavior, but with similar functionality. For example, the game application provided similar logging and recommendation features as the empathic game companion but required more manual user input and did not provide affective or motivational feedback. We created a text script for each scenario and used it to generate audio files using AI-based text-to-speech generation. In the generated audio scenarios, a narrator guides the user through the storyline. We used distinct voices for the parts spoken by the narrator and the empathic agents. For the agent voices, we ensured gender balance by using both male and female voices. Our goal was to immerse users in these scenarios so that they could imagine interacting with the systems. In our study, we presented the scenarios with both audio and text scripts. The audio playback of the scenarios lasted between 42 s and 110 s (M = 74.8, SD = 15.7). Empathic scenarios had a more extended script and playback duration, primarily due to the extended verbal communication of the systems. This should also reflect the situation in real-world applications, e.g., comparing purely functional applications with emotional CAs. We provide an overview of the scenarios in Table 5, including references to existing approaches. The complete scripts and audio for the scenarios are available at https://perceived-empathy-of-technology-scale.com and in the Supplementary Material.

5.3 Participants

After rejecting 14 responses due to failed attention checks, we evaluated data from 324 participants. The mean age was 33.6 years (SD = 10.0), with 165 participants identifying as female, 154 as male, four as non-binary, and one preferring not to say. We recruited participants from 31 countries, with 228 residing in the European Economic Area, followed by 31 in North America, 25 in South America, 25 in Africa, ten in the Asia-Pacific region, and five in the Middle East. All participants were required to be fluent in English. Regarding education, 220 participants had a university degree (BA or MA), 61 had a high school diploma, 20 participants had vocational training, 13 had a doctoral degree, and seven participants had no degree or some other degree. The average ATI score of the participants was 4.2 (SD = 0.8).

5.4 Results

The results from the 324 participants on 38 items showed promising item agreement, as presented in Figure 2. After reversing the ratings of the inverted items, the empathic scenarios generally received higher ratings (M=67.4, SD = 10.1), while the non-empathic scenarios received lower ratings (M=33.9, SD = 17.7).

6 ITEM REDUCTION: EXPLORATORY FACTOR ANALYSIS

We applied a series of steps to the data set of 324 observations with the 38 items identified in the previous step, to determine the optimal number of factors and reduce the number of items. Exploratory factor analysis is used to reveal the latent structure of the items

Table 5: Descriptions of our 18 audio- and text-based scenarios including references to related work that served as inspiration. We created two scenarios (empathic and non-empathic) for each context, based on existing approaches and input from focus group and expert interviews. All audio files and text scripts can be found in the Supplementary Material.

Context	No.	System	Audio	Ref.
Playing a board game against a friend.	1	**empathic** | Small toy-like robot (male voice) with facial expressions and speech capabilities offers encouragement, strategic advice, and empathy.	01:50	[8, 55]
	2	**non-empathic** | Training app on smartphone allows to log moves and view analysis.	01:20	
Emotional text conversation with partner.	3	**empathic** | Voice assistant app (female voice) that recognizes emotional distress, helps to draft messages, and promotes understanding and emotional support.	01:23	[45]
	4	**non-empathic** | Standard messaging app with virtual keyboard that provides auto-completion, slide-to-type functionality and animations.	01:05	
Experiencing anxiety in a doctor's waiting room.	5	**empathic** | Robot (male voice) that acts as "companion in misfortune", by talking to the user about doctor appointments and providing comfort.	01:32	[83]
	6	**non-empathic** | Digital kiosk (female voice), providing health information and data input to prepare doctors appointment.	01:10	
Working at a stressful office job.	7	**empathic** | Smart work assistant (female voice), that prevents interruptions, provides emotional support and assistance in email communication.	01:37	[21, 58]
	8	**non-empathic** | Functional office applications that send intrusive notifications and reminders.	01:09	
Being supported by a care robot.	9	**empathic** | Care robot (male voice), providing proactive assistance in daily life and providing emotional support to cope with loneliness.	01:24	[48]
	10	**non-empathic** | Functional care robot (no voice), reacting to user input to help with daily tasks.	00:52	
Asking for mental health support.	11	**empathic** | CA (female voice) that talks with user about their emotional problems, reacting empathic and providing support.	01:16	[16, 54]
	12	**non-empathic** | CA (male voice) asking multiple choice questions to analyze user's mental state.	01:05	
Asking for medical advice.	13	**empathic** | CA (female voice) analyzes symptoms and expresses care, empathy and long-term engagement.	01:20	[29, 56]
	14	**non-empathic** | CA (male voice) responds to symptoms input very functional, provides list- and selection based output.	01:15	
Streaming a movie.	15	**empathic** | CA (male voice) acting as emotional companion while watching a movie, sharing experiences and providing recommendations.	01:20	[45]
	16	**non-empathic** | Smartphone application that provides streaming recommendations and rating.	01:00	
Navigating home after a stressful day.	17	**empathic** | Intelligent CA (female voice) provides personalized navigation alternatives based on user's emotional states.	01:06	
	18	**non-empathic** | Standard navigation application, providing routes and traffic information.	00:42	

by regressing the observed variables on the latent (unobservable) factors [13]. Our exact procedure is described in the following section.

6.1 Data Preparation

As a first step, we reversed the inverse-worded items and removed one of the original 38 items that we had split into three items in the previous step. Then, using a visual approach based on histograms and density plots, we removed eight items that were too skewed, leaving us with 29 items [35]. Subsequently, we computed item-total correlation, skewness, and excess kurtosis for each of the items and removed those that were below or above the specified thresholds (< 0.5 for item-total correlation, $> |1|$ for skewness, and $> |2|$ for kurtosis) [13, 46]. We computed intercorrelations for the remaining 28 items and removed two items that correlated > 0.8 with several other items [71], leaving 26 items for exploratory factor analysis.

6.2 Exploratory Factor Analysis

We first checked the necessary prerequisites for the exploratory factor analysis (EFA). Bartlett's sphericity test was significant ($\chi^2(325) = 8070.95$, $p < 0.001$), indicating that the remaining 26 items are sufficiently correlated and thus suitable for factor analysis [50]. In addition, the Kaiser-Meyer-Olkin test yielded an overall $MSA = 0.97$, an excellent value that confirms sampling adequacy [52], and all individual items had a $MSA > 0.95$, well above the acceptable limit of 0.5, further confirming the appropriateness of conducting factor analysis. We performed an initial analysis to compute eigenvalues for each data component, to determine the number of factors to extract. Two components met Kaiser's criterion of eigenvalues greater than or equal to one [90]. Both the scree plot and the parallel analysis performed suggested two underlying factors. Based on this, we performed EFA using the *psych* package in R [73] with Promax Rotation and Principal Axis Factoring, as we expected the two factors to be correlated based on our previous theoretical considerations and analysis of the qualitative data, and oblique rotation is preferred for covarying factors [61].

The resulting loadings are shown in Table 6. The PETS item order shown in Table 1 and Table 3 is based on the ascending order of these loadings but is not relevant for the application of the scale. Based on this initial factor analysis, we removed loadings < 0.4 and cross-loadings [13], resulting in a further reduction of items to

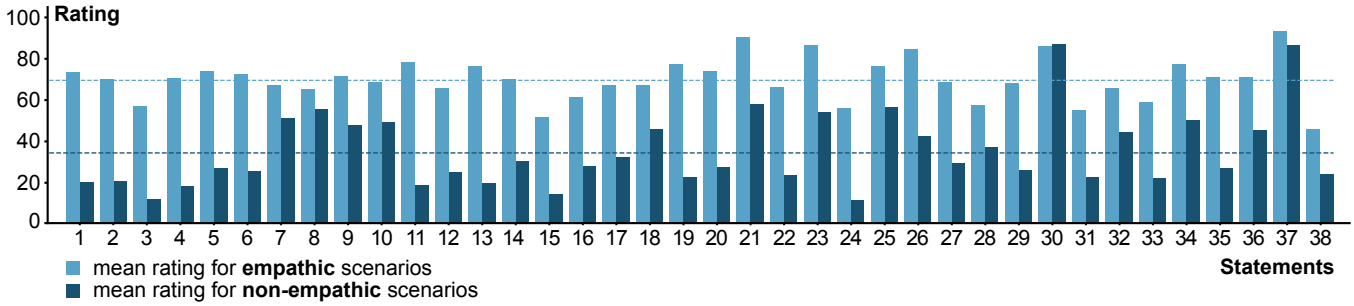

Figure 2: Mean ratings from 324 participants for each item, across empathic (M=67.4, *SD* = 10.1) and non-empathic scenarios (M=33.9, *SD* = 17.7). Values of inverse-worded items were reversed. See Table 3 for item wordings.

23. In addition, to keep the scale efficient, we kept only items that loaded >= 0.75 on the respective factor, reducing items to twelve, with eight items in Factor 1 and four items in Factor 2. The overall Cronbach's Alpha was $\alpha = 0.93$ for the entire scale, which is an excellent value and suggests that the shortened 12-item version is internally consistent [65]. The corresponding EFA model had acceptable values for the Tucker-Lewis Index ($TLI = 0.92$) [9] and the Root Mean Square Error of Approximation ($RMSEA = 0.077$) [19], so the next step was to validate the model with confirmatory factor analysis.

7 SCALE EVALUATION

To verify the factor structure found in the last step and to ensure that the measured construct of our scale is distinct from related constructs, we conducted a confirmatory factor analysis (CFA), re-examined the internal consistency, and checked test-retest reliability, discriminant validity and convergent validity [13]. To assess the generalizability of the scale, we conducted these tests with four newly designed scenarios and three new samples. For the test of dimensionality, we collected two samples ($N = 200$ and $N = 100$) as we refined the final set of items from twelve to ten due to inadequate fit in the first CFA iteration. For testing reliability, we recruited participants again who took part in the second CFA run. For construct validation, we collected another distinct sample ($N = 96$).

7.1 Scenario Design

We created four scenarios (two empathic, two non-empathic) and decided to present them not only with audio and text but with audio and visual animation, similar to related approaches [26, 28]. Our goal was to increase the range of scenario variation compared to the previous item test study and improve the system representation's depth and consistency by using a visual representation. According to participant ratings and text feedback in our previous study, the systems in the board game and work contexts were perceived as highly empathic and non-empathic, respectively. Therefore, we decided to create four scenarios: (a) an empathic robot companion named Bud-E, (b) a purely functional game training application, (c) an empathic voice assistant for office work, and (d) a purely functional office work application.

We deliberately chose systems with different forms of embodiment. In the board game scenario, the robot interacted via speech (male voice), gestures, gaze, and blushing. The empathic assistant interacted via speech (female voice) and controlling desktop elements in the work scenario. Its speech interaction was also visualized as a wave animation. The game training application featured only touch input and visual display. The office application displayed standard desktop notifications. While the systems in the opposing

Table 6: Factor loadings from EFA. For readability, only loadings greater than 0.3 are shown. As described in Section 6.2, only items that loaded >= 0.75 on the respective factor were retained for CFA. PA refers to the Principal Axis. The item labels refer to the labels in Table 3.

	Items explored in EFA	PA 1	PA 2
11	The system reacted to my emotions.	**1.056**	
E1	The system considered my mental state.	**0.939**	
E2	The system seemed emotionally intelligent.	**0.894**	
E3	The system expressed emotions.	**0.884**	
E4	The system sympathized with me.	**0.874**	
E5	The system showed interest in me.	**0.789**	
E6	The system supported me in coping with an emotional situation.	**0.779**	
12	The interaction with the system felt very social.	**0.774**	
29	The system considered my physical state.	0.743	
20	The system recognized my non-verbal cues.	0.691	
35	The system cared for my well-being.	0.700	
33	Over time, my emotional connection with the system increased.	0.699	
31	The system helped me to open up.	0.670	
26	The system reacted to my behavior.	0.653	
27	The system helped me better cope with difficult moments.	0.632	
14	The system perceived my concerns.	0.627	
22	The system cared about me.	0.679	
16	The system understood my thoughts.	0.460	0.435
17	The system understood my perspective.	0.425	0.457
34	The system seemed intelligent.	0.313	0.499
U1	The system understood my goals.		**0.953**
U2	The system understood my needs.		**0.813**
U3	I trusted the system.		**0.806**
U4	The system understood my intentions.		**0.779**
36	The system made me feel comfortable.		0.587
32	The system considered my personal preferences		0.492

(a) Bud-E, the empathic robot companion, provided support during a board game (2:00 min).

(b) A touch-based smartphone application allowed to log and view game information (1:29 min).

(c) An empathic voice assistant supported the user at their office job (2:10 min).

(d) The office application notified the user about incoming messages and events via notifications and popups (1:41 min).

Figure 3: We designed an animation sequence for each of the four evaluation scenarios showing user interaction with empathic (a, c) and non-empathic (b, d) systems. Pictures above show one exemplary frame from each scenario.

scenarios differed in appearance and empathic behavior, the basic storyline was the same. Figure 3 shows a sample frame from each scenario animation sequence, including the total duration. The depiction of individuals in the video is intentionally gender-neutral and blurred to allow participants to identify and focus on the system interaction. General considerations for our scenario design are described in Section 5.2. All videos and scripts are available at https://perceived-empathy-of-technology-scale.com and in included in the Supplementary Material.

7.2 Test of Dimensionality

As suggested by Boateng et al. [13] we used CFA to test our factor structure using a new sample of participants, collected in two iterations ($N = 200$ and $N = 100$).

Procedure. We followed a similar procedure as in Section 5.1 and first introduced the research details and obtained consent and demographic data. Next, each participant was randomly assigned to view one of four scenarios, presented using audio and visual information rather than audio and text. Participants were required to view the scenario at least once before continuing. We then asked participants to describe their scenario experience in a free text field. As in the item test study, we then asked them to rate their perception of the system with the presented items, indicating their level of agreement by moving a slider (0 to 100, *strongly disagree* to *strongly agree*). This step included 12 items in the first run ($N = 200$) and the revised 10-item set in the second run ($N = 100$). We also included an attention control slider and randomized all sliders to minimize order effects. Subsequently, we asked participants to complete the ATI [36] scale.

Participants. We conducted two independent validation runs. All participants were recruited via *Prolific*, reported being fluent in English, and had not participated in the previous studies. For the first batch, we recruited 234 participants and rejected 34 due to failed attention checks. On average, participants took 08:52 min (SD=4:08) to complete a session and received a compensation of 1.4£. The mean age was 34.9 years ($SD = 10.5$), with 51.0 % of participants identifying as male, 47.0 % as female, and four as non-binary. Most participants (141) held a university degree (BA/MA), 43 held a high school diploma, seven had vocational training, six held a Ph.D., and three held no degree or some other degree. Participants resided in 26 different countries, with the majority (169) in the European Economic Area, 13 in North America, seven in South America, four

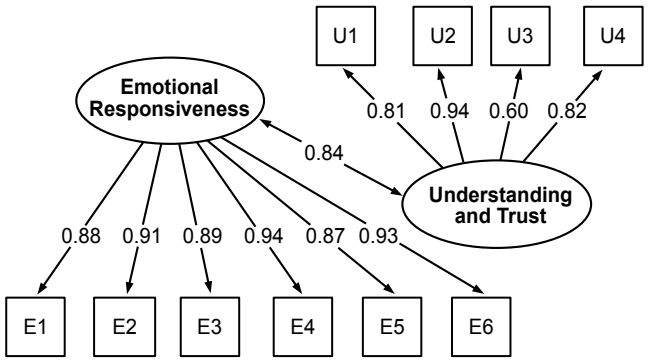

Figure 4: The final 10-item PETS factor model resulting from confirmatory factor analysis.

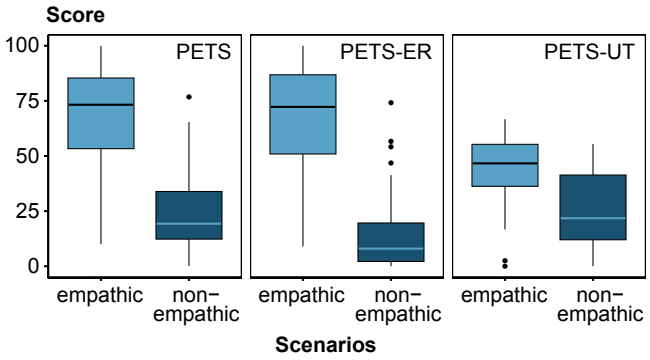

Figure 5: Comparison of empathic and non-empathic scenarios for overall PETS and the separate factors PETS-ER and PETS-UT

in the Asia-Pacific region, four in South Africa, and three in the Middle East. The average ATI score of the participants was 4.0 ($SD = 0.9$).

For the second batch, we recruited 101 participants and had to reject one, leaving 100 valid sessions. On average, participants took 08:12 min (SD = 4:01) to complete a session and received a compensation of 1.4£. The mean age of the second batch was 35.8 years ($SD = 11.4$), with 52 participants identifying as male and 48 as female. Most participants (69) held a university degree (BA/MA), 17 had a high school diploma, eight had vocational training, three had a Ph.D., and three had no or some other degree. Participants resided in 13 different countries, most (86) in the European Economic Area, 12 in North America, one in South America, and one in the Middle East. The average ATI score of the participants was 3.9 ($SD = 1.0$).

Confirmatory Factor Analysis. To compute the CFA, we used the *lavaan* package in R [76]. We allowed the two factors of eight and four items to covary in the model, as we expected the factors to be related both theoretically and based on the oblique Promax rotation in the EFA [61]. The resulting model had high loadings (all > 0.65, all but one item > 0.8) for the items on their corresponding factors. Except for $RMSEA = 0.084$, which is above the acceptable threshold of 0.08, the model had good fit indices ($TLI = 0.968, CFI = 0.974$, and $SRMR = 0.033$) [13, 46]. Since the value for $RMSEA$ suggested that the model is not optimally fit, we looked at the modification indices to determine which items might be the reason for the non-optimal fit (see Brown [18]). By excluding two items in Factor 1, we were able to significantly lower the $RMSEA$ and also improve the remaining fit indices ($RMSEA = 0.052, TLI = 0.989, CFI = 0.992$, and $SRMR = 0.024$) [13, 19, 46].

To validate this new ten-question scale configuration, we repeated the previous step with a second, independent *Prolific* sample ($N = 100$) to rate the four scenarios again. The resulting model of the CFA had optimal fit indices ($RMSEA = 0.034, TLI = 0.994, CFI = 0.996$, and $SRMR = 0.032$), indicating an adequate factor structure and thus supporting the 2-factor, 10-item solution [13, 19, 46]. The final items are shown in Table 3 and the corresponding model in Figure 4. Based on their assigned items, we named the first factor *Emotional Responsiveness* (PETS-ER) and the second factor *Understanding and Trust* (PETS-UT). Section 8.2 provides a more detailed interpretation of these two dimensions.

Internal Consistency. The internal consistency of the PETS, as measured by Cronbach's alpha, is excellent, with a value of $\alpha = 0.96$ for the entire scale and values of $\alpha = 0.96$ for PETS-ER and $\alpha = 0.87$ for PETS-UT [65]. Finally, we calculated the split-half reliability using the *psych* package in R [73] and a sample of 10000, where the items are split in half 10000 times, and the score for half of the items is correlated with the score for the other half. Here, we obtained values of $\beta = 0.90$, representing the lowest split-half reliability of all iterations, indicating excellent consistency of the PETS [74]. To examine how well PETS discriminates between empathic and non-empathic scenarios [24], we additionally conducted group comparisons between the two subscales and the overall scale using the final CFA data set via t-tests. Both PETS-ER ($t(98) = 13.23, p < .001, d = 2.65$), PETS-UT ($t(98) = 5.60, p < .001, d = 1.12$), and the total scale of the PETS ($t(98) = 10.62, p < .001, d = 2.12$) were rated significantly higher in empathic scenarios than in non-empathic scenarios Figure 5.

7.3 Test–Retest Reliability

To examine test-retest reliability, we had 51 subjects from the second sample rate the four scenarios again 12 weeks after collection for the final CFA. The study procedure was identical to the procedure described in Section 7.2 to ensure consistency when collecting data from the same participants at two time points.

Participants. Of the original 100 participants in the final CFA sample, we were able to recruit 51 participants again. The average completion time was 11:20 min (SD = 3:56) for which participants received a compensation of 1.8£. The mean age was 39.4 years ($SD = 12.2$), with 56.9 % of participants identifying as male and 43.1 % as female. 34 participants held a university degree (BA/MA), eleven held a high school diploma, five had vocational training and one held a Ph.D. Participants resided in twelve different countries, with the majority (42) in the European Economic Area, eight in North America and one in South America. The average ATI score of the participants was 4.1 ($SD = 1.0$).

Results. We calculated intraclass correlations (*ICC*) and Pearson correlations between the data from the two collection points to determine the consistency of the sum scores over time. Both *ICC* (absolute agreement: $ICC = 0.943$, 95% $CI = 0.901$ to 0.968, $p < 0.001$) and Pearson correlation ($r(48) = .89, p < .001$) had high values for the total scale, indicating high test-retest reliability. The two subscales also achieved satisfactory values for PETS-ER (absolute agreement: $ICC = 0.957$, 95% $CI = 0.925$ to 0.975, $p < 0.001$ and $r(48) = .92, p < .001$) and PETS-UT (absolute agreement: $ICC = 0.804$, 95% $CI = 0.658$ to 0.888, $p < 0.001$ and $r(48) = .67, p < .001$) [77, 84].

7.4 Construct Validity

Various of the applications we described in Section 2.2 show how empathic behavior might improve the acceptance and trust between two agents [11, 16, 20, 67, 68]. Further, research suggests that anthropomorphism increases the perceived empathy and that perceived empathy and anthropomorphism both affect the interaction quality with AI systems [68]. To test construct validity, we therefore selected measures that assess the user perception of a system regarding the concepts of trust, ease of use, anthropomorphism, and understanding.

Measures. We expected that empathic systems score high on trust in a system since empathic systems should act understandable to the user and respond to user needs. However, this does not mean that non-empathic yet reliable systems must score low on user-system trust. Therefore, we used the *Trust of Automated Systems Test* (TOAST) [85] scale to assess trust for discriminant validation.

Furthermore, empathic behavior through nonverbal cues might be associated with humanized interaction and thus also affect the usability or interaction quality of a system [68]. To ensure that PETS does not only measure perceived ease of use, we conducted a second measurement of discriminant validity using the *System Usability Scale* (SUS) questionnaire [2, 17].

To assess the effect of anthropomorphism on perceived empathy measured with PETS, we used the respective 5-item scale of the *Godspeed* questionnaire series [4] for discriminant validity. While anthropomorphic appearance might influence the perceived empathy of a system, we expected to measure high empathy ratings also with the less embodied systems in our scenarios [68]. As further described in Section 8.2, multiple PETS items relate to the system's ability to understand affective user states as well as the user's needs, goals, and intentions. The *Networked Minds Measure of Social Presence* (NMSP) [40] provides subscales for perceived affective as well as message understanding (PAU and PMU) and perceived emotional interdependence (PEI) between user and agent. As these concepts might interact with PETS-ER and PETS-UT, we used these subscales to test for convergent validity.

Procedure. The study procedure followed the same steps as in Section 7.2, where participants had to watch one of our four scenarios and subsequently rate their perception of the system. In addition to PETS and ATI [36], we asked participants to respond to the items of the scales outlined above: TOAST [85], SUS [17], *Godspeed* anthropomorphism [4] and the NMSP subscales [40].

Participants. For the validation run, we recruited 100 participants on *Prolific* that had not taken part in the previous studies. We had to reject four participants due to failed attention checks, leaving $N = 96$ participants for construct validation. On average, it took the participants 14:10 min (SD = 5:58) to complete the study, for which they received a compensation of 1.5£. The mean age was 32.7 years ($SD = 8.4$), with 51.0 % of participants identifying as male, 46.9 % as female and two participants identifying as non-binary. Most of the participants (72) held a university degree (BA/MA), 19 held a high school diploma, three had a doctoral degree, one had vocational training, and one participant had no or some other degree. The participants resided in twelve different countries, with the majority (84) in the European Economic Area, four in North America, three in South America, three in the Asia-Pacific region, and two in the Middle East. The average ATI score of the participants was 4.2 ($SD = 0.9$).

Discriminant Validity. To test the independence of our construct from the related scales, we applied the method proposed by Rönkkö and Cho [75] with the *semTools* package and the *discriminantValidity* function in R [51].

Both the PETS-ER subscale ($upperCI = 0.433$) and the PETS-UT subscale ($upperCI = 0.492$) had correlations with the SUS below the critical threshold of 0.75, suggesting that the constructs measured by PETS and SUS are independent of each other [75].

Similarly, the PETS-ER subscale was sufficiently weakly correlated with the TOAST *Understanding* subscale ($upperCI = 0.068$) and the TOAST *Performance* subscale ($upperCI = 0.459$). The PETS-UT subscale was also sufficiently low correlated with both TOAST subscales (*Understanding*: $upperCI = 0.220$, *Performance*: $upperCI = 0.673$). Thus, the constructs measured by PETS and TOAST are also sufficiently independent.

Regarding the *Godspeed* anthropomorphism scale, correlation with PETS-ER ($upperCI = 0.858$) and PETS-UT ($upperCI = 0.848$) indicated a marginal problem regarding discriminant validity [75]. Therefore, we conducted CFA to further examine the overlap between the *Godspeed* and PETS factors [1]. Table 7 shows the resulting comparably small correlation, indicating that they are mostly independent.

Convergent Validity. To test the NMSP subscales [40] PAU, PMU, and PEI for convergent validity, we again conducted CFA, with the results shown in Table 7. To account for the sample size, we calculated individual models for TOAST, SUS, and the individual subscales of the NMSP scale. As expected, all subscales show correlations with PETS-ER and PETS-UT. PAU and PEI correlate more strongly with PETS-ER than with PETS-UT, presumably reflecting the affective dimension of these factors. PMU nearly equally correlates with PETS-ER and PETS-UT. Based on the correlations, we conclude that the related constructs of understanding and emotional interdependence, as measured by the NMSP scale, are convergent with our PETS factors.

8 DISCUSSION

Finally, we provide guidelines on administering the PETS and calculating the total and factor scores. Moreover, we discuss our insights, limitations, and future work.

Table 7: Correlations of PETS subscales (ER and UT) and related scales (TOAST Understanding and Performance, SUS, Godspeed, PAU, PMU, PEI) as calculated for discriminant and convergent validity.

Scale	PETS	Corr.	χ^2	*df*	TLI	RMSEA
TOAST Under.	ER	-0.089	333.075	146	0.855	0.116
TOAST Under.	UT	-0.003	333.075	146	0.855	0.116
TOAST Perf.	ER	0.155	333.075	146	0.855	0.166
TOAST Perf.	UT	0.301	333.075	146	0.855	0.166
SUS	ER	0.102	420.432	167	0.821	0.126
SUS	UT	0.127	420.432	167	0.821	0.126
Godspeed	ER	0.347	163.071	87	0.932	0.095
Godspeed	UT	0.330	163.071	87	0.932	0.095
PAU	ER	0.526	263.891	101	0.875	0.130
PAU	UT	0.427	263.891	101	0.875	0.130
PMU	ER	0.491	281.591	101	0.844	0.136
PMU	UT	0.500	281.591	101	0.844	0.136
PEI	ER	0.508	237.417	101	0.892	0.119
PEI	UT	0.389	237.417	101	0.892	0.119

8.1 Guidelines for Implementation, Scoring and Analysis

As in development, the PETS items should be presented in a randomized order during administration of the scale to avoid order effects [78]. Items on the PETS are rated on a 0-100 scale from *strongly disagree* to *strongly agree*. Since the PETS consists of two subscales and the items load similarly on the respective factors, we propose to calculate a mean per subscale and to calculate the total score of the PETS weighted by the number of items per subscale so that values from 0 to 100 can be obtained on the subscale scores as well as on the total score. Scores are calculated as follows:

$$PETS = PETS\text{–}ER * 0.6 + PETS\text{–}UT * 0.4$$
$$PETS\text{–}ER = (E_1 + E_2 + E_3 + E_4 + E_5 + E_6)/6$$
$$PETS\text{–}UT = (U_1 + U_2 + U_3 + U_4)/4$$

This scoring method ensures that each item contributes proportionally to the total score and that each subscale is appropriately weighted to reflect its number of items. In this way, we maintain the integrity of each subscale while providing a comprehensive, aggregated measure of empathy as expressed by the PETS that allows intuitive interpretation of both the total score and the subscale scores and facilitates straightforward comparability across use cases and domains. We also suggest referring to the PETS scores of our four validation scenarios as presented in Table 8.

8.2 PETS Dimensions

As described in Section 3, we followed a bottom-up approach to develop our scale items, as human empathy models might not cover the specific features of empathy expressed through a system. Although the 2-factor structure of PETS results from our bottom-up approach and the EFA, it reflects concepts from the established two-dimensional view of affective and cognitive empathy (see Section 2.1). The first factor, *Emotional Responsiveness*

Table 8: Total and factor PETS scores of our four validation scenarios (N=100).

Scenario	PETS		PETS-ER		PETS-UT	
	M	*SD*	*M*	*SD*	*M*	*SD*
(a) empathic game companion	75.5	19.4	76.2	20.5	49.7	12.8
(b) game training app	33.0	15.8	17.7	16.5	37.3	13.8
(c) empathic work companion	60.6	22.8	61.4	24.0	39.7	16.2
(d) work application	16.3	17.2	11.2	17.8	15.9	12.4

(PETS-ER), mainly relates to the system's ability to recognize, process, and respond to the user's affective states. The items E1 and E2 relate to understanding emotional states and can, therefore, be associated with cognitive empathy, as described in Cuff et al. [27]. However, the dimensions of empathy are not always defined in the same way in related studies. For example, Concannon and Tomalin [25] assign the item *understanding of feelings/inner experience* in their scale to affective empathy. Item E3 assesses the system's emotional response, which is why we assign it to the affective empathy dimension. Item E4 refers to the system's expression of sympathy toward the user. It relates to empathic concern (see Batson [7]) and, therefore, to affective empathy or, according to Powell and Roberts [70], to a dedicated dimension of compassionate empathy. Item E5 asks whether the system shows interest in the user, referring to engagement and techniques such as active listening. It involves cognitive understanding as well as affective components. In contrast to the previous items, item E6 assesses the result of empathic behavior by asking whether the system has helped the user to cope with an emotional situation. Following Powell and Roberts [70], we associate this item with compassionate empathy in which one desires to "help the other person deal with his situation and his emotions" [70]. In human interaction, another component of affective empathy would be the actual experience of emotion triggered, for example, through emotion contagion, empathic concern, or empathic distress [7, 27]. PETS-ER does not evaluate this component as we focus on the user perception perspective, further described in Section 8.3.

The correlations of PETS-ER with affective understanding and perceived emotional interdependence as measured in Section 7.4 and Table 7 further highlight the emotional dimension of this factor.

The items in the second factor *Understanding and Trust* (PETS-UT) mainly represent cognitive empathy. However, items U1, U2, and U4 relate to understanding the user's goals, intentions, and needs. They are not necessarily connected to an affective dimension and, therefore, could also be associated with cognitive-only perspective-taking regarding the theory of mind [27]. We suggest that the higher correlation of PETS-UT with perceived message understanding compared to perceived affective understanding, as shown in Table 7, further supports that distinction. Compared to the other items, item U3 occupies some kind of special position in the factor, as it assesses trust in the system from the user's perspective. Given our bottom-up approach and the results from the EFA, we assume that this item correlates with the other PETS-UT items, as a system that appears to understand and know the

user well might result in increased perceived competence and trustworthiness.

8.3 Perspective

In the context of psychotherapy, Elliott et al. [34] divided measurement approaches for human empathy into four categories based on the subject's perspective: *observer rating*, *client rating*, *therapist rating*, and *empathic accuracy*. They suggest that a first-person perspective (client rating) performs better than other perspectives. On the other hand, human self-perception may differ from self-reported empathy [25], and in a system context, self-report would mean that the system reports its own level of empathy. Concannon and Tomalin [25] suggest that this may be of little analytical value since AI can be trained to respond positively to appropriate questions. We argue that this depends on the system's design and that, in any case, the user's perception is more important. Furthermore, we believe a third-person observer perspective would not adequately capture this perception. Therefore, PETS measures empathy as perceived by users interacting with a system, and thus, the adequacy with which the machine models/simulates empathic behavior based on its capabilities. In the future, however, there may be cases where a system needs to evaluate the perceived empathy of another system. We plan to apply and evaluate PETS in such use cases.

8.4 Perceived System Empathy

One challenge in our development process has been to create test scenarios that elicit empathic system interaction. First, the overall context is important. In our case, the scenarios were intended to provide emotional experiences to which the system could respond. We chose to cover a variety of emotions to ensure generality. In addition, based on participants' scenario summaries, we suggest using at least audio and preferably video to promote immersion and allow subjects to imagine the situation being described fully. Second, it is important to consider participants' general attitudes toward technology, AI, and systems that can detect emotional states. Some participants expressed concerns about privacy and surveillance by the empathic systems described. When implementing the PETS, it may be relevant to include scales such as the ATI to control for these influences. Another influencing factor is the appearance of the system and its interaction modalities. Although we have described different levels of system embodiment to ensure generalizability, anthropomorphic system design could positively influence the perceived empathy of a system, which we did not control for in the development of the PETS. Finally, another factor that may influence the perceived empathy of systems is the temporal dimension. Perceptions of empathy may change over time, e.g., a system may be perceived as highly empathic initially, but as users interact with it over time, weaknesses become apparent, or vice versa. This could change scores when evaluated at a different time, providing an opportunity to adapt and revise systems.

8.5 Scale Limitations

While the PETS is a valuable tool for assessing system empathy, it has some limitations that should be acknowledged. First, our sample was drawn from *Prolific*, which may not be truly representative of the broader population due to self-selection bias. This sampling strategy could potentially limit the generalizability of our findings, and future studies should examine whether the scale is valid for other samples [10]. Second, only positively coded items were included in the final scale. This approach may introduce a positive response bias that could lead to generally higher scores for the PETS compared to a version that would include negatively coded items [14]. Furthermore, the scale we developed conceptualizes and measures empathy as a strictly positive construct. However, it is worth noting that high perspective-taking, a key component of empathy, could theoretically be used with malicious intent. The PETS-UT subscale assumes a positive correlation between perspective-taking and pro-social behavior, which may not always be accurate. The potential for misuse of perspective-taking skills is not accounted for in our current study, which may limit the scope of our understanding of empathy in intelligent systems. It is unclear to what extent human conditions such as psychopathy, associated with average perspective-taking ability but low levels of emotional compassion [64], can be translated to intelligent systems. The need for this aspect in PETS needs to be discussed.

We developed and validated PETS with imaginative scenarios to allow us full control over the system design. While various researchers followed a similar approach (see [45, 56, 88]), we are aware that the lack of interactivity is a limitation compared to a first-person experience with a real system. Therefore, we intend to apply PETS to existing systems in interactive constellations to validate it further.

8.6 Future Validation

In addition to construct validity, Boateng et al. [13] also describe the investigation of criterion validity by evaluating predictive and concurrent validity as part of a scale evaluation process. As with Section 7.4, finding applicable scales for our empathic system context poses a challenge, as most related scales either focus on specific contexts or are not designed to assess the perception of a system from a user perspective.

Concurrent Validity. To the best of our knowledge, there are no validated "gold standard" measures that could be used to assess concurrent validity. Davis' IRI [30, 82] is often used to validate scales for human empathy, however, like most of the other scales presented in Section 2.3, such as the TES or the TEQ, it is designed as a self-report scale and is therefore not suitable for evaluating the perceived empathy of artificial systems from the user's perspective. Also, established scales that measure the closely related concept of emotional intelligence, such as the SSEIT [79] or the MSCEIT [60], are either self-report or task-based and, therefore, not applicable. In Section 2.3, we introduced approaches to assess system empathy, which are mostly poorly validated and often restricted to a specific context or an observer perspective. Therefore, they are also not suitable to test for concurrent validity [21, 22, 25, 45, 68, 88, 89]. So, in alignment with Boateng et al. [13], we decided to omit concurrent validation at this point due to a lack of matching measures.

Predictive Validity. Regarding predictive validity, we intend to carry out further evaluations in the future application of PETS. We argue that effects on user behavior in this context are highly dependent on the empathic system and the context of use. Park

and Whang [67] discuss empathy in the context of human-robot interaction. They describe the improvement of social relationships, long-term interaction, liking, and trust as potential purposes of empathic behavior in social robots. We suggest to assess these variables for predictive validity in corresponding scenarios. Based on the findings described in Section 2.2, we further conclude that interaction with empathic systems might affect technological acceptance or the readiness to share information with a system and, therefore, suggest assessing the predictability of these variables [16]. For systems that offer psychological support, such as the chatbot designed to help users cope with social exclusion described by de Gennaro et al. [31], we suggest observing the development of corresponding social behaviors before and after an interaction.

9 CONCLUSION

In this paper, we introduced the *Perceived Empathy of Technology Scale* (PETS), a novel scale designed to measure the empathy of systems toward the user. PETS is a 10-item scale composed of two factors, *Emotional Responsiveness* (PETS-ER) and *Understanding and Trust* (PETS-UT), all items are listed in Table 1. We developed the scale based on expert interviews, focus groups, and a series of user tests. To ensure the broad applicability of our scale, we conducted testing and validation with 22 distinct scenarios. With the PETS, we aim to establish a standardized method for rapid testing and evaluating empathic systems and thus support the advancement of emotionally intelligent technology in a wide range of domains.

REFERENCES

[1] James C. Anderson and David W. Gerbing. 1988. Structural equation modeling in practice: A review and recommended two-step approach. *Psychol. Bull.* 103, 3 (May 1988), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

[2] Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594. https://doi.org/10.1080/10447310802205776

[3] Simon Baron-Cohen and Sally Wheelwright. 2004. The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *J. Autism Dev. Disord.* 34, 2 (April 2004), 163–175. https://doi.org/10.1023/B:JADD.0000022607.19833.00

[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (Jan. 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[5] C. Daniel Batson. 1987. Prosocial Motivation: Is it ever Truly Altruistic? In *Advances in Experimental Social Psychology*, Leonard Berkowitz (Ed.). Vol. 20. Elsevier, Academic Press, Cambridge, MA, USA, 65–122. https://doi.org/10.1016/S0065-2601(08)60412-8

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

[7] C. Daniel Batson. 2010. Empathy-induced altruistic motivation. In *Prosocial motives, emotions, and behavior: The better angels of our nature , (pp,* Mario Mikulincer (Ed.). Vol. 468. American Psychological Association, xiv, Washington, DC, USA, 15–34. https://doi.org/10.1037/12061-001

[8] Howard Becker. 1931. Some forms of sympathy: a phenomenological analysis. *J. Abnorm. Soc. Psychol.* 26, 1 (April 1931), 58–68. https://doi.org/10.1037/h0072609

[9] Peter M. Bentler and Douglas G. Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin* 88, 3 (1980), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

[10] Jelke Bethlehem. 2010. Selection bias in web surveys. *International statistical review* 78, 2 (2010), 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x

[11] 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. https://doi.org/10.1145/1067860.1067867

[12] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI Research: Going Behind the Scenes.* Springer Cham, Cham, Switzerland. 51–60 pages. https://doi.org/10.2200/s00706ed1v01y201602hci034

[13] Godfred O. Boateng, Torsten B. Neilands, Edward A. Frongillo, Hugo R. Melgar-Quiñonez, and Sera L. Young. 2018. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Front Public Health* 6 (June 2018), 149. https://doi.org/10.3389/fpubh.2018.00149

[14] Kathrin Bogner and Uta Landrock. 2016. Response biases in standardised surveys. *GESIS survey guidelines* 2 (2016), 0–0. https://doi.org/10.15465/gesis-sg_en_016

[15] Jeremy Boy, Anshul Vikram Pandey, John Emerson, Margaret Satterthwaite, Oded Nov, and Enrico Bertini. 2017. Showing People Behind Data: Does Anthropomorphizing Visualizations Elicit More Empathy for Human Rights Data?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5462–5474. https://doi.org/10.1145/3025453.3025512

[16] 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. https://doi.org/10.1145/3411764.3445318

[17] John Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. *Usability evaluation in industry* 189, 3 (1996), 189–194. https://doi.org/10.1201/9781498710411-35

[18] Timothy A. Brown. 2015. *Confirmatory factor analysis for applied research.* Guilford publications, New York, NY, USA.

[19] Michael W. Browne and Robert Cudeck. 1992. Alternative ways of assessing model fit. *Sociological methods & research* 21, 2 (1992), 230–258. https://doi.org/10.1177/0049124192021002005

[20] Paolo Buono, Giovanna Castellano, Berardina De Carolis, and Nicola Macchiarulo. 2020. Social Assistive Robots in Elderly Care: Exploring the role of Empathy. In *Empathy@ Avi.* ceur-ws.org, Aachen, Germany, 12–19.

[21] Michael Burmester, Katharina Zeiner, Katharina Schippert, and Axel Platz. 2019. Creating Positive Experiences with Digital Companions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312821

[22] Laurianne Charrier, Alisa Rieger, Alexandre Galdeano, Amélie Cordier, Mathieu Lefort, and Salima Hassas. 2019. The rope scale: a measure of how empathic a robot is perceived. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, New York, NY, USA, 656–657. https://doi.org/10.1109/hri.2019.8673082

[23] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376461

[24] Gilbert A. Churchill Jr. 1979. A paradigm for developing better measures of marketing constructs. *Journal of marketing research* 16, 1 (1979), 64–73. https://doi.org/10.1177/002224377901600110

[25] Shauna Concannon and Marcus Tomalin. 2023. Measuring perceived empathy in dialogue systems. *AI & SOCIETY* Online (July 2023), 1–15. https://doi.org/10.1007/s00146-023-01715-z

[26] Henriette Cramer, Jorrit Goddijn, Bob Wielinga, and Vanessa Evers. 2010. Effects of (in)accurate empathy and situational valence on attitudes towards robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, New York, NY, USA, 141–142. https://doi.org/10.1109/HRI.2010.5453224

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

[28] Max T. Curran, Jeremy Raboff Gordon, Lily Lin, Priyashri Kamlesh Sridhar, and John Chuang. 2019. Understanding Digitally-Mediated Empathy: An Exploration of Visual, Narrative, and Biosensory Informational Cues. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300844

[29] 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. https://doi.org/10.1145/3383652.3423864

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

[31] 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. https://doi.org/10.3389/fpsyg.2019.03061

[32] Suzanne E. Decker, Charla Nich, Kathleen M. Carroll, and Steve Martino. 2014. Development of the Therapist Empathy Scale. *Behav. Cogn. Psychother.* 42, 3 (May 2014), 339–354. https://doi.org/10.1017/S1352465813000039

[33] N. Eisenberg and P. A. Miller. 1987. The relation of empathy to prosocial and related behaviors. *Psychol. Bull.* 101, 1 (Jan. 1987), 91–119. https://doi.org/10.1037/0033-2909.101.1.91

[34] Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and Leslie S. Greenberg. 2011. Empathy. *Psychotherapy* 48, 1 (March 2011), 43–49. https://doi.org/10.1037/a0022187

[35] David B. Flora, Cathy LaBrish, and R. Philip Chalmers. 2012. Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in psychology* 3 (2012), 55. https://doi.org/10.3389/fpsyg.2012.00055

[36] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (April 2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150

[37] Jérémy Frey, May Grabli, Ronit Slyper, and Jessica R. Cauchard. 2018. Breeze: Sharing Biofeedback through Wearable Technologies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174219

[38] Frederik Funke and Ulf-Dietrich Reips. 2012. Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales. *Field Methods* 24, 3 (2012), 310–327. https://doi.org/10.1177/1525822x12444061

[39] Nirit Geva, Florina Uzefovsky, and Shelly Levy-Tzedek. 2020. Touching the social robot PARO reduces pain perception and salivary oxytocin levels. *Sci. Rep.* 10, 1 (June 2020), 9814. https://doi.org/10.1038/s41598-020-66982-y

[40] Chad Harms and Frank Biocca. 2004. Internal Consistency and Reliability of the Networked Minds Measure of Social Presence. In *Seventh annual international workshop: Presence*. Vol. 2004. International Society for Presence Research, Delft, Netherlands, 246–251.

[41] Mariam Hassib, Daniel Buschek, Paweł W. Wozniak, and Florian Alt. 2017. HeartChat: Heart Rate Augmented Mobile Chat to Support Empathy and Awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2239–2251. https://doi.org/10.1145/3025453.3025758

[42] Elaine Hatfield, Richard L. Rapson, and Yen-Chi L. Le. 2011. Emotional contagion and empathy. In *The social neuroscience of empathy*. MIT Press, Cambridge, MA, USA, 19–30. https://doi.org/10.7551/mitpress/9780262012973.003.0003

[43] Grit Hein and Tania Singer. 2008. I feel how you feel but not always: the empathic brain and its modulation. *Curr. Opin. Neurobiol.* 18, 2 (April 2008), 153–158. https://doi.org/10.1016/j.conb.2008.07.012

[44] Martin L. Hoffman. 2008. Empathy and prosocial behavior. *Handbook of emotions* 3 (2008), 440–455.

[45] Jiaxiong Hu, Yun Huang, Xiaozhu Hu, and Yingqing Xu. 2021. Enhancing the Perceived Emotional Intelligence of Conversational Agents through Acoustic Cues. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 282, 7 pages. https://doi.org/10.1145/3411763.3451660

[46] Li-tze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal* 6, 1 (1999), 1–55. https://doi.org/10.1080/10705519909540118

[47] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch Your Heart: A Tone-Aware Chatbot for Customer Care on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173989

[48] Lillian Hung, Cindy Liu, Evan Woldum, Andy Au-Yeung, Annette Berndt, Christine Wallsworth, Neil Horne, Mario Gregorio, Jim Mann, and Habib Chaudhury. 2019. The benefits of and barriers to using a social robot PARO in care settings: a scoping review. *BMC Geriatr.* 19, 1 (Aug. 2019), 232. https://doi.org/10.1186/s12877-019-1244-6

[49] William Ickes. 1993. Empathic accuracy. *Journal of personality* 61, 4 (1993), 587–610.

[50] Robert I. Jennrich. 1970. An asymptotic $\chi 2$ test for the equality of two correlation matrices. *J. Amer. Statist. Assoc.* 65, 330 (1970), 904–912. https://doi.org/10.1080/01621459.1970.10481133

[51] Terrence D. Jorgensen, Sunthud Pornprasertmanit, Alexander M. Schoemann, and Yves Rosseel. 2022. `semTools`: *Useful tools for structural equation modeling*. GPL-3. https://CRAN.R-project.org/package=semTools R package version 0.5-6.

[52] Henry F. Kaiser. 1974. An Index of Factorial Simplicity. *Psychometrika* 39, 1 (1974), 31–36. https://doi.org/10.1007/BF02291575

[53] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: A Chatbot for Self-Compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300932

[54] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-Disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376175

[55] Iolanda Leite, André Pereira, Samuel Mascarenhas, Ginevra Castellano, Carlos Martinho, Rui Prada, and Ana Paiva. 2010. Closing the Loop: From Affect Recognition to Empathic Interaction. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments* (Firenze, Italy) *(AFFINE '10)*. Association for Computing Machinery, New York, NY, USA, 43–48. https://doi.org/10.1145/1877826.1877839

[56] Bingjie Liu and S. Shyam Sundar. 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10 (Oct. 2018), 625–636. https://doi.org/10.1089/cyber.2018.0110

[57] Xiaojuan Ma, Emily Yang, and Pascale Fung. 2019. Exploring Perceived Emotional Intelligence of Personality-Driven Virtual Agents in Handling User Challenges. In *The World Wide Web Conference* (San Francisco, CA, USA) *(Www '19)*. Association for Computing Machinery, New York, NY, USA, 1222–1233. https://doi.org/10.1145/3308558.3313400

[58] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, Paul Johns, Akane Sano, and Yuliya Lutchyn. 2016. Email Duration, Batching and Self-Interruption: Patterns of Email Use on Productivity and Stress. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1717–1728. https://doi.org/10.1145/2858036.2858262

[59] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5421–5432. https://doi.org/10.1145/2858036.2858063

[60] John D. Mayer, Peter Salovey, David R. Caruso, and Gill Sitarenios. 2003. Measuring emotional intelligence with the MSCEIT V2.0. *Emotion* 3, 1 (March 2003), 97–105. https://doi.org/10.1037/1528-3542.3.1.97

[61] D. Betsy McCoach, Robert K. Gable, and John P. Madura. 2013. *Instrument Development in the Affective Domain: School and Corporate Applications*. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-7135-6

[62] Nicole McDonald and Daniel Messinger. 2011. The Development of Empathy: How, When, and Why. In *Moral behavior and free will: A neurobiological and philosophical approach*. IF Press, Rome, IT, 333–359.

[63] Michael Minge, Manfred Thüring, Ingmar Wagner, and Carina V Kuhr. 2017. The meCUE Questionnaire: A Modular Tool for Measuring User Experience. In *Advances in Ergonomics Modeling, Usability & Special Populations*. Springer International Publishing, Basel, CH, 115–128. https://doi.org/10.1007/978-3-319-41685-4_11

[64] Jana L. Mullins-Nelson, Randall T. Salekin, and Anne-Marie R. Leistico. 2006. Psychopathy, empathy, and perspective-taking ability in a community sample: Implications for the successful psychopathy concept. *International Journal of Forensic Mental Health* 5, 2 (2006), 133–149. https://doi.org/10.1080/14999013.2006.10471238

[65] Jum C. Nunnally and Ira H. Bernstein. 1994. *Psychometric Theory*. McGraw, New York.

[66] 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, Article 11 (Sept. 2017), 40 pages. https://doi.org/10.1145/2912150

[67] 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. https://doi.org/10.3390/ijerph19031889

[68] Corina Pelau, Dan-Cristian Dabija, and Irina Ene. 2021. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior* 122 (2021), 106855. https://doi.org/10.1016/j.chb.2021.106855

[69] Denise F. Polit and Cheryl Tatano Beck. 2006. The content validity index: are you sure you know what's being reported? Critique and recommendations. *Res. Nurs. Health* 29, 5 (Oct. 2006), 489–497. https://doi.org/10.1002/nur.20147

[70] Philip A. Powell and Jennifer Roberts. 2017. Situational determinants of cognitive, affective, and compassionate empathy in naturalistic digital interactions. *Comput. Human Behav.* 68 (March 2017), 137–148. https://doi.org/10.1016/j.chb.2016.11.024

[71] Janice Rattray and Martyn C. Jones. 2007. Essential Elements of Questionnaire Design and Development. *Journal of Clinical Nursing* 16, 2 (Feb. 2007), 234–243. https://doi.org/10.1111/j.1365-2702.2006.01573.x

[72] Ulf-Dietrich Reips and Frederik Funke. 2008. Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods* 40, 3 (01 Aug. 2008), 699–704. https://doi.org/10.3758/brm.40.3.699

[73] William Revelle. 2022. *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psych R package version 2.2.9.

[74] William Revelle and David M. Condon. 2019. Reliability from α to ω: A tutorial. *Psychological assessment* 31, 12 (2019), 1395. https://doi.org/10.1037/pas0000754

[75] Mikko Rönkkö and Eunseong Cho. 2022. An updated guideline for assessing discriminant validity. *Organizational Research Methods* 25, 1 (2022), 6–14. https://doi.org/10.1177/1094428120968614

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

[77] Valentin Rousson, Theo Gasser, and Burkhardt Seifert. 2002. Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Statistics in medicine* 21, 22 (2002), 3431–3446. https://doi.org/10.1002/sim.1253

[78] Murat Doğan Şahin. 2021. Effect of item order on certain psychometric properties: A demonstration on a cyberloafing scale. *Frontiers in Psychology* 12 (2021), 590545. https://doi.org/10.3389/fpsyg.2021.590545

[79] Nicola S. Schutte, John M. Malouff, Lena E. Hall, Donald J. Haggerty, Joan T. Cooper, Charles J. Golden, and Liane Dornheim. 1998. Development and validation of a measure of emotional intelligence. *Pers. Individ. Dif.* 25, 2 (Aug. 1998), 167–177.

[80] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My Chatbot Companion - a Study of Human-Chatbot Relationships. *Int. J. Hum. Comput. Stud.* 149 (May 2021), 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

[81] R. Nathan Spreng, Margaret C. McKinnon, Raymond A. Mar, and Brian Levine. 2009. The Toronto Empathy Questionnaire: scale development and initial validation of a factor-analytic solution to multiple empathy measures. *J. Pers. Assess.* 91, 1 (Jan. 2009), 62–71. https://doi.org/10.1080/00223890802484381

[82] Morgan D. Stosic, Amber A. Fultz, Jill A. Brown, and Frank J. Bernieri. 2022. What is your empathy scale not measuring? The convergent, discriminant, and predictive validity of five empathy scales. *J. Soc. Psychol.* 162, 1 (Jan. 2022), 7–25. https://doi.org/10.1080/00224545.2021.1985417

[83] Daniel Ullrich, Sarah Diefenbach, and Andreas Butz. 2016. Murphy Miserable Robot: A Companion to Support Children's Well-Being in Emotionally Difficult Situations. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 3234–3240. https://doi.org/10.1145/2851581.2892409

[84] Joseph P. Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research* 19, 1 (2005), 231–240.

[85] Heather M. Wojton, Daniel Porter, Stephanie T. Lane, Chad Bieber, and Poornima Madhavan. 2020. Initial validation of the trust of automated systems test (TOAST). *The Journal of social psychology* 160, 6 (2020), 735–750. https://doi.org/10.1080/00224545.2020.1749020

[86] Christine A. Wynd, Bruce Schmidt, and Michelle Atkins Schaefer. 2003. Two quantitative approaches for estimating content validity. *West. J. Nurs. Res.* 25, 5 (Aug. 2003), 508–518. https://doi.org/10.1177/0193945903252998

[87] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3506–3510. https://doi.org/10.1145/3025453.3025496

[88] Ozge Nilay Yalçın and Steve DiPaola. 2019. Evaluating levels of emotional contagion with an embodied conversational agent. In *Proceedings of the 41st annual conference of the cognitive science society*. The Cognitive Science Society, Seattle, WA, USA, 3143–313.

[89] Yang Yang, Xiaojuan Ma, and Pascale Fung. 2017. Perceived Emotional Intelligence in Virtual Agents. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2255–2262. https://doi.org/10.1145/3027063.3053163

[90] Keith A. Yeomans and Paul A. Golder. 1982. The Guttman-Kaiser criterion as a predictor of the number of common factors. *The Statistician* 31 (1982), 221–229. https://doi.org/10.2307/2987988

[91] Muhamad Saiful Bahri Yusoff. 2019. ABC of content validation and content validity index calculation. *Education in Medicine Journal* 11, 2 (2019), 49–54. https://doi.org/10.21315/eimj2019.11.2.6