International Journal of Social Robotics (2025) 17:693–706
https://doi.org/10.1007/s12369-025-01231-8

Within, Between, Forced Choice, or Likert Scale? How Methodological Decisions Influence Recognition Rates in HRI Recognition Studies

Astrid Rosenthal-von der Pütten[1] · Julia Arndt[1] · Aleks Pieczykolan[2] · Maria Pohl[1] · Malte Jung[3]

Accepted: 26 July 2024 / Published online: 8 March 2025

Abstract
Recognition studies are a mainstay in HRI. Such studies are often used to test whether a robot's behavior is interpreted as intended by the designer. When designing recognition studies, researchers have to make important methodological decisions about the empirical study design (e.g., within-/between-subject design) and response format (e.g., forced choice, open text). Using the example of emotional expression recognition studies in HRI, we discuss (i) how theoretical conceptualizations determine methodological choices, (ii) the implications of the designs and response formats. We demonstrate in two experiments (n = 2654 and n = 419) that conclusions drawn from recognition studies are heavily dependent on study design and response format. We conclude with a set of recommendations for researchers employing recognition studies in their research.

Keywords Recognition study · Human–robot interaction · Methodology · Study design · Response format

1 Introduction

A central concern of human–robot interaction research is to develop an understanding about how people interpret and reason about the behavior of robots. Such understanding is not only crucial for advancing basic knowledge about human–robot interaction but also for our ability to design robot behavior that is easy to recognize and to interpret (e.g., [1–6]). Research examining people's interpretation of a robot's behavior often rely on judgment or recognition studies. Judgment studies are studies in which behaviors, persons, objects or concepts are evaluated by one or more judges, raters, coders, or categorizers [7]. By recognition studies we refer to that *subset* of judgment studies that examine the degree to which study participants can "recognize" a behavior as falling within a predefined set of categories.

In HRI research, recognition studies are typically used to test whether study participants can recognize a robot's expressive behavior as it was intended by the research team or designer of that expressive behavior. For example, a researcher has designed a gentle approach pattern for an autonomous car coming to a halt at a pedestrian crossing and uses a recognition study to examine whether the behavior is indeed interpreted as gentle by pedestrians. The studies often present participants with a set of behaviors (in form of a vignette, a picture, or a video but sometimes also live interactions) and participants have to indicate what kind of behavior they recognize or are asked more openly how they perceive the behavior in order to see whether their perception matches what designers wanted people to perceive (see Fig. 1). This approach has been used to evaluate robot (or virtual agent) behaviors for indicating cooperativeness [8], dominance or submissiveness [9], communicative and iconic gestures [10], emotional body language [11], emotional facial expressions [12], behaviors that convey introvert vs extrovert personalities of robots [13], behaviors that shall express a robot's

✉ Astrid Rosenthal-von der Pütten
arvdp@itec.rwth-aachen.de

Julia Arndt
julia.arndt@rwth-aachen.de

Aleks Pieczykolan
aleks.pieczykolan@rwth-aachen.de

Maria Pohl
maria.pohl@humtec.rwth-aachen.de

Malte Jung
mfj28@cornell.edu

[1] Chair Individual and Technology, RWTH Aachen University, Theaterplatz 14, 52062 Aachen, NRW, Germany

[2] Psychological Diagnostics and Intervention, RWTH Aachen University, Dennewartstrasse 25-27, 52068 Aachen, NRW, Germany

[3] Department of Information Science, Cornell University, 343 Campus Rd, Ithaca, NJ 14853, USA

"incapability" when completing a task [14], "emotion states" portrayed by a drone's flight path [15], as well as for the development of gestures used in sign language [16], just to name some of the manifold application areas for recognition studies.

When designing recognition studies researchers have to make important methodological decisions. For example, they have to decide whether each participant is shown only one type of behavior (between-subject design) or whether each participant is shown multiple types of behavior (within-subject design). Furthermore, researchers have to decide which response format to use in assessing people's ability to recognize a certain behavior, for instance, a forced-choice or Likert Scale or open text response format. Previous research has shown that such methodological decisions about the study design and response format have large implications for the conclusions we draw. Most famously, Russell [17] showed that the degree to which studies were able to demonstrate people's ability to reliably distinguish between emotion expressions was highly contingent on the particular response format a study employed. Following Russell's influential work, a number of other studies how shown that methodological choices greatly influence recognition rates (e.g., [18–20]).

Given the frequency with which recognition studies in HRI are employed and the broad variety of study designs and response formats those studies employ, it is important to establish an understanding on how methodological choices impact the inferences researchers can make.

Here, we use the case of robot emotional expression recognition to show how conclusions we draw about people's ability to accurately recognize a specific expression depend on methodological choices regarding study design and response format. We focus on emotion expression because this topic has received much attention in HRI (e.g. [21, 22]). This attention is constantly growing as researchers develop novel modalities for robots to be emotionally expressive, for example, through changes in light and sound patterns [23], flight patterns of drones [15], surface temperature [24], or surface morphology [25]. Moreover, emotion recognition studies were thoroughly investigated in psychology against the background of the concern that the popular forced-choice format artificially inflates recognition rates [18–20] which has been discussed in early HRI research as well [26] but was not influential enough to make a persistent impression in the field of HRI.

On a more general level HRI scholars have discussed the advantages and drawbacks of different methodological choices such as opting for within- or between-subjects designs, the importance of power analyses, and more specific with regard to studies on affective HRI the need to combine different measurements in HRI interaction studies [27, 28], but they did not address recognition studies and their specifics.

2 Related Work

2.1 The Purpose of Recognition Studies in HRI

An overarching goal in HRI is to design a robot's behavior in ways that is easily recognizable and interpretable by humans. Designers therefore often rely on behaviors that people are particularly familiar with. Emotional expressions have received much attention in HRI, since such expressions are thought of as an intuitive way for robots to communicate internal states [21]. Moreover, robots come in diverse forms and shapes, often not human-like, while researchers still want to provide these familiar ways of communicating. Hence, researchers develop novel modalities for robots to express emotions such as "emotionally expressive" flight paths for drones (e.g., [15, 29]). In the following, we a) give a brief review of theoretical grounding of emotion recognition research in psychology and the resulting methodologies, and b) a review on methodological fallacies that might diminish the results.

2.2 Theory Determines Method Choice: The Case of Emotion Recognition Studies

The universality thesis, which posits that emotional displays are universally expressed and understood, has split emotion psychology in two camps, those in support of the thesis (e.g., [30, 31]) and those who doubt universality (e.g., [32, 33]). However, a range of moderate opinions exists between these two extremes [34, 35, e.g.,] which might be rooted in the conceptualisations of emotion expression of these extremes. While on the one hand, Ekman [30] and Lizard [31] advocate for easy recognizable, discrete, and specific emotion categories such as happiness, surprise, fear, anger, disgust, contempt, and sadness, Woodworth and Schlosberg [34, 35], on the other hand conceptualize emotion expressions rather in terms of overlapping broad clusters and dimensions.

Against this background, it is not surprising that different theoretical conceptualizations resulted in different methodological approaches for addressing recognition of emotions. For instance, Klineberg [32, 33] demonstrated a large between-cultural variation in weeping from grief as evidence against the universality thesis. Izard [31] provided his participants with diverse emotional labels, following a one-to-many relationship between a sign (the facial expression) and what it signifies (a message about emotion). In contrast, Ekman and Friesen [36] presuppose a one-to-one correspondence between a specific emotion label and a specific facial expression.

When reviewing emotion recognition studies in HRI, it appears that the universality thesis has been broadly adopted by HRI scholars [21, 22]. Authors often refer to Ekman's six (or seven) basic emotions as starting point for their imple-

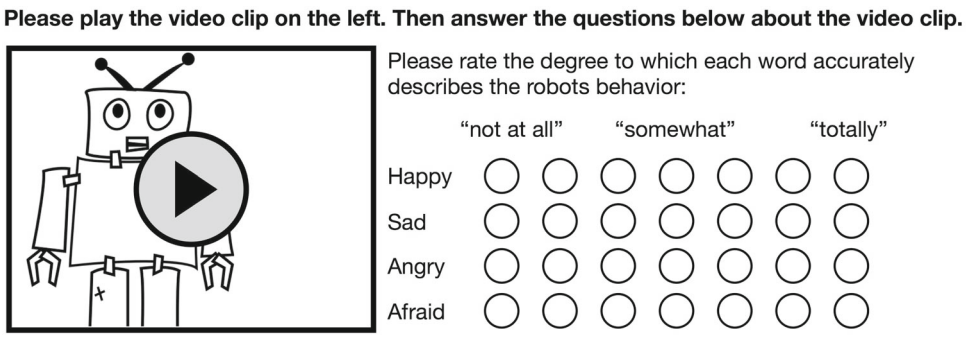

Fig. 1 Example of a typical survey used for an HRI recognition study

mentations of emotional displays in robots (e.g., [16, 26, 37–40]). As a consequence, they want to assess the success of their implementations by measuring recognition rates. Since the universality thesis is still under debate in many reviews on this issue, the trend in HRI to adopt Ekman's proposed one-to-one relationship approach neglects that there are other conceptualizations of emotional expression which might also be fruitful to consider in HRI.

In his review on cross-cultural studies in emotion recognition, Russel [17] took a close look at different methods used in this research field and raised concerns about their ecological, convergent, and internal validity. "Forced-choice response format, within-subject design, pre-selected photographs of posed facial expressions, and other features of method are each problematic. When they are altered, less supportive or nonsupportive results occur. When they are combined, these method factors may help to shape the results." [17, p.102]. Russel's results as well as later studies by other scholars suggest that certain methodological choices artificially inflate recognition rates (e.g., [18–20, 41–44]).

Just as the theoretical stance a researcher takes on when designing an emotional recognition study, any other type of recognition study will be influenced by pre-set theoretical assumptions. Researchers thus need to critically reflect about how their theoretical approach is determining their methodological approach and thus potentially limiting generalization and validity of their results.

In the following, we will briefly discuss methodological choices researchers face when designing recognition studies and the fallacies these choices might imply. We especially focus on our use case of emotion recognition. However, the problems sketched and conclusions drawn apply to all kinds of recognition studies and are therefore of wider importance.

2.3 Methodological Choices and Their Fallacies

2.3.1 Previewing & Within-Subject Design

One problematic aspect that Russel [17] identified was the practice of previewing the stimulus material and/or employing within-subject designs. Previewing involves that the participant takes a look at the complete stimulus material or parts thereof at least once before the actual recognition study starts. Within-subject designs involve that participants evaluate all emotional displays that are included in the study. Interestingly, both methods cause a similar bias, in that participants are prone to engage in more direct comparison between the various facial expressions. This bias would not emerge when utilizing a between-subjects design that is more realistic because, as Russel points out, it resembles rather everyday situations with encounters of facial expressions. Therefore, assimilation and contrasting effects can occur when participants are comparing the different emotional displays. Russel investigated this effect directly comparing a within and a between-subject design and showed lower recognition accuracy for the between-subjects design [45]. DiGirolamo and Russel proved in seven experiments the so-called elimination hypothesis: "As participants move from trial to trial and they encounter a type of expression not previously encountered in the experiment, they tend to eliminate labels they have already associated with expressions seen on previous trials; they then select among labels not previously used [46, p.538]. Other studies provided evidence for increased recognition rates for a certain stimulus (e.g., "sad") when this stimulus was preceded by another stimulus, for instance "happy" (raise from 44.2 to 58.3% [47]). Similar effects were reported by [20]. These examples show, that by using a within-subject design the order of presentation can greatly influence recognition rates, because certain preceding stimuli amplify recognition of the subsequent stimulus more than others.

2.3.2 Response Format

According to the different conceptualizations of emotional expressions (cf. Sect. 2.2, one-to-many vs. one-to-one) there are response formats that reflect these extreme views. Researchers of the one-to-one camp prefer a forced choice response format, in which participants are required to select one word from a pre-specified list of emotions (e.g., based on Ekman's basic emotions the list would be: happy, sad, afraid, angry, surprised, disgust, or contempt). The issue with this

response format is that participants have only a limited set of answers with no option to indicate that none of the labels fits or that they see something completely different in the picture. Indeed, there is empirical evidence that introducing a "none of the above" option to a forced-choice format produces significantly less agreement, while using an open-ended response format elicits even lesser agreement [19]. Forced choice can also produce artefactual agreement, for instance, when the correct emotion label is missing, participants will agree on an incorrect emotion label at rates greater than chance as demonstrated by [18].

According to Russel "forcing the observer to choose exactly one option treats the set of options as mutually exclusive, which they are not: Subjects place the same facial expression (or emotion of another or their own emotion) into more than one emotion category. Forced choice treats each option as an either-or (present-absent) choice, which they are not: Subjects reliably rate different facial expressions as belonging to a given emotion category to different degrees" [17, p. 116]. Hence, by using forced choice we are artificially simplifying the process of emotion recognition resulting in higher recognition rates.

Importantly, if emotion recognition were truly universal then all response formats should produce fairly similar recognition accuracy. Hence, researchers of the one-to-many camp, for instance, chose to ask their participants to freely label the pictures they viewed (e.g., [31]). However, in response formats involving open text answers (free label task), participants don't necessarily specify an emotion, at all, as demonstrated by Frijda [48].

As a middle way between restricted forced choice and free label response formats, researchers can also use quantitative ratings, such as Likert scales where participants indicate to which degree a picture is expressing a certain emotion. The advantage is, that participants can be asked to indicate the degree of several emotions for the same picture, which allows for assessing multiple ratings. Following the same rationale as for open text/free label tasks, we would assume that if a certain facial expression is easy to recognize and a universal signal is unique to a specific emotion, then recognition rates here should be comparable to those of the forced choice format. The question for such quantitative ratings is how to interpret multiple ratings for one facial expression. When using Likert scales, researchers should not simply look for the highest rating and treat this as successful recognition or not. They have to be aware of the issues of overall low ratings (e.g., when the highest rating is a "2" on a 6-point scale, is this really a successful recognition?) and score ties (i.e., participant assigned the same rating to two labels) when analyzing their data and have to decide on a procedure before-hand how to deal with them. Crucially, if these procedures are not reported in a recognition study using quantitative ratings, researchers should be careful when using these studies and their conclusions as basis for future work.

2.3.3 Facial Stimuli and Lack of Contextual Information

There are two concerns that are especially relevant for emotion recognition studies, though to some extent they also apply to other behaviors (e.g., gestures) used in recognition studies. Firstly, one concern of studies involving stimulus material of human faces is the preselection of pictures from a larger set (cf. [17]). Such selected pictures are by nature not representative of the population of facial expressions. Moreover, most pictures used as stimulus material are *posed and not spontaneous expressions*. While there are certain reasons for researchers to use posed pictures and preselect them (e.g., to increase experimental control), one has to be aware of the drawbacks. In fact, based on previous studies we only (partially) understand recognition of posed and preselected facial expressions, but unfortunately we know far less about how well unselected spontaneous expressions are recognized, despite of this being the default everyday situation. There is evidence that preselected pictures yield higher recognition rates compared to using a full picture set originally generated for a study, e.g., studies show that recognition rates on a full set dropped to chance levels (39.6% where chance was 33.3%) [49]. Moreover, Russel criticizes that "posed faces do not express the emotion of the poser, but what the poser chooses to pretend and in a manner most likely to be understood by the observer. According to the notion of display rules, voluntarily posed expressions are culturally influenced and have been said to originate in a different region of the brain than do spontaneous facial expressions (Rinn, 1984)" [17, p.114]. He further describes that posed expressions might be (i) exaggerated or more conventionalized, (ii) more similar within one emotion type, and (iii) more discernible from poses of other types of emotion expressions than spontaneous expressions which altogether raises concerns with respect to the ecological validity of these studies. Crucially, it has been shown that participants are sensitive to posing of emotional expressions [50]. This concern of ecological validity might be of less importance in HRI, since if researchers and designers want a robot to express a certain emotion and be sure humans understand and interpret the emotion as intended, it makes sense to use the more conventionalized versions of facial expressions. Meaning, using more striking examples of emotional expressions serves the purpose in these cases.

A second criticism regarding classic emotion recognition studies refers to the lack of contextual information. Usually, participants are presented with a series of pictures showing just the face of the emotion expresser without any information about what caused the emotion, in which setting did it occur and what is the accompanying behavior of the expresser. A

smile is not always a happy smile as suggested in the recognition studies, were a smiling face corresponds to the emotion "happy". A smile can be indeed joy, but sometimes we "smile something away" like a dispute we do not want to engage in or embarrassment, or we smile politely. We smile to greet someone without really enjoying the moment. Putting emotional displays into context can give them another meaning.

2.3.4 Summary

In the previous sections we outlined how methodological choices when designing a recognition study will influence recognition rates and supported our assumption with empirical findings within our use case of emotion recognition studies. However, the same principles extend to all types of recognition studies. Specifically, we showed that previewing and within-subject designs potentially inflate recognition rates. Moreover, forced choice answering formats potentially produce higher recognition rates than Likert scales or open-text formats. And finally, it can always be that stimuli presented without context artificially isolate the object to be recognized thus making its features more salient. Presenting it in a natural environment or paired together with other stimuli would reduce recognition rates because alternative interpretations of the same stimuli are possible.

2.4 Research Objective and Hypotheses

The goal of this work is to demonstrate the influence of methodological choices on recognition rates in recognition studies in HRI and in general. We will conduct two experiments, one in a within-subject design and one in a between-subject design, using different response formats and also using four different picture sets of a human and three robots. If the experiment design or the response formats have no influence there should be no noteworthy differences in recognition rates (***H0***). However, given the evidence presented above, we hypothesize that:

H 1 Recognition rates will be higher in a within-subject design experiment than in a between-subject design experiment.

H 2 Recognition rates will be the highest for the forced choice response format, and the lowest for the open text answers.

3 Experimental Studies

3.1 Overall Study Design

We tested the influence of four factors on emotion recognition rates with human and robotic stimuli: The four factors were experimental design (between-subject design vs. within-subject design), response format (forced choice, Likert scale, open text answer with emotion cue, open text answers without emotion cue), agent type (human face, iCat, Flobi, Kobian-RIII), and emotion type (anger, fear, surprise, sadness, happiness, neutral) resulting in a $2\times4\times4\times6$ design.

3.1.1 Two experimental designs

In order to address the influence of experimental design, we conducted two separate experiments with the same 3 factors: While agent type and response format served as between-subject factors in both experiments, the third factor emotion type was a between-subject factor in Experiment 1 but a within-subject factor in Experiment 2.

In Experiment 1, each participant was presented with only one stimulus, so there was only one trial per participant. A trial consisted of one of the 256 ($= 4 \times 4 \times 6$) possible combinations of response format, agent type, and emotion type. Each combination was randomly chosen for every participant causing slight differences in the occurrence frequencies across combinations.

In Experiment 2, each participant viewed six different emotional stimuli (anger, fear, surprise, sadness, happiness, neutral), all from the same agent. Stimuli were presented in randomized order. In total there were 16 different conditions.

3.1.2 Agent Type & Emotion Type

The study included four picture sets comprised of four agent types each displaying six different emotions (see Fig. 2). The four agent types were computer-generated human faces [51] and three different social robots (iCat [52], Flobi [53], and Kobian-RIII [54]). Each agent displayed facial expressions based on the "universal" facial expressions hypothesised by Ekman [30]. From this original emotion set, the two emotions "disgust" and "contempt" were excluded due to the lack of robot pictures displaying these emotions. Instead, a neutral face expression was included. Eventually, the following six expressions were used: (1) anger, (2) fear, (3) surprise, (4) sadness, (5) happiness, (6) neutral.

3.1.3 Response Format

The factor response format consisted of four different conditions: Forced Choice (FC), Likert Scale (LS), open text answer with emotion cue (OT-EC), and open text answers without emotion cue (OT-NoEC, see Fig. 3).

In the forced choice condition (FC), participants were presented with six emotion labels and selected one emotion label that matched the picture best. The six response options corresponded to the six emotion type conditions. There were no additional response options (e.g., none). The order of the dis-

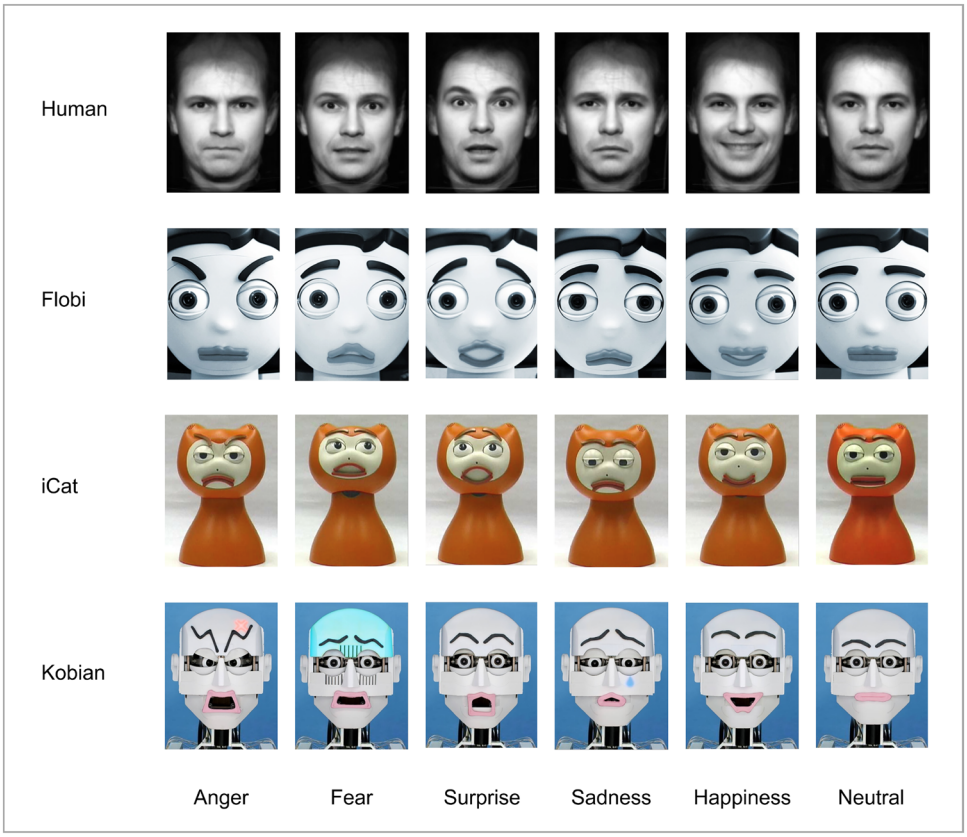

Fig. 2 Agent types & emotion types

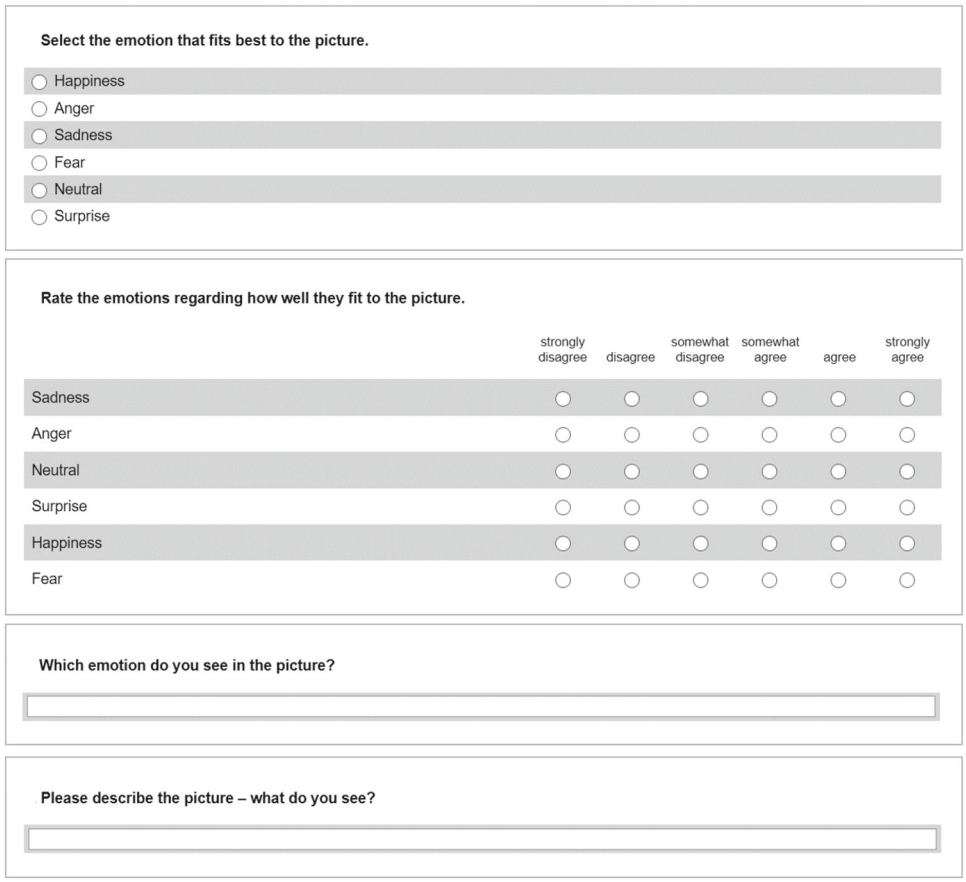

Fig. 3 Response formats

played emotion labels was randomized across participants in Experiment 2.

In the Likert Scale condition (LS), participants were presented six emotion label items and rated all six regarding the extent of the match with the picture on a 6-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (6). The order of the six items was randomized across participants in Experiment 2.

Both open text (OT-EC, OT-noEC) response formats consisted of one text question and a text box, into which participants should enter their open text response. They only differed in terms of the question phrasing which either included an emotion cue "Which emotion do you see in the picture?" or not "Please describe the picture: What do you see?".

3.2 Task & Procedure

Both experiments were conducted as online surveys. In both surveys, participants were randomly distributed to one of the 256 conditions in Experiment 1 or of the 16 conditions in Experiment 2, respectively. On the first page of the survey, participants were informed about the upcoming task, data protection, and their right to withdraw from the study at any time. They gave informed consent by clicking on the start button. To verify that participants were human and not bots, a captcha was presented. After the verification procedure, the recognition task started. Participants were instructed that on the following pages they are asked about their perception of faces. Therefore, they will be presented with one (or more) pictures. They shall take a look at this/these picture/s and answer the related question.

Trials in the recognition task consisted of one picture showing one of the six emotions displayed by one of the four agents. Below the picture, a question was presented that differed in terms of response format. For participants in Experiment 1, the recognition task was completed after one trial. In Experiment 2, there was a total of six trials according to the six different emotions displayed by one of the agent types. After the recognition task, participants were presented with three questions as attention checks. Responses to these control question were used as filters during data cleansing in further analyses. Finally, participants were asked to indicate age and gender.

3.3 Participants and Inclusion Criteria

Recruitment of participants was carried out via Amazon MTurk. The two experiments were conducted sequentially, so that Experiment 2 did not start until Experiment 1 was completed. Participants who took part in Experiment 1 were excluded from participating in Experiment 2. Inclusion criteria for participants were to complete the entire survey, to fill out the survey in English and to pass the test questions. Participants were naive about the purpose of the study. Both experiments have been conducted in accordance with the last revision of the declaration of Helsinki and in accordance with the standards of the American Psychological Association. Participants who completed the entire survey and passed the attention check (see subsequent paragraphs) were reimbursed monetarily within usual limits, and their anonymity will be guaranteed.

3.3.1 Participants in Experiment 1

In Experiment 1, 2742 participants took part. For data cleansing purposes we included three test statements, that participants had to respond to: (1) "This study is not about facial expressions" (5-point Likert scale), (2) "This study is about e-learning" (5-point Likert scale), (3) "This study is about emotions" (5-point Likert scale). With regard to the data cleansing procedure, our test questions were not applicable. Especially for participants in the condition with the open text answer type "no emotion cue" the answer "e-learning" would have been sensible, since we provided no information on emotion-related content in our instructions. Therefore, we decided to omit the test questions altogether and included all participants who completed the survey and did so in English. However, for participants in the response format condition with open text entries, we still utilized text entries as quality check. We checked all cases of non-completed surveys in which participants quit before stating their age in the conditions "open text with emotion cue" and "open text without emotion cue" (n = 24). We retained those cases with meaningful answers, which was true for all 24 cases that were examined this way.

This data cleansing procedure yielded 2654 participants (1011 male, 1616 female, 27 not indicated) with a mean age of 35.4 (SD = 11.74; $range$ = 16–89 years, based on 2630 participants).

3.3.2 Participants in Experiment 2

In Experiment 2, 541 participants took part. Based on our experiences with data cleansing in Experiment 1, we developed a new attention check procedure. First, we asked participants after the recognition task to memorize the number 42 and then asked three test questions: (1) "This study is not about facial expressions" (yes/no), (2) "This study is about e-learning" (yes/no), (3) "This study is about emotions" (yes/no). Afterwards, we asked participants to recall the number we asked them to memorize before. Participants had two chances to enter the correct number. As in Experiment 1, participants who did not complete the survey or did not respond in English were excluded from further analyses. We implemented a two-staged data cleansing procedure.

First, participants who entered the number 42 correctly and passed two of the three test questions were retained ($n = 392$). However, this resulted in a small number of participants in the condition OT-EC with the agent type "human". Hence, also those participants were retained who passed two of three test questions correctly even though they did not recall the number 42 ($n = 27$).

The final data set of Experiment 2 included 419 participants (213 male, 204 female, 2 not indicated) with a mean age of 36.5 ($SD = 11.2$; range = 19–71 years, based on 417 participants).

3.4 Coding of Open Text Answers

For the coding process of the open text answers, we implemented the following procedure: Based on the answers of the open text conditions, a set of verbal expressions was defined for each emotion that was regarded as semantically equivalent to the literally correct answer and therefore coded as correctly recognized (e.g., literal meaning "anger" with equivalent expression "mad"). A list of all equivalents can be found in the Appendix. Any other answers were coded as incorrectly recognized. This also applies to descriptions of activities typically associated with emotions (e.g., the verbs *to cry* or *to laugh*), since a verb does not unequivocally indicate whether the participants indeed recognized a particular emotion (for instance, "show a happy, nervous, sarcastic or sad laughter").

We decided to regard a fairly broad range of expressions as equivalents, even if their meaning slightly differed from the exact emotion meaning (e.g., emotion type: "sadness" with equivalent "despair"). Crucially, in order to qualify as an equivalent the expression should have a clear proximity to one of the six emotion types. Responses that could be assigned to two different emotion types were regarded as incorrectly recognized (e.g., emotion types: "fear" and "surprise" with answer "shock"). Apart from the correct recognition of one of the six emotions, 4 other classification of responses were defined. With these four additional categories we have a total of 10 different codes, from which only one represents the correct recognition of the emotion.

1. *Other Emotions* This category included entries of naming internal states that could not conclusively be assigned to one of the six emotion types (e.g., "shock", "impatience", "worry").
2. *Descriptions Without Emotional Reference*: Pooled in this category are different types of descriptions without any reference to internal states. This includes descriptive nouns for the agent (e.g. "robot") as well as adjectives not directly associated with emotional states (e.g. "thinking"). Furthermore, we decided to include entries of emotional adjectives that did not unambiguously refer to an internal state. For this reason, the grammatical form "annoyed" is included in the list of valid paraphrases, while "annoying" is regarded as a description.
3. *Descriptions with Emotional Reference* Entries that did not explicitly referred to an emotion, but instead consisted of activities directly associated with emotional states (e.g. "laughing", "crying", "frowning").
4. *Inconclusive Entries*: Entries with inconclusive or even non-verbal content (e.g., "wtf", ":D").

3.4.1 Handling response ambiguity

In the condition OT-EC, most participants responded with a single (emotion-related) word in accordance with the instructions. However, participants in the condition OT-noEC were requested to describe the picture, which resulted in longer phrases of several words. Since phrases increase the potential of ambiguous or imprecise meaning, we established a rule set to deal with such conflicting content within text entries. First, it was always checked for the presence of one of the six emotion types. If that was the case, the remaining text content was evaluated with a lower priority. For instance, a text entry such as "a robot with a tear, looking sad" resulted in the classification of "sadness", even though it included an additional description with an emotional reference. However, we differentiated between two different types of conflicts even if one of the six emotion types was present. Conflict type A: Naming the correct emotion type and at least one further from the remaining five emotions. This was still regarded as recognition of the correct emotion type. Conflict type B: Naming more than one incorrect emotion types. In order to avoid a bias for one incorrect emotion (since there were at least two), we classified this response into the category "other emotion".

3.4.2 Inter-rater Reliability

Any open text entries have been coded by two coders. We computed inter-rater reliability using Cohen's Kappa, for each emotion type separately. For study 1, Cohen's Kappa ranged between $kappa = .82$ and $= .97$ indicating excellent agreement. For study 2, Cohen's Kappa ranged between $kappa = .92$ and $kappa = .98$ indicating excellent agreement. Only codings for the "neutral" emotion stimulus were slightly lower with $kappa = .81$ in the OT-EC condition and $kappa = .71$ in OT-noEC condition.

3.5 Data Transformation of Different Response Formats for Statistical Comparison

To compare the recognition rates of the different response formats, data had to be re-coded into a common response format. We re-coded all responses into a dichotomous dependent

variable format: Correct recognition (+) and no recognition (−).

FC The choice of the correct emotion label was coded as correct recognition. All other answers were coded as no recognition.

LS Two conditions had to be met for responses from the LS condition in order to be coded as a correct recognition. First, the correct emotion hat to be rated above the mid point of the 6-point Likert scale (4 "somewhat agree", 5 "agree" or 6 "strongly agree"). Second, no other emotion should have received a higher or equally high rating.

OT-EC & OT-noEC Only answers with either literally correct label or corresponding to the defined set of equivalents expressions for each emotion were counted as correctly recognized.

3.6 Statistical Analysis and Results

3.6.1 Experiment 1: Between-Subject Design

We computed Chi square tests to examine the influence of response format, emotion type, and agent type on overall emotion recognition rates. See Table 1 for recognition frequencies for the different response formats across the six emotion types. See Table 2 for recognition frequencies for the different agent types across the six emotion types.

We found that the effect of response format on overall recognition was significant, (χ^2 (3, n = 2654) = 152.0, p <.001, Cramer's V = .24). In line with our hypothesis (**H2**), participants in the FC condition more frequently as expected correctly recognised the emotion displayed while in the condition OT-noEC participants recognised less often then expected the displayed emotion. The effect of emotion type on overall recognition was significant, too (χ^2 (3, n = 2654) = 259.1, p <.001, Cramer's V = .31). Finally, the effect of agent type on overall recognition was significant (χ^2 (3, n = 2654) = 31.71, p <.001, Cramer's V = .11) as well.

3.6.2 Experiment 2: Within-Subject Design

Since all participants completed all six emotions, we computed Chi square tests for all emotion types separately. See Table 1 for recognition rates for the different response formats across the six emotion types and Table 2 for recognition rates for the different agent types across the six emotion types.

Effect of Response Format Influence of response format on emotion recognition was significant for

- *Fear* (χ^2 (3, n = 419) = 12.0, p = .007, Cramer's V = .17), recognition was higher than expected in the FC condition and lower than expected in LS condition.
- *Surprise* (χ^2 (3, n = 419) = 14.92, p = .002, Cramer's V = .19), recognition was higher than expected for FC and lower than expected for OT-EC.
- *Neutral* (χ^2 (3, n = 419) = 81.89, p <.001, Cramer's V = .44), recognition was higher than expected for FC and LS and lower than expected for OT-EC and OT-noEC.
- and non-significant for anger, happiness, and sadness.

Effect of Agent Type Influence of agent type on emotion recognition was significant for

- *Anger* (χ^2 (3, n = 419) = 26.37, p <.001, Cramer's V = .25), recognition was higher than expected for Kobian and Flobi, and lower than expected for Human and iCat.
- *Fear* (χ^2 (3, n = 419) = 15.31, p = .002, Cramer's V = .19), recognition was lower than expected for iCat.
- *Surprise* (χ^2 (3, n = 419) = 12.89, p = .005, Cramer's V = .18), recognition was higher than expected for human and lower than expected for Flobi and iCat.
- *Sadness* (χ^2 (3, n = 419) = 11.36, p = .010, Cramer's V = .17), recognition was higher than expected for human and iCat and lower than expectd for Kobian.
- *Happiness* (χ^2 (3, n = 419) = 24.74, p <.001, Cramer's V = .24), recognition was higher than expected for human and lower than expected for Kobian.
- but not significant for neutral.

3.6.3 Comparison Between Experiments

For a comparison of recognition rates between experiments and across answering formats, we can inspect the results reported in Table 1. Across all answering formats we see that recognition rates are higher in the within-subject study than the between-subject study. In total, the recognition rates were 16.6% higher in the within-subject design. For some emotions this increase war smaller (e.g., 5.1% for fear) and others it was particularly high (e.g., 26.1% for happy). Moreover, we see that the forced choice response format in both studies lead to highest recognition rates.

4 Overall Discussion

Motivated by the large number of recognition studies employed in HRI research, we explored how methodological choices in the design of such studies affect the conclusions researchers can draw. As a use case, we focused on the recognition of facial displays of emotion in robots since recognition studies for emotionally expressive robots are particularly common in HRI.

Most importantly we found that the choice of methodology has large implications for the kinds of claims researchers are likely to be able to make. For example, confirming

Table 1 Emotion recognition rates (%) for the different response formats in both experiments

	Total	Anger	Fear	Surprise	Sadness	Happy	Neutral
Between subject design							
FC	63.7	66.0	38.0	74.7	91.4	54.6	57.5
LS	48.3	54.6	30.9	51.4	73.5	33.3	47.1
OP-EC	50.2	62.2	29.8	41.3	85.3	62.9	19.1
OT-noEC	30.1	41.0	22.7	40.0	55.1	10.7	9.0
Total	47.8	55.7	30.2	51.4	76.0	40.5	32.6
Within-subject design							
FC	75.3	83.3	42.6	76.9	89.8	80.6	78.7
LS	65.2	76.6	22.5	65.8	86.5	76.6	63.1
OP-EC	58.7	68.6	41.2	52.0	80.4	81.4	28.4
OT-noEC	57.3	74.5	35.7	60.2	78.6	67.3	27.6
Total	64.4	75.9	35.3	64.0	84.0	76.6	50.4

Table 2 Emotion recognition rates (%) for the different agent types in both experiments

	Overall	Anger	Fear	Surprise	Sadness	Happy	Neutral
Between subject design							
Human	52.2	33.9	29.6	68.8	62.7	68.1	50
Flobi	45.3	53.1	43.6	44.1	91.9	24.3	13.7
iCat	40.1	52.6	13.8	29.8	79.2	31.5	33.3
Kobian	53.7	83.6	33.6	63.9	70	38.3	33.6
Total	47.8	55.7	30.2	51.4	76	40.5	32.6
Within-subject design							
Human	68.2	62.8	37.1	78.1	91.4	80.6	59
Flobi	63.7	84.9	41.5	56.6	80.2	76.6	42.5
iCat	61.6	67.6	19.6	58.8	88.2	81.4	53.9
Kobian	63.7	87.7	42.5	62.3	76.4	67.3	46.2
Total	64.4	75.9	35.3	64.0	84.0	76.6	50.4

our hypothesis (*H1*) using a within-participants design over a between participants-design boosts recognition rates by 16.6%. Similarly influential on recognition rates is a study's response format. As predicted (*H2*) we found that using forced choice boosts recognition rate by a minimum of 13.5% (in comparison to open text field with an emotion cue; OT-EC) up to 30.6% (in comparison to open text field with no cue; OT-noEC) over other response formats in the between-subject design. In the within-subject design experiment forced choice inflated recognition rates between 10 and 18% compared to the other response formats. It is noteworthy that some recognition for happiness and neutral was specifically low in OT-noEC (around 10%) in the between-subject design, while in the within-subject design recognition rates were 27% and 67%. This is a powerful example how the within-subject design in itself provides participants with the necessary context that the study is about emotion recognition—even though no explicit information is given on the matter linking our findings to previous work on this matter (e.g., [46]). In both studies, recognition rates varied between the emotion types with lowest rates for fear (across all response formats) and high rates for anger, sadness and happiness. The recognition rates for the human picture set in the within-subjects forced-choice study are comparable to those in previous studies by Ekman and other researchers that were reviewed in Russel's meta analysis [17], except for the emotion fear. Indeed, we also observed that every agent type (not only the human) scored specifically low on one emotion in contrast to the other agents or the other emotional displays. The human scored low on fear (in both study designs, i.e. 29, .6% and 37.1%) compared to other emotions. Since we used computer generated human faces that were averaged over a large set of real human faces to reduce effects of physiognomic variability. It might be, however, that especially the recognition of fear is easier in non-prototypical (averaged), but real faces. In another judgment study using averaged nonverbal behavior and testing it against individual nonverbal behavior (in this case gestures of a virtual agent using gestures learned from individual speaker data or averaged over a group of speaker data) the individual nonverbal behavior lead to better understanding of the agent and more positive evaluations [55]. A similar effect might have emerged here

for the recognition of an averaged fear expression. For the iCat the particularly low recognition rate for fear (in both studies below 20%) corresponds to the results of the original study from which we took the picture set [52]. In that respective study, participants also rated the difficulty to recognize fear highest among all tested emotional displays. So it seems that the design of that emotional display is generally hard to recognize as "fear". In the between-subjects study only, the human scored comparably low on anger and Flobi on happiness. This might be an indicator that these two emotions (displayed by these two agents) specifically profit from within-subjects designs and the possibility to eliminate other options (elimination hypothesis; [17, 46]).

4.1 Implications for Research on Emotional Expression in HRI

Human–robot interaction research largely sees emotional expressions as a means "to make robots more understandable, likable, intuitive, and predictable (or 'believable') by using patterns that allow people to apply mental models and heuristics from interactions with people to infer a robot's internal states and intentions" [21]. Irrespective of a growing debate about the adequacy of basic emotion perspectives in both psychology (e.g., [56]) and human–robot interaction (e.g., [21, 22]), it is important for developers of expressive robots that people can reliably identify a robot's expressions (as it was intended by the designers). Prior work demonstrated that choosing within-subject designs as well as forced choice artificially inflates emotion recognition [17–19, 46]. For HRI studies, this is not a trivial issue, since most recognition studies are conducted using a within-subjects design and some studies use previewing and familiarization procedures because their participants are confronted with a robot for the first time (e.g., [57]). Moreover, the majority of the studies use the forced choice response format. This implies that the overall impression arising from these studies that humans are quite good in identifying emotional expressions in robotic behavior might be exaggerated. This notion is backed up by our experimental results demonstrating how study design influences recognition rates with drastic differences for some aspects.

In his critique, Russel [17] mentioned two more factors we briefly want to discuss here: the influence of preselected and posed pictures and the lack of contextual information. Using preselected or posed pictures in recognition studies (cf. Sect. 2.3.3) might or might not be a problem in HRI depending on the viewpoint and the goal of the team developing and implementing emotional displays for robots. We can assume that most emotional displays for robots are more close to those posed, exaggerated, and stylized displays Russel refers to, because they are built upon human posed and preselected facial stimuli. However, if the goal is to be easily readable for humans, then this might not create a problem to researchers, since these stylized expressions might just be the right way to design easy to recognize robot behavior. The lack of contextual information, however, is quite important for HRI. Putting emotional displays into context can give them another meaning: a smile can express happiness or can be used to smile away embarrassment. Thus, what we learn from recognition studies in which emotional displays are presented in isolation is limited and maybe not generalizable in the sense that universal recognition as defined by Ekman and colleagues might only be true if we encounter smiles without greater context. For HRI studies this means that researchers should be alert not to be lulled into a false sense of security that their developed displays will always be recognized as intended, in every situation, by every user. In fact, some empirical studies already demonstrated the influence of context in emotion recognition [58].

4.2 Implications for the Design of All Types of Recognition Studies in HRI

For researchers conducting recognition studies we have several recommendations for their study design. First, as have other HRI scholars before us (e.g., [28]) we suggest to use between- rather than within-subject designs to establish recognizability. When a within-subject design is used, recognition rates will likely be inflated over what can be expected in an actual context of use. Ideally studies employ both a within-subject and a between-subject design to allow a comparison of recognition rates. Second, when using a within-subject design, it is important to completely randomize the order of presentation. For example, just using two alternative orders will not eliminate sequence effects (cf. Sect. 2.3.1). Third, it is important to keep in mind that forced choice response formats are problematic when attempting to establish general recognizability. Forced choice recognition does not reflect real-world recognition and artificially inflates recognition. At a minimum, we recommend to include alternative answers into the forced choice measure such as "none of the above", "does not apply" or a combination of forced choice and open text or a comments section. Finally, even though rarely used, better options such as quantitative rates and open questions exist to assess recognition rates. When using quantitative ratings, it is important to decide how to analyze data and define "cut offs" and how to treat scored ties.

5 Limitations and Future Studies

Our study was subject to some limitations, namely that participants were not screened for their ability to accurately recognize emotion. Moreover, since our test questions were not suitably designed we had to remove our inclusion criteria

during analysis for Experiment 1 and some inclusion criteria for Experiment 2 which might have affected the results. Between response formats information on the purpose of the study were more or less obvious, i.e. in contrast to all other response format the open text with no emotion cue condition did not provide any explicit information that the study is about emotional displays and how people judge them. Hence, future studies with similar study designs need to implement check questions that can cope with this disparate information states. In sum, more careful construction of attention checks is needed for future studies. Moreover, our study is limited as it only looked into one specific use case of recognition studies—emotional displays in picture material—though we are very confident that the same tendencies will be found in other types of recognition studies on other types of behavior and other stimulus material such as videos. Therefore, also based on the discussion above future studies should explore more deeply how emotional facial expressions displayed by (a larger set of different) robots are perceived by observers when they are not provided as still pictures but presented dynamically in videos as well as in different interaction contexts. Lastly, although we discuss the possible effect of alternative answers we did not include this variation into our (already complex) study design. Hence, future work could explore the effect of including alternative answers such as "none of the above", "does not apply" o recognition rates.

6 Conclusion

Our paper makes an important contribution to the literature as we discuss how methodological choice influence recognition rates in HRI recognition studies. Taking the use case of emotion recognition studies, we briefly review and question the thesis that emotion recognition is universal. We outline how being in favor for the universality thesis of emotion recognition can determine the methodological choices for recognition studies (e.g., using forced choice response format). From that, we discuss the implications and consequences that come with different methodological choices and highlight that the study designs used by most (emotion) recognition studies in HRI does not convincingly establish recognizability. Crucially, we provide empirical evidence which shows that recognition rates are hugely dependent on methodological choices researchers make.

Appendix A List of Equivalent Emotion Words

1. *Anger* Anger, angry, annoyance, annoyed, exasperated, hate, mad, rage, upset
2. *Fear* Afraid, anxiety, fear, fearful, fright, frightened, frightful, horrified, scared, terrified, terror
3. *Surprise* Amazed, amazement, astonishment, astounded, baffled, bewilderment, caught off guard, in awe, overwhelmed, perplexed, startled, stunned, surprise, surprised, wonder, wondering, wowed
4. *Sadness* Depressed, depression, despair, down, grief, sad, sadness, sorrow, unhappiness, unhappy
5. *Happiness* Amused, amusement, cheerful, content, contentment, excited, excitement, happiness, happy, joy, joyful, overjoyed, pleased, pleasure, satisfaction, satisfied
6. *Neutral* Apathetic, apathy, at rest, blank expression, calm, calmness, dead expression, emotionless, empty expression, expressionless, flat expression, idle, indifference, indifferent, little emotion, neutral, no emotion, none, non-expressive, not any emotion, quiet, uncaring, unconcerned, uninterested

Author Contributions Astrid Rosenthal-von der Pütten and Malte Jung contributed to the study conception and design. Material preparation and data collection were performed by Julia Arndt and Maria Pohl. Analysis was performed by Aleks Pieczykolan and Astrid Rosenthal-von der Pütten. The first draft of the manuscript was written by Julia Arndt, Astrid Rosenthal-von der Pütten, and Malte Jung and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors have no relevant financial or non-financial interests to disclose.

Data Availability The data of this project will be made available upon request.

Declarations

Ethics Approval The studies were conducted guaranteeing respect for the participating volunteers and human dignity. We followed the national, EU and international ethical guidelines and conventions as laid down in the following documents: The Ethical guidelines of the German Psychological Association; The Charter of Fundamental Rights of the EU; Helsinki Declaration in its latest version. Specifically, participants gave informed consent prior to the study and were properly debriefed after completion of the study.

Conflict of interest The authors declare that they have no Conflict of interest.

References

1. Takayama L, Dooley D, Ju W (2011) Expressing thought: improving robot readability with animation principles. In: 2011 6th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 69–76
2. Wykowska A, Chaminade T, Cheng G (2016) Embodied artificial agents for understanding human social cognition. Philos Trans R Soc B Biol Sci 371(1693):20150375
3. Hoffman G, Ju W (2014) Designing robots with movement in mind. J Hum Robot Interact 3(1):91–122
4. Dautenhahn K, Nehaniv CL, Walters ML, Robins B, Kose-Bagci H, Mirza NA, Blow M (2009) Kaspar-a minimally expressive humanoid robot for human–robot interaction research. Appl Bion Biomech 6(3–4):369–397
5. Breazeal C (2003) Toward sociable robots. Robot Auton Syst 42(3–4):167–175
6. DeSteno D, Breazeal C, Frank RH, Pizarro D, Baumann J, Dickens L, Lee JJ (2012) Detecting the trustworthiness of novel partners in economic exchange. Psychol Sci 23(12):1549–1556
7. Rosenthal R, Robert R et al (1987) Judgment studies: design, analysis, and meta-analysis. Cambridge University Press, Cambridge
8. Straßmann C, Rosenthal-von der Pütten A, Yaghoubzadeh R, Kaminski R, Krämer N (2016) The effect of an intelligent virtual agent's nonverbal behavior with regard to dominance and cooperativity. In: Traum D, Swartout W, Khooshabeh P, Kopp S, Scherer S, Leuski A (eds) Intelligent virtual agents: 16th international conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, proceedings. Springer International Publishing, Cham, pp 15–28. https://doi.org/10.1007/978-3-319-47665-0
9. Rosenthal-von der Pütten AM, Straßmann C, Yaghoubzadeh R, Kopp S, Krämer NC (2019) Dominant and submissive nonverbal behavior of virtual agents and its effects on evaluation and negotiation outcome in different age groups. Comput Hum Behav 90:397–409. https://doi.org/10.1016/j.chb.2018.08.047
10. Cabibihan J-J, So W-C, Pramanik S (2012) Human-recognizable robotic gestures. IEEE Trans Auton Ment Dev 4(4):305–314. https://doi.org/10.1109/TAMD.2012.2208962
11. McColl D, Nejat G (2014) Recognizing emotional body language displayed by a human-like social robot. Int J Soc Robot 6(2):261–280. https://doi.org/10.1007/s12369-013-0226-7
12. Saldien J, Goris K, Vanderborght B, Vanderfaeillie J, Lefeber D (2010) Expressing emotions with the social robot probo. Int J Soc Robot 2(4):377–389. https://doi.org/10.1007/s12369-010-0067-6
13. Lee KM, Peng W, Jin S-A, Yan C (2006) Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. J Commun 56(4):754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x
14. Kwon M, Huang SH, Dragan AD (2018) Expressing robot incapability. In: Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction. ACM, pp 87–95
15. Cauchard JR, Zhai KY, Spadafora M, Landay JA (2016) Emotion encoding in human–drone interaction. In: 2016 11th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 263–270
16. Akalin N, Uluer P, Kose H (2014) Non-verbal communication with a social robot peer: towards robot assisted interactive sign language tutoring. In: 14th IEEE-RAS international conference on humanoid robots (humanoids), 2014. IEEE, Piscataway, pp 1122–1127. https://doi.org/10.1109/HUMANOIDS.2014.7041509
17. Russell JA (1994) Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol Bull 115(1):102
18. Frank MG, Stennett J (2001) The forced-choice paradigm and the perception of facial expressions of emotion. J Personal Soc Psychol 80(1):75–85. https://doi.org/10.1037/0022-3514.80.1.75
19. Winters A (2005) Perceptions of body posture and emotion: a question of methodology. New Sch Psychol Bull 3(2):35–45
20. Romashov V (2018) How response format artificially increases emotion recognition rates. Master Thesis, Kozimski University, Warsaw
21. Jung MF (2017) Affective grounding in human–robot interaction. In: 2017 12th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 263–273
22. Fischer K, Jung M, Jensen LC, Wieschen MV (2019) Emotion expression in HRI–when and why. In: 2019 14th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 29–38
23. Song S, Yamada S (2017) Expressing emotions through color, sound, and vibration with an appearance-constrained social robot. In: Proceedings of the 2017 ACM/IEEE international conference on human–robot interaction. ACM, pp 2–11
24. Peña D, Tanaka F (2018) Touch to feel me: designing a robot for thermo-emotional communication. In: Companion of the 2018 ACM/IEEE international conference on human–robot interaction. HRI '18. ACM, New York, pp 207–208. https://doi.org/10.1145/3173386.3177016
25. Hu Y, Zhao Z, Vimal A, Hoffman G (2018) Soft skin texture modulation for social robotics. In: 2018 IEEE international conference on soft robotics (RoboSoft). IEEE, pp 182–187
26. Schiano DJ, Ehrlich SM, Rahardja K, Sheridan K (2000) Face to interface: facial affect in (Hu)man and machine. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI '00. ACM, New York, pp 193–200. https://doi.org/10.1145/332040.332430
27. Bethel CL, Murphy RR (2009) Use of large sample sizes and multiple evaluation methods in human–robot interaction experimentation. In: presented at AAAI Spring 2009 Symposium: experiment design for real-world systems
28. Bethel CL, Murphy RR (2010) Review of human studies methods in HRI and recommendations. Int J Soc Robot 2(4):347–359. https://doi.org/10.1007/s12369-010-0064-9
29. Sharma M, Hildebrandt D, Newman G, Young JE, Eskicioglu R (2013) Communicating affect via flight path: exploring use of the Laban effort system for designing affective locomotion paths. In: Proceedings of the 8th ACM/IEEE international conference on human–robot interaction. IEEE Press, pp 293–300
30. Ekman P (1972) Universal and cultural differences in facial expression of emotion. In: Nebraska symposium on motivation, pp 207–284
31. Izard CE (1971) The face of emotion. Appleton-Century-Crofts
32. Klineberg O (1940) Social psychology. Holt, New York
33. Klineberg O (1938) Emotional expression in Chinese literature. J Abnorm Soc Psychol 33(4):517
34. Woodworth RS, Schlosberg H (1954) Experimental psychology, Revised. Henry Holt, New York
35. Woodworth RS (1937) Experimental psychology. New York: Holt, 1938. Department of Psychology Dartmouth College Hanover, New Hampshire
36. Ekman P, Friesen WV (1988) Who knows what about contempt: a reply to Izard and Haynes. Motiv Emot 12(1):17–22
37. Arroyo D, Lucho C, Roncal SJ, Cuellar F (2014) Daedalus: a sUAV for human–robot interaction. In: Proceedings of the 2014 ACM/IEEE international conference on human–robot interaction.

HRI '14. ACM, New York, pp 116–117. https://doi.org/10.1145/2559636.2563709
38. Bennett CC, Šabanović S (2013) Perceptions of affective expression in a minimalist robotic face. In: Proceedings of the 8th ACM/IEEE international conference on human–robot interaction. HRI '13. IEEE Press, Piscataway, pp 81–82. http://dl.acm.org/citation.cfm?id=2447556.2447577
39. Cohen I, Looije R, Neerincx MA (2011) Child's recognition of emotions in robot's face and body. In: Proceedings of the 6th international conference on human–robot interaction. HRI '11 ACM, New York, pp 123–124.. https://doi.org/10.1145/1957656.1957692
40. Novikova J, Watts L (2014) A design model of emotional body expressions in non-humanoid robots. In: Proceedings of the second international conference on human–agent interaction. HAI '14. ACM, New York, pp 353–360. https://doi.org/10.1145/2658861.2658892
41. Gendron M, Crivelli C, Barrett LF (2018) Universality reconsidered: diversity in making meaning of facial expressions. Curr Dir Psychol Sci 27(4):211–219
42. Zupan B, Dempsey L, Hartwell K (2023) Categorising emotion words: the influence of response options. Lang Cognit 15(1):29–52
43. Kollareth D, Esposito J, Ma Y, Brownell H, Russell JA (2021) On evidence for a dozen new basic emotions: a methodological critique. Emotion 21(5):1074
44. Haidt J, Keltner D (1999) Culture and facial expression: open-ended methods find more expressions and a gradient of recognition. Cognit Emot 13(3):225–266
45. Russell JA (1991) Negative results on a reported facial expression of contempt. Motiv Emot 15(4):281–291
46. DiGirolamo MA, Russell JA (2017) The emotion seen in a face can be a methodological artifact: the process of elimination hypothesis. Emotion (Washington, D.C.) 17(3):538–546. https://doi.org/10.1037/emo0000247
47. Tanaka-Matsumi J, Attivissimo D, Nelson S, D'Urso T (1995) Context effects on the judgment of basic emotions in the face. Motiv Emot 19(2):139–155
48. Frijda NH (1953) The understanding of facial expression of emotion. Acta Psychol 9:294–362
49. Winkelmayer R, Gottheil E, Exline RV, Paredes A (1978) The relative accuracy of US, British, and Mexican raters in judging the emotional displays of schizophrenic and normal US women. J Clin Psychol 34(3):600–608
50. Reuter-Lorenz P, Davidson RJ (1981) Differential contributions of the two cerebral hemispheres to the perception of happy and sad faces. Neuropsychologia 19(4):609–613. https://doi.org/10.1016/0028-3932(81)90030-0
51. Vanger P, Hoenlinger R, Haken H (1998) Computer aided generation of prototypical facial expressions of emotion. Methods Psychol Res Online 3(1):25–38
52. Bartneck C, Reichenbach J, Breemen A (2004) In your face, robot! The influence of a character's embodiment on how users perceive its emotional expressions. In: Proceedings of the design and emotion
53. Lütkebohle I, Hegel F, Schulz S, Hackel M, Wrede B, Wachsmuth S, Sagerer G (2010) The bielefeld anthropomorphic robot head "Flobi". In: IEEE international conference on robotics and automation (ICRA), 2010. IEEE, Piscataway, pp 3384–3391. https://doi.org/10.1109/ROBOT.2010.5509173
54. Zecca M, Endo N, Momoki S, Itoh K, Takanishi A (2008) Design of the humanoid robot KOBIAN—preliminary analysis of facial and whole body emotion expression capabilities-. In: Humanoids 2008. IEEE, Piscataway, pp 487–492. https://doi.org/10.1109/ICHR.2008.4755969
55. Kopp S, Bergmann K (2012) Individualized gesture production in embodied conversational agents. In: Zacarias M, de Oliveira JV (eds) Human–computer interaction: the agency perspective. Springer, Berlin, Heidelberg, pp 287–301
56. Crivelli C, Fridlund AJ (2019) Inside-out: from basic emotions theory to the behavioral ecology view. J Nonverbal Behav 43(2):161–194
57. Ammi M, Demulier V, Caillou S, Gaffary Y, Tsalamlal Y, Martin J-C, Tapus A (2015) Haptic human–robot affective interaction in a handshaking social protocol. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction. ACM, pp 263–270
58. Read R, Belpaeme T (2014) Situational context directs how people affectively interpret robotic non-linguistic utterances. In: 2014 9th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 41–48