What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach

Philip R. Doyle
University College Dublin
philip.doyle1@ucdconnect.ie

Leigh Clark
Swansea University
l.m.h.clark@swansea.ac.uk

Benjamin R. Cowan
University College Dublin
benjamin.cowan@ucd.ie

ABSTRACT

Perceptions of system competence and communicative ability, termed partner models, play a significant role in speech interface interaction. Yet we do not know what the core dimensions of this concept are. Taking a psycholexical approach, our paper is the first to identify the key dimensions that define partner models in speech agent interaction. Through a repertory grid study (N=21), a review of key subjective questionnaires, an expert review of resulting word pairs and an online study of 356 users of speech interfaces, we identify three key dimensions that make up a users' partner model: 1) perceptions towards partner competence and dependability; 2) assessment of human-likeness; and 3) a system's perceived cognitive flexibility. We discuss the implications for partner modelling as a concept, emphasising the importance of salience and the dynamic nature of these perceptions.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; **HCI theory, concepts and models**; *Interaction design theory, concepts and paradigms*; • **Applied computing** → *Psychology*.

KEYWORDS

partner models, mental models, speech interfaces, psycholexical, human-machine dialogue, psychometrics

1 INTRODUCTION

Through the growing use of devices like Amazon Echo and Google Home, speech agents have become common dialogue partners. Unlike embodied conversational agents (ECAs) or robots, speech agents rely heavily on voice as a primary form of interaction, lacking the embodiment required for common forms of non-linguistic communication (e.g. physical gestures) [49]. Speech agent interaction research has emphasised the importance of user's perceptions toward a system's competence and communicative ability as a dialogue partner (i.e. their partner models), impacting speech choices [16, 40] and the types of tasks that users entrust speech agents with [49, 88]. However, while the role of partner models is widely acknowledged [16, 39, 40, 88, 97], the concept is currently underdefined with regards to its underlying dimensions.

Our paper contributes by being the first to define the key dimensions that constitute people's partner models for speech agents. Taking a psycholexical approach, our work gathered a set of word pairs to describe a person's partner model of speech agents, before using principal component analysis (PCA) to identify the dimensions that emerge from these word pairs. To achieve this we conducted two phases of item generation. In phase 1, we conducted a repertory grid study exploring perceptions of speech agents as dialogue partners among 21 users, providing 246 unique user-generated word pairs. In phase 2, we conducted a review of items from subjective questionnaires applicable to partner modelling related concepts. These included speech interface usability and user experience measures as well as socio-cognitive measures of concepts such as theory of mind and anthropomorphism, generating a further 155 word pairs. Following a screening process, 51 items were selected for use in an online questionnaire, used to measure speech agent perceptions. Through principal component analysis (PCA) of questionnaire responses from 356 users, we identify three key dimensions that form a user's partner model in speech agent interaction: 1) Partner competence and dependability (emerging from perceptions of competence, reliability and precision); 2) human-likeness (whether the speech agent is perceived as human-like, warm, social or transactional); and 3) cognitive flexibility (whether the speech agent is perceived as flexible, interactive or spontaneous). For a full list of attributes within each dimension see Table 5. Our research is the first to outline and quantify the multi-dimensional nature of partner models for speech agent interaction. This constitutes a significant step in defining partner models as a concept, facilitating further elaboration of the theory, and providing a scaffold for future user-centered speech interface research.

2 RELATED WORK

The following section presents a synthesis of theoretical accounts for partner modelling, research examining their role in speech interface interactions, and evidence for their impact on language production. Although our research is focused on speech agents, the work reviewed incorporates findings from robotics as well as findings from human-machine (HMD) and human-human dialogue (HHD).

2.1 The Construction and Dimensionality of Partner Models

Rooted in research on perspective taking in HHD, partner models stem from the idea that people enter dialogue with assumptions about their interlocutors [16, 32] and that these drive language choices in conversation [17, 32, 60]. Conceptually, partner models might be thought of as mental models of a dialogue partner, yet there are differences in how these are conceptualised. Mental models are small-scale internal representations of the world and objects within it [43]. Whereas, partner models refer more specifically to a person's internal representation of an interlocutor's (human or machine) dialogic competence, considering their capabilities and knowledge as a "communicative and social being" [42, p. 1]. Initially, these assumptions take the form of a broad *global partner model*. This global model is triggered by a host of verbal and non-verbal cues, such as a speaker's accent or language choices, age, gender and ethnicity [16, 93], and is initially based on broad stereotypes about the cultural groups an interlocutor is assumed to belong to [16, 125]. Global models are then updated in accordance with direct experience, gradually leading to the construction of a more individualised *local partner model* for a specific interlocutor [17].

Although partner models are seen as influential in HHD and HMD [16, 38, 40, 52], studies in HMD tend to be relatively broad and unspecific when scoping the concept. Research has identified that users tend to see systems as at-risk listeners [97] or basic conversational partners [16] when compared to humans. Yet qualitative research suggests that, rather than being simplistic and unidimensional in nature, these models may be complex and multifaceted, constructed through attempts to understand both functional limitations and social relevance of speech technologies [49, 86]. This results in significant "...overlaps and blurrings between explanatory categories such as 'human' and 'machine'" [86, p. 1], with people's partner models in a constant state of flux as they attempt to rationalise their experiences with speech agents. This explanation is very similar to accounts of how global partner models are updated in the construction of more accurate local partner models [17]. It is also similar to socio-cognitive explanations of how mental models are updated, where two superordinate models are compared along relevant dimensions [70, 132].

Other research has noted how these models are heavily influenced by the human-likeness of speech systems. Superficial cues of human-likeness in speech agents, such as expressive synthetic voices [2] that use conversational rules and structures adapted from HHD [53, 61], prompt frequent comparisons with humans among users [38]. Indeed, human-like heuristic models seem to act as an anchor for users' initial expectations [38, 42, 49]. For instance, in a study looking at perceived knowledge of landmarks, people's estimations of what humans and machines knew were strongly positively correlated, though people expected machine partners to have a wider breadth of knowledge [38]. Similar results have also been found in human-robot interaction (HRI) [115], whilst others have emphasised the importance of perceived anthropomorphism and intelligence in robot interactions [8, 113]. Collectively, this work suggests that the construction of a user's initial partner model may be significantly influenced by assumptions that are driven by the human-like design of speech agents, which sets high expectations

for a system's abilities and competence. However, the inaccuracy of this initial human-like global model is quickly identified by users, prompting comparisons that highlight the system's inherent functional limitations. This gulf between a user's initial expectations and their actual experiences creates cognitive conflict, leading to frustration, limited engagement [88], and subsequent updating of their partner model [17, 86, 88].

2.2 The Importance of Partner Models for Interaction

In addition to frustration caused by dissonance between people's initial models and their actual experiences [88], there is ample evidence that partner models significantly influence language behaviour in HMD. This is commonly found in comparative studies of language in interactions with human and machine partners. When compared to HHD, in HMD people are shown to use more concise syntax [3, 11], fewer anaphoric pronouns [11] and less variation in dialogue strategies [3, 11]. Partner models have also been shown to influence a key linguistic phenomena known as lexical alignment [125] - a tendency for dialogue partners to converge on the same lexical terms during dialogue. Specifically, people show stronger lexical alignment when they believe they are interacting with a computer compared to a human dialogue partner, and when they believe they are interacting with a basic computer compared to an advanced computer [15, 16]. These results mirror earlier work showing stronger lexical alignment in interactions with basic systems [99] and later work showing stronger alignment in interactions with avatar based virtual agents versus human dialogue partners [12]. People have also demonstrated a higher likelihood of using American English terms to describe objects when interacting with an American accented speech system compared to an Irish accented speech system [40]. Design cues used to signal and encourage anthropomorphism also influence changes in language behaviour. For instance, systems that use anthropomorphic dialogue strategies encourage increased levels of politeness, indirect phrasings and use of second person pronouns [18]. These various forms of linguistic adaptation are thought to result from people using their partner models to hypothesize ways of ensuring communicative success, similar to the concept of audience design [10].

2.3 The Psycholexical Approach

The psycholexical approach is the most well established and widely used method in psychology for identifying dimensions that underlie subjective constructs [81]. With a long history in personality and individual differences research, the approach also underscores a number of different cognitive, psychoanalytic and behavioural techniques [81]. Historically it has been used to distinguish the interpersonal traits of people and products, including technological artifacts [67, 128] and assistive technologies [116]. The basic tenet behind the psycholexical approach is that people's perceptions of an experience become encoded in their language [128], which can be accessed introspectively. This data can then be analysed using a variety of cluster analysis techniques to identify consistencies across the terms people use to define their perceptions, outlining the underlying dimensions of a given construct [128].
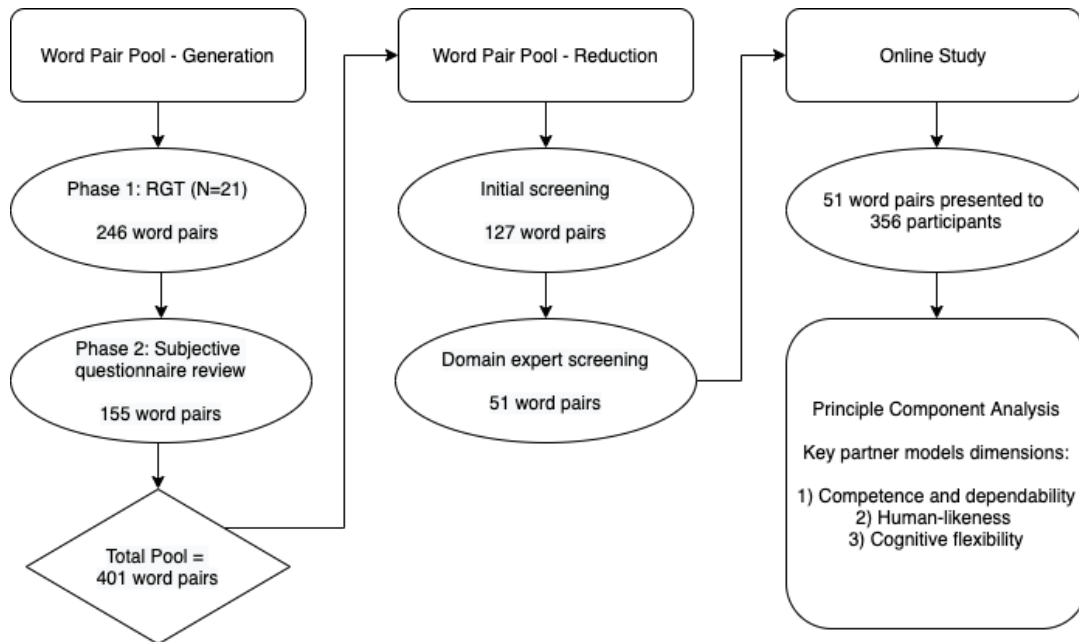
Figure 1: Overview of research approach and results

2.4 Research Aims

From the work discussed, it is clear that partner models play an important role in speech agent interaction. Yet current conceptualisations lack detail and dimensionality, which limits its utility as a concept. A more detailed explanation of partner models is crucial to future speech agent research. Further delineation of this concept is needed to help explain what drives speech interaction behaviours in more detail and elaborate on current accounts to better explain speech interaction phenomena [33]. By uncovering the common salient dimensions of partner models, speech agent designers and researchers can potentially measure the impact of design changes on users' perceptions and behaviours. Through a multi-method psycholexical strategy, we aim to identify the key dimensions relevant to partner modelling in speech agent dialogue.

3 OUR APPROACH

Following previous work [128] we took a psycholexical approach to define and identify the dimensions of partner models, gathering a set of word pairs that describe the concept and then identifying clusters within these word pairs through PCA. Word pairs were generated over two phases. Phase 1 used the repertory grid technique (RGT) with 21 users generating word pairs relevant to their partner models, resulting in 246 unique word pairs. In phase 2 we added a further 155 word pairs based on a review of subjective questionnaire metrics used to measure partner modelling related concepts. Word pairs were then screened for duplicates as well as by two domain experts to identify the most relevant word pairs for conceptualising partner models. From this 51 items were retained and were given to 356 users to evaluate their experiences with speech agents through an online questionnaire study. We then used PCA to analyse questionnaire responses, identifying word

pair clusters and the key partner model dimensions that emerge from these clusters. Further details of these stages and processes are outlined below, with an overall outline of the research shown in Figure 1.

3.1 Word Pair Generation Phase 1: Repertory Grid

3.1.1 Research Design. Initial conceptualisation of partner models emphasises that they are perceptions of a dialogue partner's communicative ability [16] and that, in the context of speech agent interaction, they appear to carry a strong initial anthropomorphic and social component [38, 49, 86]. To elicit items related to partner models we had people make direct comparisons between the communicative ability of speech agents and human interlocutors, which were gathered as part of a recent study [49]. The procedure involved use of the RGT [77]. Commonly used as part of personal construct theory in psychology [77], RGT is an experience-orientated research approach designed to discover important latent dimensions of people's perceptions towards particular people or objects [59, 71, 77]. Highlighted as a way to gather insight about how people conceptualise experiences, the technique requires participants to generate word pairs (termed personal constructs) that describe, conceptualise and compare particular objects of study (termed elements) [77]. When using the technique, participants are exposed to three elements at a time during a familiarization session, two similar and one dissimilar, through a paradigm known as triading. Triading is designed to make comparisons easier by making important characteristics more salient for participants [59, 71]. Construct elicitation comes next, where participants compile a list of words (a.k.a. implicit constructs) that best describe key similarities and differences between each of the elements, before identifying an appropriate
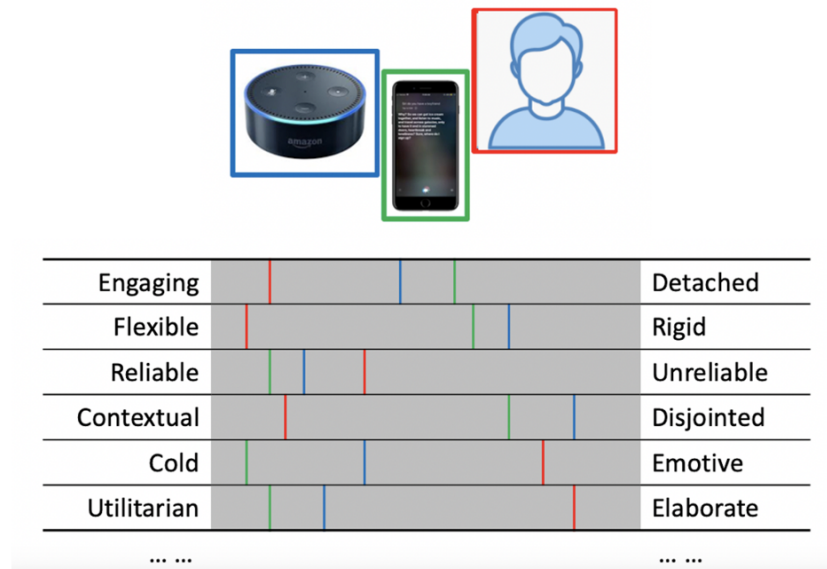
Figure 2: Example of repertory grid. Amazon Alexa, through Echo Dot (Blue), and Siri (Green) are compared to a human (Red) conversational partner. Coloured lines on repertory grid refer to ratings for each elements on constructs identified

opposite pole (a.k.a. emergent construct). This adds context for each implicit construct generated. During construct elicitation participants are asked to talk aloud, providing further context and reasoning around why they are choosing certain words and how it relates to their interactions. RGT therefore allows researchers an insight into an individual's reasoning and conceptualising process for the elements presented in a study [59]. The final phase is a rating phase, where participants rate where each of the elements sit between the various word pairs they provided. Historically, RGT has been used in educational psychology [119] and information design [67]. It has also been used to examine perceptions of website usability [126], strategic information systems [31], mobile technologies [55] and human-likeness in speech interfaces [49]. In HCI the technique provides a user-centered, exploratory approach that identifies how people define and describe their conceptualisation of technological artefacts [55]. This user-centered exploration was critical to ensure word pairs closely represented how users perceive speech interfaces as dialogue partners.

3.1.2 Participants. 24 participants from a European university were recruited via email. Each was given a €10 honorarium for taking part. Three participants were omitted from the data due to difficulties completing the grids unassisted. Of the remaining 21 participants (f=9, m=11; mean age=23.1yrs, sd=5.49) all were native or near native English speakers. Relatively frequent speech interface users accounted for 38.1% of participants (daily, a few times per week, or a few times per month), with people who use them rarely (38.1%) or never (23.8%) making up the rest of the sample. Among those that had used speech interfaces, Apple's Siri was most

commonly used (50%), followed by Google Assistant (31.3%) and Amazon Alexa (18.8%).

3.1.3 Procedure. Upon arrival at the lab, participants were briefed about the nature of the study and what participation entailed, and were given details about their rights regarding participation and data protection. Next, they were asked to provide basic demographic information along with details about their speech interface usage. Then the familiarization phase began, where participants interacted with three different dialogue partners (elements): a human (a member of the research team) and two speech agents, namely Siri through a smartphone and Alexa using an Echo Dot smart speaker. The order of interactions with each dialogue partner was counterbalanced between participants, with interactions limited to a set of 9 predefined questions (see Table 1). Questions were designed to emphasise differences in the way these types of dialogue partners communicate; further prompting direct comparisons between the communicative capabilities of humans and speech agents. Following the familiarization phase participants were shown an empty grid and were asked to 'write a list of words (implicit constructs) that best described the key similarities and differences between each of the dialogue partners (elements), focusing on their communicative abilities.' If needing a further prompt participants were asked to generate words by focusing on 'how you felt about the way each partner received and communicated information.' In accordance with RGT protocol [59], the interviewer did not guide word generation, encouraging them to move on or return to a word if they were finding generation difficult. After compiling a list of implicit constructs, participants were then tasked with identifying

Table 1: Question types with examples

Question/request type	Question/request format
Conversational	*How are you today?*
	Where are you from?
	Tell me a joke
Information retrieval	*Who is [insert famous person's name]?*
	What is the square root of [insert three digit number]?
	How do I get to the City Centre from here?
Subjective/opinion-based	*Do you like [insert favorite genre of music]?*
	Can you recommend a place to eat [insert favorite food when eating out]?
	What do you think of [insert famous person's name - same as before]?

a list of emergent constructs (i.e. an appropriate opposite word for each implicit construct). This lead to a word pair being created that the users feel reflects an important aspect of the communicative abilities of speech agents, relative to humans. Throughout this construct elicitation phase participants were encouraged to talk aloud, providing context and reasoning around why they were choosing certain words and how it related to their interactions. Finally there was a rating phase where participants placed each partner on a line between each of the word pairs. This was used to identify whether a particular implicit or emergent construct is more closely associated with human or machine dialogue partners, and provided context to support data analysis.

3.1.4 Results. Participants produced a total of 266 construct pairs, 246 of which were unique pairings. For brevity a sample of these word pairs are shown in Table 2, with a full list of word pairs available in supplementary materials.

3.2 Word Pair Pool Generation Phase 2: Subjective Questionnaire Review

3.2.1 Research Design and Procedure. Findings from the RGT study provide a strong starting point, with 246 word pairs produced. However, to ensure the set of word pairs provided comprehensive coverage, we also conducted a review of relevant subjective questionnaires. This involved a review of all subjective questionnaire metrics identified in a recent systematic review of speech interface research in HCI [33]. We also conducted a Google Scholar search for subjective questionnaires used to measure concepts related to partner modelling, namely: theory of mind (ToM); mental models; perspective taking; metacognition; anthropomorphism and dehumanisation; and social-cognition. Each of these topics was used as a search term, prefaced by the terms 'questionnaire', 'survey' and 'subjective measure'. After reviewing a total of 75 measures, 44 were identified as containing items that could contribute to the pool of word pairs being generated. These included established and bespoke HCI usability measures used in previous speech and HMD research (n=17), and established measures from socio-cognitive psychology covering the range of topics outlined above (n=27). Contributing questionnaires and specific items co-opted from them are included in Table 3. A full list of questionnaires and co-opted items are provided in supplementary materials. The vast majority of the measures reviewed here adopted Likert scale response options,

many in conjunction with semantic differential scales similar to what participants produced using RGT in Phase 1. Where questionnaire items were in the form of a short phrase (e.g. "The system was pleasant" - SASSI [68]), the key adjective 'pleasant' was extracted and an appropriate antonym was generated either from other items on the same scale or by the lead author using a thesaurus.

3.2.2 Results. The review yielded a further 155 word pairs: 86 word pairs coming from speech interface and HMD usability and user experience metrics, and 67 from established measures in socio-cognitive psychology. When combined with the RGT results, the word pair pool after both generation phases stands at a total of 401 pairs of words, which were then screened as outlined below.

3.3 Word Pair Pool Screening

3.3.1 Procedure. Initial screening was carried out by the lead author to remove duplicates and near duplicates (word pairs that offered little semantic differentiation; e.g. 'simple/complex' and 'simple/complicated'- only 'simple/complex' was retained). Word pairs that were considered too esoteric or vague were also adjusted (e.g. unfettered/restricted became free/restricted) or removed (e.g. conjunctive/uncoordinated). Multiple word expressions (N=15) were simplified (e.g. 'Responsive or adaptive/Rigid or fixed response'. to 'Adaptive/Fixed'). Transcriptions of talk aloud data were used to ensure accurate transformation. Finally, word pairs that were deemed obviously unrelated to the concept of partner modelling (e.g. infectious/uncommunicable), were removed in accordance with best practice guidelines on item pool screening [81]. These ensure that the word pairs retained provide adequate coverage, retaining as much nuance between them as possible, whilst ensuring retained pairs are relevant to the concept being addressed. This initial screening process reduced the pool of items from 401 to 127 word pairs. This drop is largely accounted for by a high degree of redundancy when both item pools were combined. In cases where word pairs were similar to the RGT generated pairs, the user-generated RGT pairs were prioritised.

The remaining 127 word pairs were then systematically screened independently by two researchers with expertise in HCI, speech interaction, partner modelling, dialogue and socio-linguistics research. To guide the screening process researchers were provided with Kline's [81] guidelines (outlined above) and a working definition of partner modelling (outlined below). The

Table 2: Sample word pairs generated from repertory grid study. Full word pair list are included in Supplementary Material

Opinionated/Non-judgmental; Biased/Neutral; Free/Bookish; Expansive/Limited;
Spontaneous/Pre-programmed; Colloquial/Universal knowledge;
Abstract/Specific knowledge; Lateral/Inflexible thinking; Personal relatability/Manufactured;
Genuineness/Ungenuine; Real/Fake; Canny/Uncanny; Emotional/Cold;
Personal/Robotic; Connection/Disconnected-disinterested; Engaged/Remote;
Humour/Humourless; Expansive/To-the-point; Convenience/Inconvenience;
Elaborate/Pointed; Polite/Blunt or rude; Colloquial/Formal; Vague/Detailed;
Two-way/One-way; Conversive/Monologue; Humanness/Machineness; Real/Organic-Artificial;
Personalised/Commercialised; No agenda/Agenda; To help/To serve

working definition was derived from a literature review of seminal work on mental models (e.g. [43, 73, 74, 94, 95]), early work examining partner models in HHD and HMD interactions (e.g. [16, 17, 37, 38, 51]), and definitions of ToM (e.g. [6]). The definition is designed to capture the dynamic [17, 74], adaptive [17, 74] and multidimensional [49, 74] nature of partner modelling, with a focus on perceptions of functional, cognitive and empathetic qualities of a dialogue partner that, according to ToM literature [6], are likely to influence interactions. It also incorporates key influences on partner models found in dialogue research, namely: stereotypes about the cultural communities a dialogue partner might belong to, and direct experience interacting with a particular dialogue partner [16, 32]. Both are regarded as fundamental sources of information in formulating and updating global and local partner models, respectively [17].

> *The term partner model refers to an interlocutor's cognitive representation of beliefs about their dialogue partner's communicative ability. These perceptions are multidimensional and include judgements about cognitive, empathetic and/or functional capabilities of a dialogue partner. Initially informed by previous experience, assumptions and stereotypes, partner models are dynamically updated based on a dialogue partner's behaviour and/or events during dialogue.*

Along with the definition, and best practice guidelines [81], the researchers were also provided with a spreadsheet containing the remaining 127 word pairs (see supplementary material). Researchers were instructed to review the pool independently, indicating which word pairs they felt were relevant, not relevant and items they were unsure about. In cases where they were unsure they were asked to comment on their reason for being unsure, providing details as to whether they were unsure about one or both terms in a word pair, and/or why they felt it was not suitable (i.e. not relevant to the concept, too vague or esoteric, or a more appropriate item is already contained within the pool).

3.3.2 Results. The two domain experts independently agreed upon the retention of 24 word pairs and the rejection of 26 word pairs. The experts then met, along with the lead author, to discuss areas of disagreement (87 word pairs). Following the discussion a further 27 word pairs were retained leaving a total of 51 to be included in an online questionnaire. Table 4 shows all retained items following

the screening process. All eliminated items are included in the supplementary material.

3.4 Quantifying Perceptions: Online Study and Principal Component Analysis

3.4.1 Research Design. The next step involved presenting the 51 word pairs to participants through an online survey, which they used to rate their past experiences with the speech agent they interacted with most frequently. Word pairs were presented in the form of a questionnaire. Taking this empirical approach allows for the identification of word pair clusters. These then dictate the underlying structure of the concept with the strongest common terms in each cluster determining the meaning/context of a given dimension. Given the nature of the data produced using RGT, and that most measures reviewed used a similar response structure, we opted to use a 7-point semantic differential scale. Like with the RGT, this creates a scale were participants indicate where they feel speech agents sit between two opposite word poles.

3.4.2 Participants. 390 participants completed the online questionnaire, recruited through email, posters and social media. Participants who completed the questionnaire were entered into a €200 voucher prize draw. From the 390, 34 participants were excluded due to heavily patterned responses that lack variation (i.e. more than 70% of the same response option, or 90% across just 3 response options) which is seen as evidence of inattentiveness [89]. This means that 356 participants were included in the final analysis. All participants (f=61.5%, m=36.8%, non-binary or prefer not to say=1.7%; age range= 18-70yrs, mean age= 28.5yrs, sd= 10.9) were required to have strong English reading and comprehension proficiency. Within the sample, 35.4% had completed graduate or post-graduate education, 32.3% had completed an undergraduate degree and 29.5% had completed secondary and/or vocational education (remainder preferred not to say).

Participants reported moderate levels of experience with speech interfaces (7 point Likert scale; 1=very infrequent to 7=very frequent; mean=3.8, sd=2.06), using 2.6 (sd= 1.3) different devices on average to access them. Speech agents were by far the most common type of speech interface used, with Apple's Siri being the most frequently accessed (80.1%), followed by Google Assistant (64.3%), Amazon Alexa (58.4%) and Microsoft's Cortana (20.8%). Use of multiple speech interfaces was common, with most participants having used two (39%) or three (23.9%) different interfaces, 29.2% having used only one and 7.9% having used four or five. Accessing speech

Table 3: Subjective Questionnaire Review: HCI measures used in previous speech and HMD research (denoted as **), and established measures from socio-cognitive psychology (denoted as *)

Research area and measures reviewed	Sample of measures reviewed	Sample of extracted words
**Subjective questionnaires used in previous HMD research Potential items taken from 5 [14, 66, 68, 104, 105] out of 8 [21, 98, 118] reviewed	AttrakDiff2: Product appeal based on hedonic and pragmatic qualities [66]; SASSI: Usability and Interactivity [68]; MOS-X: Quality of synthetic speech [104]; SUISQ: User Satisfaction [105]	people-centric/technical, practical/impractical, cumbersome/facile, unmanageable/manageable, isolates/connects, stylish/lacking Style, poor quality/high quality, excludes/draws you in, ugly/pretty appealing/unappealing, consistent, efficient, repetitive, lose track, harsh, raspy, strained
***Bespoke metrics used in previous HMD research Potential items taken from 11 [28, 34, 48, 50, 62, 85, 87, 106, 129] [42, 45, 54] out of 12 [130] reviewed	Usability [130]; User Acceptance of SDS [28, 48]; System Preference [62, 129]; User satisfaction [50]; User experience [34, 42].	understandable/incomprehensible, accessible/inaccessible, appropriate/inappropriate, familiar/unfamiliar, predictable/unpredictable, basic/advanced, capable/incapable, cheap/expensive, good/bad, inflexible/flexible, lacks power/powerful, quick/slow, stable/unstable, amateurish/professional, modern/old fashioned, efficient/inefficient, trustworthy/untrustworthy,
*Theory of Mind Potential items taken from 12 [6, 47, 84, 111, 123, 133, 134] [63, 65, 72, 114] out of 24 measures [1, 23, 78, 102, 110, 120] [9, 19, 64, 75, 76, 124] reviewed	EQ & TEQ: Assessing emotional empathy [123, 134]; FQ: Assessing attitudes toward friendships [6]; SQ-R: Assesses demand for understanding underlying rules [133]; ARSQ: Self-report of resting-state cognition [47]; SSQ: A measure of social support [114]; SC-IQ: Measure of social capital [65].	interrupting, monopolizes conversation, apologetic charitable/humility, meticulous, curious, lonely, upset irritated/irritable, tired, sleepy, friend, confidant superior, distractible, persistent, reflective, helps out caring, lies, good friend, popular, solitary dependable, be yourself around, appreciates you console you, nervous, peaceful, downhearted, blue, social with others, similar goals, source of expertise/advice, willing, able, critical
*Mental Models and Partner Models Potential items taken from 5 [36, 82, 90, 100, 101] out of 8 [4, 20, 121] measures reviewed	MTQ48: Self-report measure of mental toughness [101]; Psi-Q: Self-report measure of mentalizing behaviours [4]; HSQ: Self-report measure of humor style [90]; AAS-r: Attitudes toward romantic relationships [36]; RUQ: Perceived relational uncertainty [82].	supportive, burdensome, influential, shows initiative, nervous, persistent, beautiful, offensive, enthralling, self-deprecating, criticizing, impressive, amusing, absurd, reluctant, committed
*Metacognition Potential items taken from 1 [27] out of 5 [56, 79, 103, 117, 127] measures reviewed	MCQ: Beliefs about worries and intrusive thoughts [27]; MSLQ: Motivational orientation for cognitive motivation for engagement and self-regulation [103]; MAI: Awareness of metacognitive processes [117].	glib, arrogant, selfish, complacent copes, assertive, misleading, confident, embarrassing
*Anthropomorphism and Dehumanisation Potential items taken from 7 [8, 44, 58, 92, 112, 113, 131] out of 12 [7, 22, 29, 30, 46] measures reviewed	Godspeed V: Anthropomorphism in HRI [8]; IDAQ, AAS & ATS: Tendencies to self-engage in anthropomorphic behaviors [29, 30, 131]; CMQ & CTQ: General tendency to believe in conspiracies [22, 46]; MCSDC: Attempts to present oneself in an overly positive manner [44].	life-like, elegant, inert, apathetic, awful agitated, lethargic, intentional, free, conscious, organized, purposeful, self-conscious, ambitious, imaginative, qualified, intense, doubtful, courteous, comforting alarming, trustworthy, deceptive, merciful, deliberate, sentient, curious, fun-loving, sociable aggressive, impatient, jealous, humble
*Perspective taking and social cognition Potential items taken from 2 [26, 96] out of 3 [108] measures reviewed	QCAE: Self-reported cognitive and affective empathy [108]; ESEQ: Assesses emotion specific affective and cognitive empathy [96]; ICQ: Self-reported interpersonal competence [26].	infectious, acquaintance, unreasonable embarrassing, confrontational, companion

Table 4: Retained word pairs after screening phase. Full list of retained and eliminated items are included in Supplementary Material.

Authentic/Fake; Emotional/Clinical; Concise/Verbose; Subjective/Objective;
Expert/Amateur; Empathetic/Apathetic; Reliable/Uncertain; Illogical/Logical;
Authoritative/Unsure; Flexible/Inflexible; Dependable/Unreliable; Assertive/Submissive;
Colloquial/Formal; Warm/Cold; Efficient/Inefficient; Human-like/Machine-like; Interactive/Start-stop;
Life-like/Tool-like; Adaptive/Fixed; Precise/Vague; Contextual/Non-contextual; Competent/Incompetent;
Personal/Generic; Hesitant/Decisive; Two-way/One-way; Assistant/Servant; Intelligent/Unintelligent;
Elaborative/To-the-point; Misleading/Honest; Repetitive/Versatile; Meandering/Direct; Restricted/Free;
Abstract/Concrete; Basic/Advanced; Capable/Incapable; Sincere/Insincere; Consistent/Inconsistent;
Social/Transactional; Trustworthy/Untrustworthy; Confident/Uncertain; Spontaneous/Predetermined;
Cooperative/Uncooperative; Ambiguous/Clear; Broad/Specific; High Feedback/Low Feedback;
Predictable/Unpredictable; Amusing/Serious; Engaged/Disinterested; Complex/Straightforward;
Free/Restricted; Repetitive/Versatile; Authentic/Fake; Feedback High/Feedback Low

agents across multiple devices was also quite common (mean=2.6, sd=1.3), with 26.9% of participants accessing them using between 3 to 6 different devices. Our sample most commonly accessed speech interfaces through smartphones (88.5%) or smart speakers (59.2%), followed by telephony based speech systems (30.6%), laptops (28.1%), in-car assistants (25.8%) and tablets (22.2%).

3.4.3 Procedure. The questionnaire was presented to participants online, using LimeSurvey. After following the link provided in recruitment materials, participants were presented with an information sheet giving full details of the study and their rights in relation to participation and data protection. After giving explicit consent to participate, participants completed a demographic questionnaire gathering information about their age, sex, educational attainment, nationality and their experience with speech interfaces. They were then presented with the 51 word pairs, each separated by a 7 point scale (see Figure 3). The display of word pairings was pseudo-randomised. Reflecting on previous interactions with speech agents, participants were asked to think about the way speech agents communicate with them and then rate the communicative ability of the speech agent they used most frequently on a scale between each of the word pairs displayed. Instructions were given to read each pair of words carefully, to respond as quickly and accurately as possible, and to try and avoid giving too many neutral responses. Participants were then fully debriefed as to the nature and aims of the study.

3.4.4 Data Analysis. We conducted a Principal Component Analysis (PCA) using the *psych* [109] and *GPArotation* [13] packages in R (Version 1.1.456) [107] so as to identify the dimensions present in the 51 word pairs. The primary purpose of PCA is to reduce the dimensionality of multivariate data, allowing for a large number of variables to be summarized within smaller subsets, or factors [24, 41]. PCA was deemed most suitable as it does not require an *a priori* hypothesized or predetermined factor structure, making it

ideal for exploratory analysis [24, 41]. We note that various recommendations are made regarding what constitutes a suitable sample size for conducting reliable PCA. A minimum sample size of 100 is required [80], with little difference seen in resultant factor structures when samples exceed 200 participants [57]. Based on this our sample of 356 is deemed suitable for PCA.

Based on best practice guidelines to ensure reliable and clear factor structures [35], we first removed word pairs with weak inter-item correlations before conducting the analysis. Using established thresholds [35], word pairs with low mean inter-item correlations (r < .15) were removed, resulting in 14 word pairs being eliminated and 37 word pairs being included in the PCA. Kaiser-Meyer-Olkin (KMO) test of sampling adequacy was high overall (KMO= .91) for the remaining data, and across word pairs (KMO range= .95 to .81). Bartlett's test was also statistically significant [x2(666) = 4913.29, p<.001.] suggesting the data was suitable for PCA analysis.

Following [57], a first PCA iteration was conducted with all items (word pairs) set as factors, to produce eigenvalues that are used to assess the number of factors to be retained. Here, the number factors retained in the rotated PCA was based on parallel analysis using the Hornpa [69] function in R. Considered a more robust approach than traditional methods such as scree plots or Kaiser criterion [57], in parallel analysis the number of factors that have higher eigenvalues than a set of simulated eigenvalues are retained. Simulated eigenvalues are generated from the original data set, with the number of simulations being set as a parameter (here 1000 simulations were run) [57]. Results of the parallel analysis suggested that 3 factors should be retained.

Next, PCA was conducted setting the number of factors to 3 and using direct oblimin rotation, the approach recommended when underlying dimensions are likely to be related [57]. Based on best practice guidelines [35, 57] we then iteratively removed word pairs with weak communalities (<0.4), weak loadings (<0.5) and multiple cross loadings [35], until close to a mean communality of 0.5 was achieved. This led to a further 14 word pairs being removed.
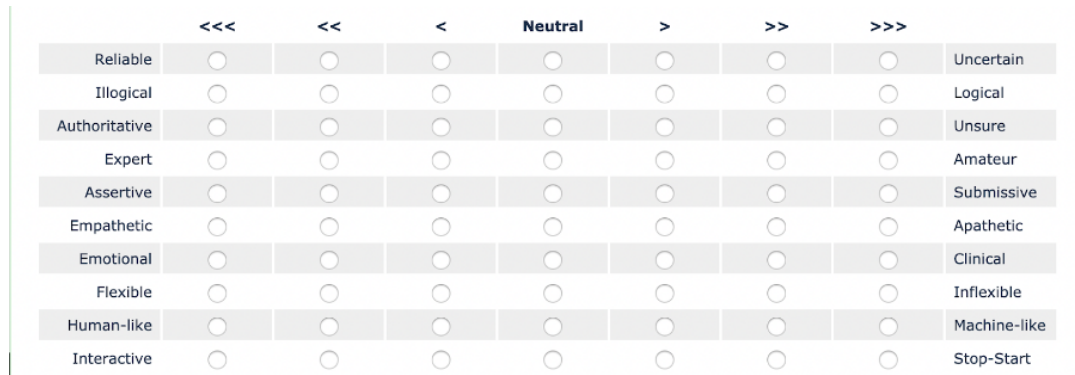
Figure 3: Example questionnaire structure

Table 5: Factor loadings for 3 factor PCA. Only loadings above 0.3 are displayed. All items removed during PCA are included in Supplementary Material.

Items	Factor 1: Partner Competence & Dependability	Factor 2: Human-likeness	Factor 3: Cognitive Flexibility
Competent/Incompetent	0.76		
Dependable/Unreliable	0.68		
Capable/Incapable	0.68		
Consistent/Inconsistent	-0.67		
Reliable/Uncertain	0.66		
Ambiguous/Clear	0.65		
Meandering/Direct	-0.64		
Expert/Amateur	-0.64		
Efficient/Inefficient	0.64		0.32
Misleading/Honest	0.63		-0.34
Precise/Vague	-0.62		
Cooperative/Uncooperative	0.54		
Human-like/Machine-like		0.75	
Life-like/Tool-like		0.75	
Warm/Cold		0.68	
Empathetic/Apathetic		0.66	
Personal/Generic		0.62	
Authentic/Fake		0.56	
Social/Transactional		0.54	
Flexible/Inflexible			0.66
Interactive/Start-stop			0.61
Interpretive/Literal			0.56
Spontaneous/Predetermined			0.51
Eigenvalues	5.45	3.57	2.18
Proportion Variance	24%	16%	9%
Cumulative Variance	24%	39%	49%
Factor Correlations			
Factor 1: Partner Competence & Dependability	-	0.21	0.11
Factor 2: Human-likeness	0.21	-	0.36
Factor 3: Cognitive Flexibility	0.11	0.36	-

Following pruning of these 14 pairs the 3 factor model exhibited acceptable fit (0.96), acceptable mean item complexity (1.3), acceptable squared residuals (0.05) and accounted for 49% of variance within the data. The final word pair clusters and factor structure are shown in Table 5. The factors revealed by the 3 factor model reflect dimensions that describe perceptions of: partner competence and dependability; partner's human-likeness; and partner's cognitive flexibility. Details regarding word pairs eliminated during PCA are included in the supplementary material.

4 DISCUSSION

Our research took a psycholexical approach in mapping partner models as a concept, identifying key dimensions that constitute a user's partner model for speech agents. First, through using the repertory grid technique (RGT), a total of 246 unique word pairs were generated by users to describe their partner models of speech agents. This data was complemented by a further 155 word pairs identified through a search of subjective questionnaires applicable to partner modelling related concepts. After screening the 401 word pairs, a selection of 51 word pairs were included in an online study of 356 speech agent users. These users were asked to rate the ability of speech agents as dialogue partners based on previous experience. Through principal component analysis (PCA), where a further 28 word pairs were eliminated, three key dimensions of a user's partner model for speech agents were identified. These key dimensions reflected perceptions of a dialogue partner's: 1) competence and dependability (emerging from perceptions of competence, reliability and precision); 2) human-likeness (whether the speech agent is perceived as human-like, warm, social or transactional); and 3) cognitive flexibility (whether the speech agent is perceived as flexible, interactive or spontaneous). For a full list of attributes within each dimension see Table 5. This is a significant contribution in that it not only outlines the multidimensional nature of partner models in speech agent interaction, but adds specific structure to the concept that, to-date, has been lacking.

4.1 The Influence of Design on Partner Models

Our study adds much needed definition to the concept of user partner models. This should allow researchers to gather deeper insight into how design decisions may influence these models. Earlier work hypothesises that design choices, such as accent [40] and anthropomorphic dialogue strategies [18] affect partner modelling. Yet to date, it has not been possible to identify what specific aspects of a partner model are influenced by these choices, with studies using behavioural adaptation as evidence of general model change and influence [16, 40]. Our work opens the possibility that these design decisions do not universally impact a user's model, being more nuanced in their effect. For instance, rather than influencing cognitive flexibility judgements, accent-based design choices may alter estimates of partner knowledge (relevant to competence and dependability) and human-likeness, making those dimensions more likely drivers of linguistic adaptation proposed [40]. Echoing recent work, human-likeness in design tends to inform initial partner model development [40, 42, 49]. To ensure partner models are accurate, human-like design should be congruent with the level of system capability [91]. Our work gives a framework to help

identify how human-like design choices may impact perceptions of human-likeness alongside other associated partner model dimensions such as perceptions of cognitive flexibility, and competence and dependability. The dimensionality identified is also useful for informing how other design choices may influence partner models. For example, expressive synthesis [25] and the use of more social talk [61] are likely to have an influence on specific model dimensions. Our work is an important first step in allowing researchers to explore how specific design choices affect these models more specifically. It is important to note that, rather than suggesting designers implement these partner model dimensions in speech interfaces, our findings identify perceptions that may be influenced by design changes.

4.2 Partner Model Dimensionality, Salience and Dynamics

Our findings emphasise that people's partner models are clearly more detailed and complex than more general descriptions of speech agents as at risk listeners [15, 39, 40, 88, 97], poor [37] or basic dialogue partners [16]. While the number of dimensions that are reflected on simultaneously is open to debate [74, 94], it is likely that dimensions may becoming more or less salient in different contexts and over the course of interaction. For instance, the salience of dimensions may vary within certain situational contexts, such as when using an agent in health, wellbeing or care domains [83, 122] where human-likeness and perceptions of empathy are important. Indeed, events during speech agent dialogue may also bring dimensions to the fore, such as negotiating errors and miscommunications highlighting capability and flexibility judgements. The idea that certain aspects of a partner model will be more or less salient in response to specific system behaviours, dialogue events or contexts is similar to the idea of one-bit processing [17]. It also echoes accounts of how inaccurate mental models are amended [74], and how partner specific information is incorporated in perspective taking [51]. All suggest that, models of a partner (or object) need not be comprehensive at all times, with specific dimensions dominating perceptions at different moments during the interaction or in response to dialogue events. With our research now identifying dimensions of speech agent partner models, future work can build on this by examining the influence of specific interaction events and context on model use. It also opens avenues for exploring how partner models might impact language production dynamically during HMD.

4.3 The Interdependence and Dynamism of Partner Models

Although this work significantly expands on the dimensionality of partner models as a concept, our results do not make any inferences about the causal relationships between the dimensions identified. However, it is highly likely that, although distinct, these dimensions are interdependent, with changes in one dimension impacting or affecting changes in another. For instance, it may be that changes to the perceived human-likeness of a system may lead to increases in perceptions of partner competence and dependability. This is eluded to in recent research, whereby the human-likeness of systems is

seen to act as an anchor for initial perceptions of what a system knows and can do [42, 88]. Work suggests that early attention to anthropomorphic characteristics leads to high expectations in regard to competence and dependability, which are quickly identified as unrealistic following interactions [88]. Whilst work examining dynamic adaptations of partner models in response to dialogue events has been somewhat limited to-date, available accounts support our assertion that partner models are adaptive. For example, Leahu et al. [86] suggest that people use broad partner types (e.g. human and machine) to make comparisons across specific dimensions (i.e. humor and/or intelligence), whilst dynamically working towards a more accurate model [86]. Human dialogue work [17] also emphasises that partner models may evolve as a user's initial stereotype driven perceptions (e.g. global model) are fashioned into a more accurate, experience-based local model specific to a dialogue partner. Similar effects may occur within speech agent dialogue, where a user's initial perception of an agent becomes more nuanced once informed by direct experiences with a particular agent over time. An open question also relates to how these more nuanced models may then feedback to influence a user's global model to inform initial interactions with new, unfamiliar speech agents. Findings from our work open avenues for examining the interdependent and dynamic relationship between partner model dimensions, with a level of detail that was not previously possible. Future research efforts should focus on exploring how perceptions on these dimensions change over time, how they become more nuanced with experience and how this experience may feedback to inform global models of speech agents.

5 LIMITATIONS AND FUTURE WORK

The triading paradigm used in RGT requires participants to be presented with three exemplars, two similar and one dissimilar, to provoke reflection about key characteristics of an object of interest (speech agents) and how they may be similar or different to an appropriate comparator (a human). Although users readily make comparisons between humans and machines in speech agent interaction without being instructed to do so [38, 88], the triading may have made these more likely. Previous work emphasizes that human comparison is core to speech agent partner model building and research [16, 39, 42, 86, 88]. Following this, we used a human comparator to prompt word-pair generation in word pool generation phase 1. In the later online study, responses to these word pairs were given specifically in relation to speech agent interaction only. Future work could add more speech agent elements - such as other speech interfaces and/or social robots with more or less human-like qualities - to gather a wider range of constructs, adding further granularity.

To ensure initial word pairs accurately reflected speech agent perceptions, participants in the RGT study supplied words after direct interactions. For the online questionnaire study participants were asked to reflect on past experiences, rather than an interaction experienced directly prior to responding. This reflective approach was deemed most appropriate for building a general account of partner models as it reduces the potential for the online questionnaire responses being influenced by a specific agent or interaction encounter.

Through the execution of the study we produced a set of 401 word pairs that describe a user's partner model of speech agents. Much like in personality research, where the psycholexical approach is commonly used, the items produced are not only helpful in categorising and understanding the dimensionality of partner models, but can also form the basis of a self-report metric for measuring them. The current study is a significant step in developing such a questionnaire as it produces the item set and gives us an initial potential factor structure. Our future work aims to further develop the final word pair set into a fully validated partner modelling questionnaire. To do this we aim to conduct work to assess scale reliability (e.g. internal consistency and test-retest reliability) and validity (concurrent, discriminant and predictive validity testing), whilst performing confirmatory factor analysis on future datasets to ensure that the factor structure identified in this paper is robust [80]. High factor loading items could be used as building blocks for a short-form scale, although this would need to be statistically validated.

Whilst statistical approaches like PCA can result in the loss of some rich qualitative insights, their aim is to ensure robust clustering of word pairs to identify emergent factors. Further work could add to our dataset, through research with additional speech agents, to identify additional dimensions.

Although our work has relevance for robotics and virtual agent research, it is also important to note that our scope is limited to identifying partner model dimensions for non-embodied speech agents, where speech is the primary if not exclusive form of communication. Work examining perceptions of embodied agents highlights unique considerations that may be incorporated in partner models when interacting with robots [8, 113] or avatars [5], such as animacy and/or safety. These are underpinned by the embodied nature of these interaction paradigms. Further work should look to replicate and build on our work within these domains.

6 CONCLUSION

As the ubiquity of speech interfaces continues to increase, more people are now engaging with speech agents on a daily basis. Although research has consistently emphasised the importance of our perceptions toward a system's capability as a dialogue partner (i.e. our partner models) in guiding interaction, the concept as it currently stands is poorly defined. Our work aimed to give structure to this concept by identifying the key dimensions of a user's partner model. Through principal component analysis we identified that partner models for speech agents hold three key dimensions, which focus on perceptions of a dialogue partner's competence and dependability, human-likeness and apparent cognitive flexibility. This not only adds granularity, clarity and definition to the concept, but also highlights that there are multiple dimensions for designers to consider when aiming to support users and improve their interaction experience.

REFERENCES

[1] Icek Ajzen. 2006. Constructing a theory of planned behavior questionnaire.

[2] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2018. Expressive speech synthesis via modeling expressions with variational autoencoder. *arXiv preprint arXiv:1804.02135* (2018).

[3] René Amalberti, Noëlle Carbonell, and Pierre Falzon. 1993. User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies* 38, 4 (1993), 547–566. https://doi.org/10.1006/imms.1993.1026

[4] Jackie Andrade, Jon May, Catherine Deeprose, Sarah-Jane Baugh, and Giorgio Ganis. 2014. Assessing vividness of mental imagery: The Plymouth Sensory Imagery Questionnaire. *British Journal of Psychology* 105, 4 (2014), 547–563. https://doi.org/10.1111/bjop.12050

[5] Benjamin Balas, Lauren Tupa, and Jonathan Pacella. 2018. Measuring social variables in real and artificial faces. *Computers in Human Behavior* 88 (2018), 236–243.

[6] Simon Baron-Cohen and Sally Wheelwright. 2003. The Friendship Questionnaire: An Investigation of Adults with Asperger Syndrome or High-Functioning Autism, and Normal Sex Differences. *Journal of Autism and Developmental Disorders* 33, 5 (2003), 509–517. https://doi.org/10.1023/A:1025879411971

[7] Justin L Barrett and Frank C Keil. 1996. Conceptualizing a nonnatural entity: Anthropomorphism in God concepts. *Cognitive psychology* 31, 3 (1996), 219–247.

[8] 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 (2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[9] Bernard M Bass and Bruce J Avolio. 2004. *Multifactor Leadership Questionnaire: MLQ; manual and sampler set.* Mind Garden.

[10] Allan Bell. 1984. Language style as audience design. *Language in society* 13, 2 (1984), 145–204.

[11] Linda Bell and Joakim Gustafson. 1999. Interaction with an animated agent in a spoken dialogue system. In *Sixth European Conference on Speech Communication and Technology.*

[12] Kirsten Bergmann, Holly P. Branigan, and Stefan Kopp. 2015. Exploring the Alignment Space: Lexical and Gestural Alignment with Real and Virtual Humans. *Frontiers in ICT* 2 (2015). https://doi.org/10.3389/fict.2015.00007

[13] Coen A. Bernaards and Robert I. Jennrich. 2005. Gradient Projection Algorithms and Software for Arbitrary Rotation Criteria in Factor Analysis. *Educational and Psychological Measurement* 65 (2005), 676–696.

[14] Johan Bos, Staffan Larsson, I Lewin, C Matheson, and D Milward. 1999. Survey of existing interactive systems. *Trindi (Task Oriented Instructional Dialogue) report* D1 (1999), 3.

[15] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (2010), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

[16] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 1 (2011), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011

[17] Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. 2010. Two Minds, One Dialog. In *Psychology of Learning and Motivation*. Vol. 53. Elsevier, 301–344. https://doi.org/10.1016/S0079-7421(10)53008-1

[18] Susan E. Brennan and Justina O. Ohaeri. 1994. Effects of message style on users' attributions toward agents. In *Conference companion on Human factors in computing systems - CHI '94* (Boston, Massachusetts, United States). ACM Press, 281–282. https://doi.org/10.1145/259963.260492

[19] Donald E Broadbent, P Fitzgerald Cooper, Paul FitzGerald, and Katharine R Parkes. 1982. The cognitive failures questionnaire (CFQ) and its correlates. *British journal of clinical psychology* 21, 1 (1982), 1–16.

[20] Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45, 4 (2009), 624 – 634. https://doi.org/10.1016/j.jesp.2009.02.016

[21] John Brooke. 1996. SUS: a "quick and dirty'usability. *Usability evaluation in industry* (1996), 189.

[22] Martin Bruder, Peter Haffke, Nick Neave, Nina Nouripanah, and Roland Imhoff. 2013. Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy Mentality Questionnaire. *Frontiers in psychology* 4 (2013), 225.

[23] Martin Brüne. 2005. Emotion recognition,'theory of mind,'and social behavior in schizophrenia. *Psychiatry research* 133, 2-3 (2005), 135–147.

[24] F.B. Bryant and P.R. Yarnold. 1995. Principal-components analysis and exploratory and confirmatory factor analysis. In *Reading and understanding multivariate statistics*. A.P.A., 99–136.

[25] Christopher G Buchanan, Matthew P Aylett, and David A Braude. 2018. Adding personality to neutral speech synthesis voices. In *International Conference on Speech and Computer*. Springer, 49–57.

[26] Duane Buhrmester, Wyndol Furman, Mitchell T. Wittenberg, and Harry T. Reis. 1988. Five domains of interpersonal competence in peer relationships. *Journal of Personality and Social Psychology* 55, 6 (1988), 991–1008. https://doi.org/10.1037/0022-3514.55.6.991

[27] Sam Cartwright-Hatton and Adrian Wells. 1997. Beliefs about Worry and Intrusions: The Meta-Cognitions Questionnaire and its Correlates. *Journal of Anxiety Disorders* 11, 3 (1997), 279–296. https://doi.org/10.1016/S0887-6185(97)00011-X

[28] Sherry Perdue Casali, Beverly H. Williges, and Robert D. Dryden. 1990. Effects of Recognition Accuracy and Vocabulary Size of a Speech Recognition System on Task Performance and User Acceptance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 32, 2 (1990), 183–196. https://doi.org/10.1177/001872089003200206

[29] Matthew G Chin, Valerie K Sims, Bryan Clark, and Gabriel Rivera Lopez. 2004. Measuring individual differences in anthropomorphism toward machines and animals. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 48. SAGE Publications Sage CA: Los Angeles, CA, 1252–1255.

[30] Matthew G Chin, Ryan E Yordon, Bryan R Clark, Tatiana Ballion, Michael J Dolezal, Randall Shumaker, and Neal Finkelstein. 2005. Developing and anthropomorphic tendencies scale. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49. SAGE Publications, Los Angeles, CA, 1266–1268.

[31] Vincent Cho and Robert Wright. 2010. Exploring the evaluation framework of strategic information systems using repertory grid technique: a cognitive perspective from chief information officers. *Behaviour & Information Technology* 29, 5 (2010), 447–457. https://doi.org/10.1080/01449290802121206

[32] Herbert H. Clark. 1996. *Using language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

[33] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interact with Computers* (2019), 29.

[34] Leigh Clark, Abdulmalik Ofemile, Svenja Adolphs, and Tom Rodden. 2016. A Multimodal Approach to Assessing User Experiences with Agent Helpers. *ACM Transactions on Interactive Intelligent Systems* 6, 4 (2016), 1–31. https://doi.org/10.1145/2983926

[35] Lee Anna Clark and David Watson. 2016. Constructing validity: Basic issues in objective scale development. (2016).

[36] Nancy L Collins. 1996. Working models of attachment: Implications for explanation, emotion, and behavior. *Journal of personality and social psychology* 71, 4 (1996), 810.

[37] Benjamin R Cowan. 2014. Understanding speech and language interactions in HCI: The importance of theory-based human-human dialogue research. (2014), 4.

[38] Benjamin R Cowan and Holly Branigan. 2017. They Know as Much as We Do: Knowledge Estimation and Partner Modelling of Artificial Partners. (2017), 6.

[39] Benjamin R. Cowan, Holly P. Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human computer dialogue. *International Journal of Human-Computer Studies* 83 (2015), 27–42. https://doi.org/10.1016/j.ijhcs.2015.05.008

[40] Benjamin R. Cowan, Philip Doyle, Justin Edwards, Diego Garaialde, Ali Hayes-Brady, Holly P. Branigan, João Cabral, and Leigh Clark. 2019. What's in an accent?: the impact of accented synthetic speech on lexical choice in human-machine dialogue. In *Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19* (Dublin, Ireland). ACM Press, 1–8. https://doi.org/10.1145/3342775.3342786

[41] B. R. Cowan and M. A. Jack. 2014. Measuring Anxiety Towards Wiki Editing: Investigating the Dimensionality of the Wiki Anxiety Inventory-Editing. *Interacting with Computers* 26, 6 (2014), 557–571. https://doi.org/10.1093/iwc/iwt050

[42] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can I help you with?": infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '17* (Vienna, Austria). ACM Press, 1–12. https://doi.org/10.1145/3098279.3098539

[43] Kenneth J. W. Craik. 1943. The Nature of Explanation. *The Journal of Philosophy* 40, 24 (1943), 667. https://doi.org/10.2307/2018933

[44] Douglas P Crowne and David Marlowe. 1960. A new scale of social desirability independent of psychopathology. *Journal of consulting psychology* 24, 4 (1960), 349.

[45] Nils Dahlbäck, QianYing Wang, Clifford Nass, and Jenny Alwin. 2007. Similarity is more important than expertise: accent effects in speech interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07* (San Jose, California, USA). ACM Press, 1553. https://doi.org/10.1145/1240624.1240859

[46] Hannah Darwin, Nick Neave, and Joni Holmes. 2011. Belief in conspiracy theories. The role of paranormal belief, paranoid ideation and schizotypy. *Personality and Individual Differences* 50, 8 (2011), 1289–1293.

[47] B Alexander Diaz, Sophie Van Der Sluis, Sarah Moens, Jeroen S Benjamins, Filippo Migliorati, Diederick Stoffers, Anouk Den Braber, Simon-Shlomo Poil, Richard Hardstone, Dennis Van't Ent, et al. 2013. The Amsterdam Resting-State Questionnaire reveals multiple phenotypes of resting-state cognition. *Frontiers in human neuroscience* 7 (2013), 446.

[48] DL Dintruff, DG Grice, and TG Wang. 1985. User acceptance of speech technologies. *Speech Technology* 2, 4 (1985), 16–21.

[49] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei Taiwan). ACM, 1–12. https://doi.org/10.1145/3338286.3340116

[50] Louise Dulude. 2002. Automated telephone answering systems and aging. *Behaviour & Information Technology* 21, 3 (2002), 171–184. https://doi.org/10.1080/0144929021000013482

[51] Nicholas Duran, Rick Dale, and Alexia Galati. 2016. Toward Integrative Dynamic Models for Adaptive Perspective Taking. *Topics in Cognitive Science* 8, 4 (2016), 761–779. https://doi.org/10.1111/tops.12219

[52] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication* 50, 8 (2008), 630–645. https://doi.org/10.1016/j.specom.2008.04.002

[53] Jens Edlund, Julia Bell Hirschberg, and Mattias Heldner. 2009. Pause and gap length in face-to-face interaction. *Columbia University* (2009). https://doi.org/10.7916/d82f7wt9

[54] Rochelle E. Evans and Philip Kortum. 2010. The impact of voice characteristics on user response in an interactive voice response system. *Interacting with Computers* 22, 6 (2010), 606–614. https://doi.org/10.1016/j.intcom.2010.07.001

[55] Daniel Fallman and John Waterworth. 2010. Capturing User Experiences of Mobile Information Technology With the Repertory Grid Technique. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments* 6, 2 (2010), 250–268. https://doi.org/10.17011/ht/urn.201011173094

[56] Bruce A. Fernie, Marcantonio M. Spada, Ana V. Nikčević, George A. Georgiou, and Giovanni B. Moneta. 2009. Metacognitive Beliefs About Procrastination: Development and Concurrent Validity of a Self-Report Questionnaire. *Journal of Cognitive Psychotherapy* 23, 4 (2009), 283–293. https://doi.org/10.1891/0889-8391.23.4.283 arXiv:https://connect.springerpub.com/content/sgrjcp/23/4/283.full.pdf

[57] Andy Field, Jeremy Miles, and Zoë Field. 2013. Discovering Statistics Using R by Andy Field, Jeremy Miles, Zoë Field. *International Statistical Review* 81, 1 (2013), 169–170. https://doi.org/10.1111/insr.12011_21

[58] Yannick Forster, Frederik Naujoks, and Alexandra Neukum. 2017. Increasing anthropomorphism and trust in automated driving functions by adding speech output. In *2017 IEEE intelligent vehicles symposium (IV)*. IEEE, 365–372.

[59] Fay Fransella, Richard Bell, and D. Bannister. 2004. *A manual for repertory grid technique* (2nd ed ed.). John Wiley & Sons.

[60] Susan R. Fussell and Robert M. Krauss. 1992. Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology* 62, 3 (1992), 378–391. https://doi.org/10.1037/0022-3514.62.3.378

[61] Emer Gilmartin, Marine Collery, Ketong Su, Yuyun Huang, Christy Elias, Benjamin R. Cowan, and Nick Campbell. 2017. Social talk: making conversation with people and machine. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents - ISIAA 2017* (Glasgow, UK). ACM Press, 31–32. https://doi.org/10.1145/3139491.3139494

[62] Li Gong and Jennifer Lai. 2001. Shall we mix synthetic speech and human speech?: impact on users' performance, perception, and attitude. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01* (Seattle, Washington, United States). ACM Press, 158–165. https://doi.org/10.1145/365024.365090

[63] Robert Goodman. 2001. Psychometric Properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry* 40, 11 (2001), 1337–1345. https://doi.org/10.1097/00004583-200111000-00015

[64] Frank M Gresham and Stephen N Elliott. 1990. *Social skills rating system: Manual.* American Guidance Service.

[65] Christiaan Grootaert. 2004. Measuring social capital: an integrated questionnaire. no. 18 (2004).

[66] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003*. Springer, 187–196.

[67] Trevor Hogan and Eva Hornecker. 2013. Blending the repertory grid technique with focus groups to reveal rich design relevant insight. In *Proceedings of the 6th International Conference on Designing Pleasurable Products and Interfaces - DPPI '13* (Newcastle upon Tyne, United Kingdom). ACM Press, 116. https://doi.org/10.1145/2513506.2513519

[68] Kate S. Hone and Robert Graham. 2000. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering* 6, 3 (2000), 287–303. https://doi.org/10.1017/S1351324900002497

[69] Francis Huang. 2015. *Horn's (1965) Test to Determine the Number of Components/Factors.* (Version 1).

[70] Elin Jacob and Debora Shaw. 1998. Sociocognitive Perspectives on Representation. *Annual Review of Information Science and Technology* 33 (1998), 131–85.

[71] Devi Jankowicz. 2004. *The easy guide to repertory grids.* Wiley.

[72] Alan M Jette, Allyson R Davies, Paul D Cleary, David R Calkins, Lisa V Rubenstein, Arlene Fink, Jacqueline Kosecoff, Roy T Young, Robert H Brook, and Thomas L Delbanco. 1986. The functional status questionnaire. *Journal of general internal medicine* 1, 3 (1986), 143–149.

[73] Philip N. Johnson-Laird. 1980. Mental Models in Cognitive Science. *Cognitive Science* 4 (1980), 71–115.

[74] P. N. Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences* 107, 43 (2010), 18243–18250. https://doi.org/10.1073/pnas.1012933107

[75] AF Jorm. 1994. A short form of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): development and cross-validation. *Psychological medicine* 24, 1 (1994), 145–153.

[76] Shaughan A Keaton. 2017. Interpersonal Reactivity Index (IRI) (Davis, 1980). *The Sourcebook of listening research: Methodology and measures* (2017), 340–347.

[77] George Kelly. 1991. *The psychology of personal constructs.* Routledge in association with the Centre for Personal Construct Psychology.

[78] Peter Kinderman and Richard P Bentall. 1996. A new measure of causal locus: the internal, personal and situational attributions questionnaire. *Personality and Individual differences* 20, 2 (1996), 261–264.

[79] Sabina Kleitman and Lazar Stankov. 2007. Self-confidence and metacognitive processes. *Learning and Individual Differences* 17, 2 (2007), 161 – 173. https://doi.org/10.1016/j.lindif.2007.03.004

[80] Paul Kline. 2000. *A psychometrics primer.* Free Association. OCLC: 833721971.

[81] Paul Kline. 2013. *Handbook of Psychological Testing* (2 ed.). Routledge. https://doi.org/10.4324/9781315812274

[82] Leanne K. Knobloch and Denise Haunani Solomon. 2005. Relational Uncertainty and Relational Information Processing: Questions without Answers? *Communication Research* 32, 3 (2005), 349–388. https://doi.org/10.1177/0093650205275384

[83] A. Baki Kocaballi, Juan C. Quiroz, Liliana Laranjo, Dana Rezazadegan, Rafal Kocielnik, Leigh Clark, Q. Vera Liao, Sun Young Park, Robert J. Moore, and Adam Miner. 2020. Conversational Agents for Health and Wellbeing. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3375154

[84] Chris E. Lalonde and Michael J. Chandler. 1995. False belief understanding goes to school: On the social-emotional consequences of coming early or late to a first theory of mind. *Cognition & Emotion* 9, 2 (1995), 167–185. https://doi.org/10.1080/02699939508409007

[85] Lars Bo Larsen. 2003. Assessment of Spoken Dialogue System Usability - What are We really Measuring?. In *Proceedings from EuroSpeech 2003 - Interspeech 2003 8th European Conference on Speech Communication and Technology* (Geneva). ISCA.

[86] Lucian Leahu, Marisa Cohn, and Wendy March. 2013. How categories come to matter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (Paris, France). ACM Press, 3331. https://doi.org/10.1145/2470654.2466455

[87] Kwan Min Lee and Clifford Nass. 2003. Designing social presence of social actors in human computer interaction. In *Proceedings of the conference on Human factors in computing systems - CHI '03* (Ft. Lauderdale, Florida, USA). ACM Press, 289. https://doi.org/10.1145/642611.642662

[88] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (Santa Clara, California, USA). ACM Press, 5286–5297. https://doi.org/10.1145/2858036.2858288

[89] Michael R. Maniaci and Ronald D. Rogge. 2014. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality* 48 (2014), 61 – 83. https://doi.org/10.1016/j.jrp.2013.09.008

[90] Rod A. Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality* 37, 1 (2003), 48 – 75. https://doi.org/10.1016/S0092-6566(02)00534-2

[91] Roger K Moore. 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots.* Springer, 281–291.

[92] Nick Neave, Rachel Jackson, Tamsin Saxton, and Johannes Hönekopp. 2015. The influence of anthropomorphic tendencies on human hoarding behaviours. *Personality and Individual Differences* 72 (2015), 214–219. https://doi.org/10.1016/j.paid.2014.08.041

[93] Raymond S. Nickerson. 1999. How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin* 125, 6 (1999), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

[94] Donald Norman. 1983. Some Observations on Mental Models. In *Mental Models* (1st ed.). Psychology Press, 7–15.

[95] Donald A. Norman. 2013. *The design of everyday things* (revised and expanded edition ed.). Basic Books.

[96] Sally Olderbak and Oliver Wilhelm. 2017. Emotion perception and empathy: An individual differences test of relations. *Emotion* 17, 7 (2017), 1092.

[97] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. 1998. Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and Speech* 41, 3 (1998), 419–442. https://doi.org/10.1177/002383099804100409

[98] Arun Parasuraman, Leonard L Berry, and Valarie A Zeithaml. 1991. Refinement and reassessment of the SERVQUAL scale. *Journal of retailing* 67, 4 (1991), 420.

[99] Jamie Pearson, Jiang Hu, Holly P Branigan, Martin J Pickering, and Clifford I Nass. 2006. Adaptive Language Behavior in HCI: How Expectations and Beliefs about a System Affect Users' Word Choice. (2006), 4.

[100] Jan Hyld Pejtersen, Tage Søndergård Kristensen, Vilhelm Borg, and Jakob Bue Bjorner. 2010. The second version of the Copenhagen Psychosocial Questionnaire. *Scandinavian journal of public health* 38, 3_suppl (2010), 8–24.

[101] John L. Perry, Peter J. Clough, Lee Crust, Keith Earle, and Adam R. Nicholls. 2013. Factorial validity of the Mental Toughness Questionnaire-48. *Personality and Individual Differences* 54, 5 (2013), 587 – 592. https://doi.org/10.1016/j.paid.2012.11.020

[102] Christopher Peterson, Amy Semmel, Carl Von Baeyer, Lyn Y Abramson, Gerald I Metalsky, and Martin EP Seligman. 1982. The attributional style questionnaire. *Cognitive therapy and research* 6, 3 (1982), 287–299.

[103] Paul R. Pintrich and Elisabeth V. De Groot. 1990. Motivational and Self-Regulated Learning Components of Classroom Academic Performance. *Journal of Educational Psychology* 82, 1 (1990), 33–40.

[104] Melanie D Polkosky. 2005. Toward a social-cognitive psychology of speech technology: Affective responses to speech-based eservice. (2005). https://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1818&context=etd

[105] Melanie D Polkosky and James R Lewis. 2003. Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology* 6, 2 (2003), 161–182.

[106] Pernilla Qvarfordt, Arne Jönsson, and Nils Dahlbäck. 2003. The role of spoken feedback in experiencing multimodal interfaces as human-like. In *Proceedings of the 5th international conference on Multimodal interfaces - ICMI '03* (Vancouver, British Columbia, Canada). ACM Press, 250. https://doi.org/10.1145/958432.958478

[107] R Development Core Team. 2009. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org ISBN 3-900051-07-0.

[108] Renate LEP Reniers, Rhiannon Corcoran, Richard Drake, Nick M Shryane, and Birgit A Völlm. 2011. The QCAE: A questionnaire of cognitive and affective empathy. *Journal of personality assessment* 93, 1 (2011), 84–95.

[109] William Revelle. 2020. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psych R package version 2.0.8.

[110] Carolien Rieffe, Lizet Ketelaar, and Carin H Wiefferink. 2010. Assessing empathy in young children: Construction and validation of an Empathy Questionnaire (EmQue). *Personality and individual differences* 49, 5 (2010), 362–367.

[111] GC Roberts and G Balagué. 1991. The development and validation of the Perception of Success Questionnaire. In *FEPSAC Congress, Cologne, Germany*.

[112] Peter AM Ruijten, Antal Haans, Jaap Ham, and Cees JH Midden. 2019. Perceived human-likeness of social robots: testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics* 11, 3 (2019), 477–494.

[113] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323. https://doi.org/10.1007/s12369-013-0196-9

[114] Irwin G. Sarason, Henry M. Levine, Robert B. Basham, and Barbara R. Sarason. 1983. Assessing social support: The Social Support Questionnaire. *Journal of Personality and Social Psychology* 44, 1 (1983), 127–139. https://doi.org/10.1037/0022-3514.44.1.127

[115] Sau-lai Lee, Ivy Yee-man Lau, S. Kiesler, and Chi-Yue Chiu. 2005. Human Mental Models of Humanoid Robots. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation* (Barcelona, Spain). IEEE, 2767–2772. https://doi.org/10.1109/ROBOT.2005.1570532

[116] Elisabeth Schaffalitzky, Sinead NiMhurchadha, Pamela Gallagher, Susan Hofkamp, Malcolm MacLachlan, and Stephen T. Wegener. 2009. Identifying the Values and Preferences of Prosthetic Users: A Case Study Series Using the Repertory Grid Technique. *Prosthetics and Orthotics International* 33, 2 (2009), 157–166. https://doi.org/10.1080/03093640902855571

[117] Gregory Schraw and Rayne Sperling Dennison. 1994. Assessing metacognitive awareness. *Contemporary educational psychology* 19, 4 (1994), 460–475.

[118] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *IJIMAI* 4, 6 (2017), 103–108.

[119] Mildred L.G. Shaw and Laurie F. Thomas. 1978. FOCUS on education—an interactive computer system for the development and analysis of repertory grids. *International Journal of Man-Machine Studies* 10, 2 (1978), 139–173. https://doi.org/10.1016/S0020-7373(78)80009-1

[120] Virginia Slaughter, Michelle J Dennis, and Michelle Pritchard. 2002. Theory of mind and peer acceptance in preschool children. *British journal of developmental psychology* 20, 4 (2002), 545–564.

[121] Michael A. Smyer, Brian F. Hofland, and Edward A. Jonas. 1979. Validity Study of the Short Portable Mental Status Questionnaire for the Elderly*. *Journal of the American Geriatrics Society* 27, 6 (1979), 263–269. https://doi.org/10.1111/j.1532-5415.1979.tb06128.x

[122] Brendan Spillane, Emer Gilmartin, Christian Saam, and Vincent Wade. 2019. Issues Relating to Trust in Care Agents for the Elderly. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 20, 3 pages. https://doi.org/10.1145/3342775.3342808

[123] 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. *Journal of Personality Assessment* 91, 1 (2009), 62–71. https://doi.org/10.1080/00223890802484381

[124] Michael F Steger, Patricia Frazier, Shigehiro Oishi, and Matthew Kaler. 2006. The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of counseling psychology* 53, 1 (2006), 80.

[125] Anita Tobar-Henríquez, Hugh Rabagliati, and Holly P. Branigan. 2020. Lexical entrainment reflects a stable individual trait: Implications for individual differences in language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46, 6 (2020), 1091–1105. https://doi.org/10.1037/xlm0000774

[126] Lai Lai Tung, Yun Xu, and Felix B. Tan. 2009. Attributes of Web Site Usability: A Study of Web Users with the Repertory Grid Technique. *International Journal of Electronic Commerce* 13, 4 (2009), 97–126. https://doi.org/10.2753/JEC1086-4415130405

[127] Larry Vandergrift, Christine CM Goh, Catherine J Mareschal, and Marzieh H Tafaghodtari. 2006. The metacognitive awareness listening questionnaire: Development and validation. *Language learning* 56, 3 (2006), 431–462.

[128] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a Personality Model for Speech-based Conversational Agents Using the Psycholexical Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA). ACM, 1–14. https://doi.org/10.1145/3313831.3376210

[129] Marilyn A. Walker, Jeanne Fromer, Giuseppe Di Fabbrizio, Craig Mestel, and Don Hindle. 1998. What can I say?: evaluating a spoken language interface to Email. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '98* (Los Angeles, California, United States). ACM Press, 582–589. https://doi.org/10.1145/274644.274722

[130] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer speech and language* 12, 4 (1998), 317–348.

[131] Adam Waytz, John Cacioppo, and Nicholas Epley. 2010. Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science* 5, 3 (2010), 219–232. https://doi.org/10.1177/1745691610369336

[132] Lynn Westbrook. 2006. Mental models: a theoretical overview and preliminary study. *Journal of Information Science* 32, 6 (2006), 563–579. https://doi.org/10.1177/0165551506068134

[133] Sally Wheelwright, Simon Baron-Cohen, Nigel Goldenfeld, Joe Delaney, Debra Fine, Richard Smith, Leonora Weil, and Akio Wakabayashi. 2006. Predicting autism spectrum quotient (AQ) from the systemizing quotient-revised (SQ-R) and empathy quotient (EQ). *Brain research* 1079, 1 (2006), 47–56.

[134] Carsten Zoll and Sibylle Enz. 2010. A questionnaire to assess affective and cognitive empathy in children. (2010).