Studying Mind Perception in Social Robotics Implicitly: The Need for Validation and Norming

Tuğçe Nur Pekçetin
Department of Cognitive Science
Middle East Technical University
Ankara, Turkey
tugce.bozkurt@metu.edu.tr

Badel Barinal
Bilkent University
Ankara, Turkey
badelbarinal@gmail.com

Jana Tunç
Department of Psychology
Middle East Technical University
Ankara, Turkey
bjanatunc@gmail.com

Cengiz Acarturk
Department of Cognitive Science
Jagiellonian University
Krakow, Poland
cengiz.acarturk@uj.edu.pl

Burcu A. Urgen
Department of Psychology
Neuroscience Program
Bilkent University, UMRAM
Ankara, Turkey
burcu.urgen@bilkent.edu.tr

ABSTRACT

The recent shift towards incorporating implicit measurements into the mind perception studies in social robotics has come along with its promises and challenges. The implicit tasks can go beyond the limited scope of the explicit tasks and increase the robustness of empirical investigations in human-robot interaction (HRI). However, designing valid and reliable implicit tasks requires norming and validating all stimuli to ensure no confounding factors interfere with the experimental manipulations. We conducted a lexical norming study to systematically explore the concepts suitable for an implicit task that measures mind perception induced by social robots. Two-hundred seventy-four participants rated an expanded and strictly selected list of forty mental capacities in two categories: Agency and Experience, and in two levels of capacities: High and Low. We used the partitioning around medoids algorithm as an objective way of revealing the clusters. We discussed the different clustering solutions in light of the previous findings. We consulted on frequency-based natural language processing (NLP) on the answers to the open-ended questions. The NLP analyses verified the significance of clear instructions and the presence of some common conceptualizations across dimensions. We proposed a systematic approach that encourages validation and norming studies, which will further improve the reliability and reproducibility of HRI studies.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

social robotics, human-robot interaction, mind perception, implicit association test, norming

HRI '23, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9964-7/23/03.
https://doi.org/10.1145/3568162.3577001

ACM Reference Format:
Tuğçe Nur Pekçetin, Badel Barinal, Jana Tunç, Cengiz Acarturk, and Burcu A. Urgen. 2023. Studying Mind Perception in Social Robotics Implicitly: The Need for Validation and Norming. In *Proceedings of the 2023 ACM/IEEE Int'l Conference on Human-Robot Interaction (HRI '23), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3568162.3577001

1 INTRODUCTION

Current research on Human-Robot Interaction (HRI) needs development on several fronts to grow as an independent and coherent research domain. Aside from abundant results reported in experimental studies, a major research gap is the scope and validity of the empirical findings. The scope of empirical research in HRI is enriched not only by the wide variety of robot features but also by ramifications in human cognition and behavior. The evident complexity of the interaction indicates the need for systematic approaches to developing empirical work in HRI. In the present study, we focus on a crucial aspect of this need, validation and norming, in the context of human perception of social robots. More specifically, we report a lexical study that norms and validates the concepts to be used in tasks that measure mind perception explicitly and implicitly.

1.1 Human Perception of Robots

One of the most significant questions for human-robot interaction is how people perceive robots compared to other agents and entities. The answers to this question are crucial since the way social robots are perceived modulates how they are accepted as new members of society, thus, influencing the future roles they will take. Investigating the human perception of robots can also yield insights regarding the design and development of social robots. Last but not least, the way people perceive the capabilities of the robots or the extent they attribute capacities to them is likely to affect how they would predict their behavior, hence how people would interact with them [38].

Despite being inanimate, robots may possess certain agentic and perceptible abilities depending on their design and affordances, rendering them suitable candidates for social interaction. In social interaction, an essential ability is the capacity to assume the presence of others' mental states and to recognize that others' mental states may differ from ours. This recognition has been known as the Theory of Mind (ToM) [29]. A set of well-established ToM tasks has been employed to assess whether natural cognitive agents, such as babies [1] and animals [29], have the capacity to impute mental states to others. The previous research on ToM has also investigated the other side of the coin, asking how humans ascribe mental states to other agents, such as babies, God, animals, and social robots [10, 22, 42]. Other minds are evidently perceived in the eye of the beholder [7], reflecting individual traits and preferences of the perceiver and the perceived. Nevertheless, a systematic investigation of the role of ToM in HRI requires an extensive investigation of the likely causes and consequences of mind perception from the perspective of both sides [41].

1.2 Mind Perception Research in Social Robotics

What are the elements of mind perception in social robotics? In a seminal work, Gray et al. [10] proposed that mind perception has two dimensions: *Agency* – the ability to do, and *Experience* – the ability to feel. The evidence came from the principal component analysis (PCA) on experimental findings, which demonstrated that people attribute low or medium levels of Agency and Experience to robots [10]. On the other hand, several other studies built on Gray et al. and proposed new conceptualizations of mind perception. For instance, Malle [22] pointed out the grey areas between Agency and Experience dimensions and suggested three new dimensions: *Affect*, *Moral* and *Mental Regulation*, and *Reality Interaction*, some of which could be further divided into sub-dimensions. Similarly, Weisman et al. identified at least three dimensions, *Body*, *Mind*, and *Heart-related capacities* [42], each of which relates to certain aspects of the Agency and Experience dimensions of Gray et al. [10]. Although these studies have been instrumental in developing systematic research on the elements of mind perception, further research is needed to reconcile these findings and provide valid assessments of mind perception.

One of the potential reasons for the discrepancy in the conceptualization of mind perception between different studies is the use of explicit measures such as self-reports. Human judgments, expressed in self-reports, usually provide valid assessments of certain aspects of social interaction. Nevertheless, their scope is limited, especially in predicting behavior for various reasons [9, 27]. For instance, the level of motivation, awareness, or expression ability of the participants in their introspective assessments and the presence of interruption may have a significant impact, influencing the results of experimental studies that rely on those measures [27]. Furthermore, self-reports usually fall short in providing a mechanistic understanding of human cognition and behavior.

Numerous tasks have been developed in line with Greenwald and Banaji's call [13] for increasing the use of implicit measurements in social behavior research as an alternative to explicit measures. We propose that a similar shift from explicit measures, such as self-reported judgments, to implicit measures would improve the robustness of empirical findings in HRI research in general and mind perception research in particular.

Although most HRI literature has been based on explicit measurements [2, 26, 31], several studies used implicit measurements ranging from semantic priming [36] to implicit association tasks (IAT) [21, 35]. Recently, Li et al. developed an IAT task, Mind-Perception IAT (MP-IAT), as an initiative to extend the implicit measurements of mind perception to social robotics research [21]. Their approach stands out as they suggest addressing the *High* and *Low* ends of the Agency and Experience dimensions by choosing related items as stimuli. They also reported the results of both implicit and explicit measurements, which could help evaluate whether the automatic and controlled mental processes were interacting or functioning separately. A strength of their study is that they selected some of the original concepts associated with the Agency and Experience dimensions in the literature. However, we consider that their verbal stimuli selection could be improved since they 1. revised the definitions of the Agency and Experience only according to their study's needs, 2. chose the concepts from narcissists' self-perception and gender stereotypes studies rather than mind perception literature, 3. created the antonyms for the concepts themselves 4. did not mention a normative process for the first three steps. They also acknowledged the lack of methods to validate and optimize their stimuli as a limitation. We believe using normative tests would improve the generalizability of the findings. Given that the focus of their study was the development of the task rather than the stimuli, here, we emphasize the other side of the coin and draw attention to the need for using concepts identified by normative analyses in implicit measurements.

1.3 The Motivation of the Study

Developing an implicit association task to assess mind perception is challenging due to the lack of consensus regarding which concepts represent which aspect of mind perception. Such an ambiguity calls for validation and norming of the concepts used in such tasks. In social sciences and psychology research, it is common to norm the stimuli materials, such as images and vignettes. However, norming is usually ignored in designing verbal stimuli, including concepts.

Norming is a practice in which researchers reach out to a representative sample and assess the individuals to construct norms about the concepts, visual items, or any materials in question [3, 33]. Having norm-referenced evaluations is crucial since using normed stimuli would reduce the influence of confounding factors, ensuring that the outcomes of the study are due to the experimental manipulations [34]. Norming is a vital practice for the materials' reproducibility, hence improving the studies' replicability [3].

In the present study, we report a lexical norming study that provides a basis for any implicit task as well as an explicit one aiming to measure mind perception. We explored which concepts would be reliably used to represent the Agency and Experience dimensions along with the High and Low ends as introduced by the implicit association task by Li et al. [21].

In our methodology, we followed best practice guidelines suggested by Greenwald et al. [14]: 1. familiarize the participants with the concepts using clear definitions and instructions, 2. choose concepts that are easy to sort quickly to avoid confounds, 3. perform pilot testing before IAT using the intended category classification 4. use an objective method to cluster the concepts into the intended categories 5. conduct complementary analyses to reveal the evaluation processes of the participants as a manipulation check.

2 METHODOLOGY

2.1 Participants

The number of participants who completed the study thoroughly was 280. The data of six participants were excluded before further analyses since they were monolingual in a language other than Turkish, which is the language of the study. The remaining sample size was 274 (163 participants defined themselves as females, 109 as males, and two as non-binary). Turkish was the only native language of the 263 participants, while it was one of the native languages of the eleven participants. The sample's age range was between 18 and 65 (M = 33.86, SD = 12.17).

Before starting the study, all participants read and digitally signed the informed consent form approved by the Human Research Ethics Committee of Bilkent University. The structure of our study was based on participants' judgments on two options rather than pre-defined conditions and comparisons, so a typical power analysis would not be suitable. We aimed to reach as many participants as possible by consulting on the sample size of similar norming studies, which roughly changed between 50-200+ [4, 34].

The participants were volunteers who responded to the participation call for an online study. The call was distributed as a flyer with a short description of the study and a QR code directing to the URL of the direct study link. There were 463 attempts for the survey, of which 280 were complete submissions. Two hundred seventy-four participants used the anonymous link, while the remaining six used the QR code to reach the study.

The participants were from various locations and occupational and educational backgrounds. Nearly half of them (49.64%) held a bachelor's degree, 21.90% held a master's degree, and a similar number of participants (19.34%) were high school graduates. Among the remaining participants, 14 held a Ph.D. degree, nine were graduates of an associate degree, one was a middle school, and one was a primary school graduate, comprising 9.12% of the total sample.

2.2 Item Selection

The implicit association task measures the differential association of two target concepts with an attribute [13]. In this task, the aim is to measure the participants' implicit attitudes towards targets based on the comparison of the reaction times spent during each target's association with the attributes.

To decide the concepts we would include in our conceptual norming study, we started choosing our target concepts as Agency and Experience, the supposedly two prominent dimensions of mind perception suggested by Gray et al. [10]. We used these two as umbrella terms and created a large pool of possible attribute concepts under each. Our item pool creation steps are as follows:

We consulted the verbal stimuli of the studies investigating the dimensions of mind perception [10, 22, 42]. Then, we extended our pool with the stimuli used in studies on the mind perception of robots and mind attributions to robots [5, 19, 21, 23, 24, 37, 39], and acceptance of robots [17]. We also reviewed the materials of several studies on the mind perception induced by other agents [11, 12, 16, 20, 32]. We benefited from the verbal stimuli selections of some studies on anthropomorphism [6, 40] and emotion recognition [35, 44].

There were 80 candidate concepts in our initial pool. We tentatively classified these concepts into the High and Low ends of the Agency and Experience dimensions based on the definitions of the two main categories and the two ends. We defined the Low dimension by referring to concepts representing relatively primitive acts (e.g., a skill that could be achieved even by non-human animals). In contrast, we defined the High dimension by referring to concepts representing relatively higher order and complex acts (e.g., a skill only humans could achieve). The tentative division of the initial pool was: 26 concepts for High Agency and 22 for Low Agency, while 17 for High Experience and 15 for Low Experience.

Since we aimed to norm the concepts in the native languages of the participants, we translated the initial pool of English concepts. Seven native speakers with backgrounds in cognitive sciences with C1 English level took part in the translation process. Two translators had a Linguistics background with professional localization experience of three to six years. We normed the concepts in the native language to avoid misconceptions while conducting such a high-level conceptual study and to include participants from various social backgrounds and age groups, as suggested as one of the best practices [14].

To have a consistent presentation and not bias the participants toward any category, we translated the concepts in their infinitive form since the concepts of Agency and Experience are defined as 'the ability to do' and 'the ability to feel' [10].

We excluded concepts such as 'personality,' 'consciousness,' and 'thought,' which loaded high in both dimensions in Gray et al.'s study [10]. We also excluded the synonyms or concepts with similar meanings to avoid repetition in the stimulus. For example, the Turkish translations of 'communicate' and 'interact' nearly overlap, so we excluded the term 'interact' since 'communicate' has a more common usage. We also excluded terms based on the length of their translated versions. Some concepts, such as 'deliberate,' can be expressed with five words in Turkish, and such a length would not be suitable for an implicit task. We thus excluded the terms whose translated versions were longer than three words. Another exclusion criterion we applied was about the nature of the concepts: we excluded the concepts which could potentially elicit discussions (e.g., 'deserve punishment') and which have philosophical implications (e.g., 'have values') since such concepts can be hard to internalize and assess without a contextual background, as it is in the case of an implicit task.

In the end, we proceeded with concepts with the highest frequency of usage in the literature mentioned above and with the ones that were statistically found to be most strongly associated with Agency or Experience dimensions rather than those that might be in the grey area [10, 22]. We tentatively assigned the remaining concepts to the High and Low ends of the Agency and Experience dimensions. Our final list included 20 concepts for Agency: half of which is in the Low dimension (Low Agency), the other half in the High dimension (High Agency), and 20 concepts for Experience, half of which is in the Low dimension (Low Experience), and the other half in the High dimension (High Experience), comprising 40 items in total.

2.3 Tasks and Procedure

The study was prepared on the Qualtrics XM platform, a software for creating and conducting online experiments as well as collecting and storing data (https://www.qualtrics.com/). The participants who clicked on the study link on their own devices started to evaluate the lexical items after some steps. They first read a short introductory paragraph about the study. When they confirmed that they would voluntarily participate in the study, they read and digitally signed the informed consent form. Then, for the demographic questions part, the participants provided their initials and birth dates so that we could verify that there were no duplicate submissions from the same people. After the participants answered the questions about their gender identity, native language(s), educational status, and education and work field(s), they came across the instructions page.

In the instructions, the two Turkish concepts used to refer to the concepts 'Agency' and 'Experience' were introduced and explained with their own examples in the target language. Agency was defined as 'the ability to do something using cognitive abilities and processes' while 'Experience' was described as 'the ability to sense sensory, emotional, or physiological states.' Then, the definitions of the concepts referring to 'High and Low Agency' and 'High and Low Experience' were presented, and the concepts were further explained with their respective examples. The concept of 'planning' was used to describe the High and Low ends of the Agency capacity. The 'long-term planning for the unseen future' was used to exemplify a concept that requires complex cognitive processes; thus, High Agency capacity, while 'the instant, simple or instinctive plans for needs such as nutrition and shelter' were used to explain a concept which represents the low end of the agency dimension (Low Agency). 'Mental fatigue, i.e., the feeling of burnout' was provided as an example of a concept that requires High Experience capacity, while 'physical tiredness' was used to exemplify a concept that requires Low Experience capacity. The example concepts were selected to have both high and low-capacity versions that would help highlight the difference between the High and Low ends of the recently introduced and relatively unfamiliar concepts of Agency and Experience. We wanted to ensure that the participants would not confuse Agency and Experience and that they could make their judgments about the scale of these capacities based on the instructions.

The participants were instructed that they would see a concept in the middle of the screen and evaluate this concept in two steps. There were 40 concepts in total, and the participants evaluated each concept in its own section; when they finished one concept, they could continue with the next. Every section included questions for each evaluation step: the first question was: "In which category do you evaluate this concept?" The participant would choose Agency or Experience by clicking on the left or right buttons. The following question was: "What degree of capacity does this concept require in the category you chose?" The participant would choose High or Low by clicking the left or right buttons. The concepts were presented in a randomized order to avoid the possible fatigue effect towards the end of the survey.

At the end of the study, there was an open-ended part that included four questions about the participants' thought processes. The structures of the questions were the same as "What did you consider when deciding that a concept requires High/Low capacity of the Agency/Experience ability?" The only parts changing were the High/Low and Agency/Experience parts. These questions were used for two reasons: first, they would allow checking whether the participants could evaluate the concepts based on the definitions provided, and second, we could learn more about the reasoning process behind the assignment of the concepts used in mind perception.

3 RESULTS

3.1 Clustering Analyses

We calculated the total capacity and capacity scaling ratings for each concept. Each concept has a rating score out of 274, either as 'Agency' or 'Experience,' representing this concept's category, and 'High' or 'Low,' representing the scale of the capacity necessary for this concept.

3.1.1 Two dimensions. The design of the present study calls for a division of the total items into four; however, when we applied the silhouette method [30] to the data points in two dimensions which represent: 1. the concept category (in this dimension, we used the Agency ratings since the Experience ratings were complements of them) and 2. the scale capacity (in this dimension, we used the High ratings since the Low ratings were numerical complements of them), it suggested clustering the data into two (Figure 1).

We then applied the PAM (Partitioning Around Medoids) algorithm [18], which can be considered a k-means clustering algorithm variant. In the k-means algorithm, it is possible to reach a "local" optima. The process starts with k many clusters, each consisting of a single random point, and continues by adding each new point to the cluster with the nearest mean to this new point to reduce the within-cluster sum of squares [15]. Although k-means is suitable for creating a small number of clusters from a large number of observations, we preferred PAM for our case since it is a robust procedure that determines a representative object called "medoid" for each cluster.

In Table 1, the number of ratings that each concept elicited in the four dimensions is provided according to the PAM clustering with the requested cluster number of two. In the **Item** columns, the

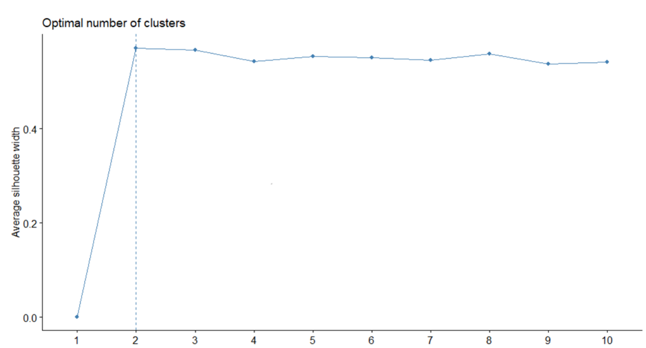

Figure 1: Silhouette method offering the optimal number of clusters as two.

Table 1: The concepts are presented with their ratings in four dimensions and their clusters according to the first PAM results (*k*=2). The items with an (*) indicate the medoids (the representative item of a cluster).

Item (to ...)	Cluster	Agency	Experience	High	Low	Item (to ...)	Cluster	Agency	Experience	High	Low
imitate	1	262	12	209	65	feel upset	2	14	260	91	183
learn	1	260	14	220	54	be afraid	2	18	256	66	208
have purposes	1	257	17	241	33	feel shame	2	23	251	151	123
express	1	257	17	239	35	feel rage	2	23	251	79	195
reason	1	252	22	257	17	feel nervous	2	25	249	122	152
prefer	1	250	24	216	58	love	2	27	247	178	96
communicate	1	247	27	209	65	feel hunger	2	31	243	61	213
be deductive	1	245	29	245	29	be in physical pain*	2	32	242	74	200
decide freely	1	240	34	249	25	feel stress	2	33	241	122	152
exercise self-control*	1	239	35	245	29	feel joy	2	39	235	129	145
compete	1	230	44	195	79	feel pride	2	46	228	202	72
tell right from wrong	1	226	48	249	25	experience pleasure	2	48	226	146	128
seek survival	1	225	49	152	122	feel thirsty	2	56	218	58	216
understand	1	214	60	238	36	taste things	2	67	207	63	211
make predictions	1	206	68	225	49	dream	2	82	192	76	198
regulate emotions	1	195	79	264	10	need sleep	2	88	186	69	205
remember	1	191	83	170	104	smell	2	98	176	59	215
imagine	1	160	114	213	61	desire things	2	118	156	94	180
recognize emotion	1	106	168	251	23	hear	2	136	138	65	209
have empathy for others	1	113	161	253	21	see	2	178	96	73	201

Table 2: The concepts are presented with their clusters according to the different PAM results (*k*=3,4,5) and the number of ratings in the capacity and the capacity scaling dimensions. The numbered superscripts indicate the medoid for that number of clusters. Different clusters and dimensions are highlighted with distinct colors.

Item (to ...)	PAM Clusterings *k*=3	PAM Clusterings *k*=4	PAM Clusterings *k*=5	Above 50% Assignments Category	Above 50% Assignments Scale	Item (to ...)	PAM Clusterings *k*=3	PAM Clusterings *k*=4	PAM Clusterings *k*=5	Above 50% Assignments Category	Above 50% Assignments Scale
imitate	1	1	1	Agency	High	feel upset	2	2	5	Experience	Low
learn	1	1	1	Agency	High	be afraid	2	2	5	Experience	Low
have purposes	1	1	1	Agency	High	feel shame	2	2	2	Experience	High
express	1	1	1	Agency	High	feel rage[5]	2	2	5	Experience	Low
reason	1	1	1	Agency	High	feel nervous	2	2	2	Experience	Low
prefer	1	1	1	Agency	High	love	2	2	2	Experience	High
communicate	1	1	1	Agency	High	feel hunger	3	3	5	Experience	Low
be deductive	1	1	1	Agency	High	be in physical pain[2]	2	2	5	Experience	Low
decide freely	1	1	1	Agency	High	feel stress[3,4]	2	2	2	Experience	Low
exercise self-control[2,3,4,5]	1	1	1	Agency	High	feel joy[5]	2	2	2	Experience	Low
compete	1	1	1	Agency	High	feel pride	2	2	2	Experience	High
tell right from wrong	1	1	1	Agency	High	experience pleasure	2	2	2	Experience	High
seek survival	1	1	1	Agency	Low	feel thirsty	3	3	3	Experience	Low
understand	1	1	1	Agency	High	taste things	3	3	3	Experience	Low
make predictions	1	1	1	Agency	High	dream	3	3	3	Experience	Low
regulate emotions	1	1	1	Agency	High	need sleep[3,4,5]	3	3	3	Experience	Low
remember	1	1	1	Agency	Low	smell	3	3	3	Experience	Low
imagine	1	4	4	Agency	High	desire things	3	3	3	Experience	Low
recognize emotion	1	4	4	Experience	High	hear	3	3	3	Experience	Low
have empathy for others[4,5]	1	4	4	Experience	High	see	3	3	3	Agency	Low

item names are presented, **Cluster** column presents the cluster that each item placed according to PAM analyses. The **Agency** column presents the number of ratings each concept received, while the **Experience** column presents the number of ratings towards the Experience dimension. The **High** and **Low** columns present the number of capacity scaling ratings. For instance, the item 'imitate' received 262 ratings favoring the Agency dimension and 209 favoring the High option, meaning that most participants considered this concept as belonging to the Agency category and requiring High capacity.

The results of the first PAM analyses suggested a clustering dividing the whole stimuli into two sets with an equal number of members. Also, the members of clusters were almost divided into concepts belonging to the Agency and Experience categories. In the first cluster, all the items except for the 'recognize emotion' and 'have empathy for others' had the Agency rating of more than half of the participants, which are highlighted in Table 1. Similarly, in the second cluster, all the items except for the item 'see' had the Experience rating from more than half of the participants. The medoids of the clusters in the Agency dimension were 239 and 32, and they were 245 and 74 for the High dimension, pointing out the 'exercise self-control' and 'be in physical pain' as the representatives of the Agency and Experience clusters, respectively.

3.1.2 More than two dimensions. As stated earlier, our design of the stimuli set implies four clusters; thus, we applied the PAM algorithm three more times with different k values decided beforehand. Table 2 presents the PAM method's clustering results with different ks. We also presented the possible categories if the judgments were based solely on the number of ratings towards the two categories and the scales.

The solution with three clusters suggested that the initial division towards Agency and Experience remains the same, but a new cluster with the concepts such as 'feel hunger,' 'need sleep,' and 'desire things', separated from the concepts like 'feel joy,' 'feel pride,' and 'feel nervous' in the Experience dimension. The concepts 'feel stress' and 'need sleep' appeared to be the medoids of those clusters. In the solution with four clusters, we had the same structure within the Experience dimension while a new cluster emerged within the Agency dimension with the concepts 'imagine,' 'recognize emotion,' and 'have empathy for others' and with the last one as the medoid of the new cluster. We also checked for the five-cluster solution and observed that the Experience dimension was further divided into three clusters, separating concepts such as 'feel hunger' and 'be in physical pain' from the concepts such as 'feel love' and 'feel shame'. The medoids of the related clusters were 'feel rage' and 'feel joy,' respectively.

Table 3: The summary table for the open-ended questions.

Category	Scale	# of Valid Answers	Word Count	Character Count
Agency	High	254	1934	15333
Agency	Low	251	1812	14320
Experience	High	252	1963	15834
Experience	Low	251	1702	13505

Another way of clustering this data set could be conducted based on the number of ratings. It is possible to complete the item selection according to the majority of the ratings. For instance, from the Experience dimension, there are six alternatives to be selected as High Experience items, while there are even more (n=15) candidates to be selected as target concepts representing the Low end of this dimension. In the Agency cluster, although 19 concepts were rated as Agency by most participants, all those ratings were towards the High level except for the concept 'see'. However, this would not cause a problem since it could be possible to select from the bottom end of the High Agency group, such as the items 'seek survival', 'remember,' and 'see.' Even the item 'compete' could be modified to 'intra-species competition', which would be suitable to represent the Low Agency capacity.

3.2 Language Analyses

We separately compiled the answers to the four open-ended questions at the end of the survey. The forms of the questions were the same across the conditions: "What did you consider when deciding that a concept requires High/Low capacity of the Agency/Experience ability?". Among the 274 participants, 257 answered these last four questions. The characteristics of the valid answers, i.e., the remaining ones after omitting the filler answers that included only one letter or random letters, are summarized in Table 3.

We used the term frequency-inverse document frequency (TF-IDF) model to investigate whether there are any most preferred words or phrases. The TF-IDF analysis was conducted using the Tfidfvectorizer class from the scikit-learn library [28]. We aimed to explore the attribution process of the participants when they were asked to evaluate a category as Agency or Experience and decide the level of mental capacity necessary for that concept. After exploring the most frequent unigrams, bigrams, trigrams, and quadgrams, we used the first fifty quadgrams to extract the possible topics since in quadgrams, it was possible to infer the primary motivation or the driving idea.

To decide the prevailing concepts in the evaluation processes of the participants, two coders blind to the purposes of the study reviewed the first fifty quadgrams for the four dimensions and came up with their own set of topics. This step was necessary to derive the thematic categories from the data since we did not have pre-determined codes for this open-ended exploratory part. Then, they conducted a session in which they shared their topic assignments and discussed the finalized versions. In the end, they agreed on three prevailing topics across four dimensions. They then coded the quadgrams; they individually assigned each quadgram to one of the topics they had pre-determined in the previous step. Finally, they compared their ratings: they gave one point for the items they agreed on and zero points for the others.

Cohen's κ was run to determine the inter-rater agreement on the topic codes. There was an almost perfect agreement [25] on the 50 items for each dimension and on the 200 items when it was checked overall. The results for the High Agency κ = .972 (95% CI, .917 to 1.000), Low Agency κ = .914 (95 % CI, .820 to 1.000), High Experience κ = .946 (95 % CI, .873 to 1.000), Low Experience κ = .945 (95 % CI, .871 to 1.000) and for the overall agreement κ = .945. (95% CI, .908 to .983) (p < .001 for all conditions and the total agreement).

Table 4 below presents the main topics or the umbrella terms on which the two raters agreed to be the leading ones from the first fifty quadgrams: 1. properties of the concept and the process, 2. type of the agent, and 3. cognitive and physical requirements for the concept and agent. Under each main topic, we presented the most frequently occurred expressions as the representative items. Table 4 summarizes the aspects the participants considered while deciding which capacity and how much capacity would be required for a concept in the study.

4 DISCUSSION

There is a growing body of research on mind perception in social robotics since the seminal study by Gray et al., which construes mind perception with two dimensions, Agency and Experience [10]. However, there is little consensus in the field about which concepts characterize these dimensions [10, 22, 42]. Furthermore, because of their inherent nature, self-reports may fall short in predicting and providing a mechanistic understanding of social behavior [9, 27]. Therefore, there is a need to develop tasks and methods to measure mind perception implicitly.

The present study aimed to norm and validate concepts associated with Agency and Experience dimensions of mind perception, which could, in turn, be used in implicit tasks reliably. To this end, we first pooled together a broad set of concepts used in the mind perception literature. We then defined clear exclusion criteria to narrow down this set, which included eliminating concepts found to stay in the grey area between Agency and Experience in different studies [10, 22]. Next, we aimed to sub-categorize the Agency and Experience dimensions into High and Low dimensions following the work of Li et al. [21] since these former dimensions cover a broad range of concepts within themselves. We thoroughly defined each of the four dimensions and then conducted an online norming study in which the participants were instructed to categorize given concepts into these dimensions. We used an objective method (PAM algorithm) to reveal the clusters of concepts for each dimension as an alternative to the widely used PCAs in the literature. We also complemented the norming study with a simplistic, frequency-based NLP analysis of the open-ended questions to understand participants' criteria in categorizing the concepts.

4.1 Theoretical and Practical Contributions

Our results demonstrate the significance of norming and validation in the study of mind perception. The PAM results with two clusters were almost perfectly in line with our a priori Agency and Experience categorization in the item selection stage, which validated the selection criteria we followed while determining our set of stimuli. However, in our study, we aimed to divide these two main categories Agency and Experience, into further clusters according to the capacity levels they would require, namely to the High and Low ends of them, which called for a solution with four clusters. So, we increased the cluster number gradually to see which concepts would be in the same group when we have four clusters and whether we have similar distributions with the previous studies.

Although our item pool consisted of some common concepts with the items tested previously [8, 16, 22, 42], a direct comparison of the results would not be possible since their methodologies and the item pools were different. However, the distributions of the concepts across clusters can also be interpreted with their terminologies. While the Agency cluster from the first PAM analysis with k=2 stayed the same with k=3, the Experience cluster was divided into two, creating Clusters 2 and 3. In these two new clusters, the cluster of the *Physiological capacities* (Cluster 3) sharply separated from the cluster with the capacities belonging to the *Affective* category (Cluster 2) in Malle's [22] terminology. Weisman et al. [42] would name these clusters as *Body and Heart*, respectively, while D'Andrade [8] and Haslam et al. [16] would label such a separation as *Perceptions* vs. *Feelings/Emotions*, respectively.

Although the Agency cluster stayed as a whole, bearing the *Agentic* categories such as *Thoughts and Intentions* [8], *Cognitive or Agentic* [22], or *Mind* [42], this structure changed when a solution with k=4 was applied, eliciting Cluster 4 with the concepts which would require the theory of mind related capacities to some extent, which Malle [22] would label under *Mental Regulations*, Weisman et al. [42] under *Heart and Mind*, and D'Andrade [8] under *Thoughts*. The solution with k=5 did not further divide the Agency class but sharply separated the *Primary Emotions* from the *Secondary Emotions* and these two from the Perceptions in compliance with the division proposed by D'Andrade [8].

The interpretations of the different clusters elicited by the results of the present study with the overlaps and variances of the category names and contents point out the degree that these investigations are open to manipulation. The results showed that applying the two-dimensional approach suggested by Gray et al. [10] was possible since all the sub-categories offered in the remaining literature could also be represented under those two labels. According to our results, the Agency dimension bore both the agentic and cognitive concepts, while the Experience dimension included the perceptual or emotional concepts. Further possible divisions also signal that some concepts in each dimension require a high-capacity level to accomplish such complex cognitive processes while others are physical or instinct-like. So, classifying the concepts with respect to the level of capacities needed for them could be a helpful way of

Table 4: The results of the topic analyses based on the most frequent fifty quadgrams of four dimensions.

Dimension	Topic 1: Properties of the concept/process	Topic 2: Type of the agent	Topic 3: Cognitive/physical requirements for the concept/agent
High Agency	difficult, hard, complex	mentally and physically healthy humans can do	requires effort, reasoning, learning, mental prerequisites
Low Agency	instinct, automatic, primitive, reflexive, easy	agents other than humans can also do, among all humans	no prerequisite, planning, and low or no effort
High Experience	arising emotions or having an emotional intensity	specific to mentally and physically healthy humans	requires mental and cognitive maturity
Low Experience	innate, instinct-like, physiological, automatic, evolutionary, easy	common across all species and among all humans	no effort, straightforward

fostering discussions on determining where different agents land on the dimensions of the mind perception.

The results support the need for systematically selected conceptual pools and norming and validating processes to ensure the usability of the stimuli. Furthermore, our results demonstrate how strongly a concept is associated with one of the four dimensions. For future studies, researchers can consult on the representative items of the clusters, i.e., medoids, and decide which concepts they will select from that cluster. As an alternative, researchers can make their item selections based on the number of ratings for the main category and the capacity levels.

For now, adopting a two-dimensional structure is functional because of the design of the IAT task. However, a further division within each dimension would enrich the evaluations. Subcategorizing the Agency and Experience dimensions into High and Low dimensions was first suggested by Li et al. [21] in their MP-IAT task. Another significance of this step was that it allowed the researchers to have absolute measures for the Agency and Experience dimensions rather than measures relative to each other, as in the classical IAT, which helped them compare the implicit measures to the explicit ones characterized by the same concepts. Our study takes this further and provides a proper and systematic way of selecting concepts for both implicit and explicit tasks of mind perception.

Our study also provides insights into the decision-making process while participants categorize the concepts. Our NLP analysis on the open-ended questions allowed us to assess our instructions' clarity and showed whether participants used additional criteria in the categorization. We realized that our instructions were generally well-received, and our study acknowledges the complementary benefits and further insights provided by the language analyses.

4.2 Limitations and Future Work

The current study was conducted in a language other than English, unlike the majority of the studies in the field. Despite this fact, the concepts of Agency and Experience were consistent with their counterparts in English. The present study has also taken a significant step forward to test and validate the concepts used in mind perception research cross-culturally [22, 43]. Nevertheless, validating the implications and the results of the present study by repeating the same procedure in different languages, thus different cultures [16] would be informative.

We did not include some 'abstract' concepts with moral and philosophical implications for various reasons, such as the lack of context in our task and the length of the translated concepts into Turkish. However, the main reason was that we aimed to norm the concepts suitable for use while measuring the attribution of mental capacities to robots, not mental states. Accordingly, a common limitation of mind perception tasks might be that they usually focus on the mental capacity attributions, not the states, which are crucial to be explored to predict the future interactions of humans with robots.

In the survey, we presented the choices side by side below each question without changing their positions since, in our setting, it was not an association task, but a multiple-choice format, and our study lasted 5–10 minutes. However, counterbalancing the positions of the choices could avoid any confounds stemming from the stimulus-response association for future studies with longer durations. Also, presenting the definitions ready for the participants throughout the experiment could help prevent confusion or forgetting. We aimed to foster careful evaluations regarding the concepts and establish our selection on the ratings of the majority, so we let the participants take their time while they decide on their final responses. However, measuring response times could be an excellent opportunity to see how easy or difficult to associate a concept with one of the choices in future conceptual norming studies. Also, future investigations of individual differences could help reveal further aspects of the mind perception process. Finally, this study attempted to norm the concepts before their usage in implicit tasks, hopefully mitigating the problems stemming from self-report measures.

In a recent review, Thellman et al. [38] drew attention to the possible shortcomings of the implicit but still verbal judgments, pointing out that they are insufficient in capturing the participants' indeliberate thought processes while they give responses. They refer to some alternative measurements, such as eye tracking or neuroimaging techniques, as non-verbal ways of implicit investigations. We acknowledge the presence of alternative opportunities, and our research agenda aligns with the proposed direction toward non-verbal measures. However, as Thellman et al. [38] also pointed out, currently, neuroimaging techniques can also fall short in explaining some aspects of the mind perception process. So, we presume that with carefully selected stimuli and a normative procedure, implicit measurements serve us to gain insights about complex mechanisms before they are further investigated with non-verbal measurements. So, we take this opportunity to draw attention to normative studies before using any stimuli in non-verbal measurements since they are more costly and labor-intensive procedures than implicit tasks.

5 CONCLUSION

Overall, in this study, we suggest that all materials – not only pictures, vignettes, or videos – including the concepts and lexical resources, should be normed and validated before using them in an experimental setting. We consider this issue vital since HRI is a relatively young discipline, open to many investigations and collaborations between engineers, designers, and researchers from various fields. We foresee that creating valid and safe stimuli before the application of any tests could be a standard step in the field, and the experimental studies' reliability, replicability, and reproducibility would improve accordingly.

ACKNOWLEDGMENTS

This study is funded by The Scientific and Technological Research Council of Türkiye, TUBITAK (Project Number: 120K913). We thank Gaye Aşkın for her help with the literature review, Tuvana Karaduman, Aslı Eroğlu, Şeyda Evsen for their contributions to discussions of the study, Harun Olur, F. Elif Müjdeci, Duru N. Han, M. İlteriş Bozkurt for their contributions in item selection and translations, Serkan Pekçetin for his support in implementing the TF-IDF and PAM algorithms and Gizem Çınar and H. Papatya Yeşilbaş for being the raters of the language data. We thank the anonymous reviewers of the study for their insightful comments and suggestions.

REFERENCES

[1] Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind.* MIT press.

[2] Christoph Bartneck, Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kennsuke Kato. 2005. Cultural differences in attitudes towards robots. AISB.

[3] Mathieu B. Brodeur, Katherine Guérard, and Maria Bouras. 2014. Bank of Standardized Stimuli (BOSS) Phase II: 930 New Normative Photos. *PLOS ONE* 9, 9 (09 2014), 1–10. https://doi.org/10.1371/journal.pone.0106953

[4] Alan S. Brown. 1976. Catalog of scaled verbal material. *Memory & Cognition* 4, 1 (01 Jan 1976), S1–S45. https://doi.org/10.3758/BF03213263

[5] Rebecca Butler, Zoe Pruitt, and Eva Wiese. 2019. The Effect of Social Context on the Mind Perception of Robots. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63, 1 (2019), 230–234. https://doi.org/10.1177/1071181319631010 arXiv:https://doi.org/10.1177/1071181319631010

[6] Charles R. Crowell, Jason C. Deska, Michael Villano, Julaine Zenk, and John T. Roddy Jr. 2019. Anthropomorphism of Robots: Study of Appearance and Agency. *JMIR Hum Factors* 6, 2 (10 May 2019), e12629. https://doi.org/10.2196/12629

[7] Daniel C Dennett. 1996. *Kinds of minds: Toward an understanding of consciousness.* Basic Books.

[8] Roy d'Andrade. 1987. A folk model of the mind. *Cultural Models in Language and Thought* (1987), 112–148.

[9] Russell H. Fazio and Michael A. Olson. 2003. Implicit Measures in Social Cognition Research: Their Meaning and Use. *Annual Review of Psychology* 54, 1 (2003), 297–327. https://doi.org/10.1146/annurev.psych.54.101601.145225 arXiv:https://doi.org/10.1146/annurev.psych.54.101601.145225 PMID: 12172003.

[10] Heather M. Gray, Kurt Gray, and Daniel M. Wegner. 2007. Dimensions of Mind Perception. *Science* 315, 5812 (2007), 619–619. https://doi.org/10.1126/science.1134475 arXiv:https://www.science.org/doi/pdf/10.1126/science.1134475

[11] Kurt Gray, Joshua Knobe, Mark Sheskin, Paul Bloom, and Lisa Feldman Barrett. 2011. More than a body: Mind perception and the nature of objectification. *Journal of personality and social psychology* 101, 6 (2011), 1207.

[12] Kurt Gray and Daniel M. Wegner. 2012. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* 125, 1 (2012), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

[13] Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological review* 102, 1 (1995), 4.

[14] Anthony G. Greenwald, Miguel Brendl, Huajian Cai, Dario Cvencek, John F. Dovidio, Malte Friese, Adam Hahn, Eric Hehman, Wilhelm Hofmann, Sean Hughes, et al. 2022. Best research practices for using the Implicit Association Test. *Behavior research methods* 54, 3 (2022), 1161–1180.

[15] John A. Hartigan and Manchek A. Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[16] Nick Haslam, Yoshihisa Kashima, Stephen Loughnan, Junqi Shi, and Caterina Suitner. 2008. Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social Cognition* 26, 2 (2008), 248.

[17] Marcel Heerink, Ben Krose, Vanessa Evers, and Bob Wielinga. 2009. Measuring acceptance of an assistive social robot: A suggested toolkit. In *RO-MAN 2009 – The 18th IEEE International Symposium on Robot and Human Interactive Communication.* 528–533. https://doi.org/10.1109/ROMAN.2009.5326320

[18] Leonard Kaufman and Peter J Rousseeuw. 1990. Partitioning around medoids (program pam). *Finding groups in data: An introduction to cluster analysis* 344 (1990), 68–125.

[19] Dennis Küster and Aleksandra Swiderska. 2021. Seeing the mind of robots: Harm augments mind perception but benevolent intentions reduce dehumanisation of artificial entities in visual vignettes. *International Journal of Psychology* 56, 3 (2021), 454–465. https://doi.org/10.1002/ijop.12715 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijop.12715

[20] Minha Lee, Gale Lucas, and Jonathan Gratch. 2021. Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator and ultimatum games. *Journal on Multimodal User Interfaces* 15, 2 (2021), 201–214.

[21] Zhenni Li, Leonie Terfurth, Joshua Pepe Woller, and Eva Wiese. 2022. Mind the Machines: Applying Implicit Measures of Mind Perception to Social Robotics. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* 236–245. https://doi.org/10.1109/HRI53351.2022.9889356

[22] Bertram F. Malle. 2019. How many dimensions of mind perception really are there?. In *CogSci.* 2268–2274.

[23] Bertram F. Malle and Stuti Thapa Magar. 2017. What Kind of Mind Do I Want in My Robot? Developing a Measure of Desired Mental Capacities in Social Robots. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17).* Association for Computing Machinery, New York, NY, USA, 195–196. https://doi.org/10.1145/3029798.3038378

[24] Federico Manzi, Giulia Peretti, Cinzia Di Dio, Angelo Cangelosi, Shoji Itakura, Takayuki Kanda, Hiroshi Ishiguro, Davide Massaro, and Antonella Marchetti. 2020. A Robot Is Not Worth Another: Exploring Children's Mental State Attribution to Different Humanoid Robots. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.02011

[25] Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[26] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interaction Studies* 7, 3 (2006), 437–454.

[27] Brian A. Nosek, Carlee Beth Hawkins, and Rebecca S. Frazier. 2011. Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences* 15, 4 (2011), 152–159. https://doi.org/10.1016/j.tics.2011.01.005

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[29] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526. https://doi.org/10.1017/S0140525X00076512

[30] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[31] Imge Saltik, Deniz Erdil, and Burcu A. Urgen. 2021. Mind Perception and Social Robots: The Role of Agent Appearance and Action Types. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21 Companion).* Association for Computing Machinery, New York, NY, USA, 210–214. https://doi.org/10.1145/3434074.3447161

[32] Daniel B. Shank, Mallory North, Carson Arnold, and Patrick Gamez. 2021. Can Mind Perception Explain Virtuous Character Judgments of Artificial Intelligence? *Technology, Mind, and Behavior* 2, 3 (Aug 13 2021). https://tmb.apaopen.org/pub/6lnhwjhe.

[33] Joan G. Snodgrass and Mary Vanderwart. 1980. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory* 6 (1980), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

[34] Cristiane Souza, Margarida V. Garrido, and Joana C. Carmo. 2020. A Systematic Review of Normative Studies Using Images of Common Objects. *Frontiers in Psychology* 11 (2020). https://doi.org/10.3389/fpsyg.2020.573314

[35] Nicolas Spatola and Olga A. Wudarczyk. 2021. Ascribing emotions to robots: Explicit and implicit attribution of emotions and perceived robot anthropomorphism. *Computers in Human Behavior* 124 (2021), 106934. https://doi.org/10.1016/j.chb.2021.106934

[36] Nicolas Spatola and Olga A. Wudarczyk. 2021. Implicit attitudes towards robots predict explicit attitudes, semantic distance between robots and humans, anthropomorphism, and prosocial behavior: From attitudes to human–robot interaction. *International Journal of Social Robotics* 13, 5 (2021), 1149–1159.

[37] Tetsushi Tanibe, Takaaki Hashimoto, Tobu Tomabechi, Taku Masamoto, and Kaori Karasawa. 2019. Attributing Mind to Groups and Their Members on Two Dimensions. *Frontiers in Psychology* 10 (2019). https://doi.org/10.3389/fpsyg.2019.00840

[38] Sam Thellman, Maartje de Graaf, and Tom Ziemke. 2022. Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *J. Hum.-Robot Interact.* 11, 4, Article 41 (Sep 2022), 51 pages. https://doi.org/10.1145/3526112

[39] Sophie van der Woerdt and Pim Haselager. 2019. When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology* 54 (2019), 93–100. https://doi.org/10.1016/j.newideapsych.2017.11.001

[40] 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 arXiv:https://doi.org/10.1177/1745691610369336 PMID: 24839457.

[41] Adam Waytz, Kurt Gray, Nicholas Epley, and Daniel M. Wegner. 2010. Causes and consequences of mind perception. *Trends in Cognitive Sciences* 14, 8 (2010), 383–388. https://doi.org/10.1016/j.tics.2010.05.006

[42] Kara Weisman, Carol S. Dweck, and Ellen M. Markman. 2017. Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences* 114, 43 (2017), 11374–11379. https://doi.org/10.1073/pnas.1704347114 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1704347114

[43] Kara Weisman, Cristine H. Legare, Rachel E. Smith, Vivian A. Dzokoto, Felicity Aulino, Emily Ng, John C. Dulin, Nicole Ross-Zehnder, Joshua D. Brahinsky, and Tanya Marie Luhrmann. 2021. Similarities and differences in concepts of mental life among adults and children in five cultures. *Nature Human Behaviour* 5, 10 (2021), 1358–1368.

[44] Kubilay Özyer and Öznur Aşan. 2003. Duygusal Zekaya Etki Eden Demografik Faktörlerin Saptanmasına Yönelik Ampirik Bir Çalışma. *Hacettepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi* 21, 1 (2003). http://search/yayin/detay/40310