Developing 14 Animated Characters for Non-verbal Self-Report of Categorical Emotions

Gaël Laurans*
Pieter M.A. Desmet

Industrial Design Engineering
Delft University of Technology
Landbergstraat 15, 2628 CE Delft, The Netherlands
Email: glaurans@gmail.com
Email: p.m.a.desmet@tudelft.nl
*Corresponding author

Abstract

Graphical self-report tools are increasingly used to collect data on users' emotional responses to products, yet most of these tools have only undergone minimal validation. A systematic set of animations was developed to allow participants in design research and other fields to report their feelings without relying on the nuances of a particular language's affective lexicon. The animations were revised based on 8 studies across 4 countries (total N = 826). The set includes well-recognized animations representing desire/love, satisfaction/approval, pride/self-esteem, hope/optimism, interest/curiosity, surprise/excitement, disgust/aversion, embarrassment/shyness, fear/shock, and boredom/dullness. Two other emotions (joy/happiness and contempt/disrespect) were recognized by about half of the participants in the final study.

Keywords

Non-verbal; self-report; emotion; feelings; user experience.

Reference to this paper

Laurans, G. and Desmet, P.M.A. (2017) 'Developing 14 animated characters for non-verbal self-report of categorical emotions', J. Design Research, Vol. 15, Nos. 3/4, pp.214–233.

Biographical notes

Gaël Laurans is a researcher in human factors with Thales Research and Technology in Delft, The Netherlands. His work focuses on complex safety-critical systems in domains such as command-and-control, security monitoring, disaster management, and air traffic management. He is especially interested in making visual analytics and artificial intelligence usable for a wide variety of people and organisations. Prior to his current position, he was a researcher at Delft University of Technology, where he completed a PhD on the measurement of emotion applied to human-person interaction. His research covered theoretical and methodological issues when tracking the course of users' affective responses over time.

Pieter Desmet is a Full Professor of Design for Experience at the Faculty of Industrial Design Engineering at Delft University of Technology. He chairs a research group that studies design for emotion and subjective wellbeing. He is a Chairman of the International Design for Emotion Society and Co-Founder of Delft Institute of Positive Design, a scientific institute that initiates and stimulates the development of knowledge that supports designers in their attempts to design for human flourishing. Besides his research, he contributes to community projects, such as the Rotterdam-based cultural 'House of Happiness' foundation.

1. Introduction

Emotion has become a major focus of empirical research on the interaction between products and their users. This growth of interest in affective experience was accompanied by the development of many self-report tools to measure feelings. Since early psychological research on the topic (e.g. Zuckerman, 1960), many such instruments use adjectives or emotion names to describe specific feelings. These instruments therefore rely on the participants' grasp of subtle nuances of the affective lexicon and require a careful translation process before using them in other cultures, curtailing their applicability and usefulness for practitioners.

Perhaps for these reasons, non-verbal self-report instruments (e.g. Desmet, 2002) have proven popular with researchers, especially in design research. Unfortunately, the scope and empirical validation of many of these tools is still quite limited. This paper reports a series of studies aiming at improving and refining one such instrument and ultimately at providing a firmer empirical ground for the non-verbal self-report of feelings in design research and related fields.

2. Non-verbal self-report instruments

Self-report questionnaires rely on the participants in a research study reporting their feelings themselves using a pre-defined set of items. These items represent particular feelings, and respondents can select one or more item(s) to describe their current state or indicate how well each item matches their current experience. In verbal emotion self-report tools, items are simply words or lists of words. Non-verbal self-report relies on another type of representation than words, often stylized facial displays or cartoons (for an overview, see Desmet et al., 2016).

These non-verbal representations are particularly attractive for the measurement of feelings, as research has found emotions and expressive non-verbal behaviour to be intimately linked. Consequently, non-verbal self-report instruments allow research participants to describe their subjective experience without resorting to words, which offers important practical advantages when working with non-English speaking participants or other groups of people who might have difficulty verbalizing their feelings.

The most popular of these tools, the self-assessment manikin (SAM), assesses the three main dimensions of affect: pleasure (valence), arousal and dominance (Bradley and Lang, 1994). Because the meaning of these dimensions is difficult to grasp and to represent graphically, the administration of the SAM typically relies on extensive instructions including a long list of adjectives to "anchor" the scales, thus making even this ostensibly non-verbal tool highly dependent on verbal descriptions of feelings. The AffectButton (Broekens and Brinkman, 2013) uses computerized administration to address this issue and allow user-friendly self-report of the same three affect dimensions. The button presents a single dynamically changing iconic facial expression that changes based on the coordinates of the user's pointer in the button. Participants can report their feeling on all three affect dimensions with a single click. However, these tools are limited to the three main dimensions of affect and do not discriminate between distinct categorical emotions that are often more meaningful for research participants and design practitioners alike.

A long lamented limitation of common checklists and rating scales measuring categorical emotions is the lack of discrimination between positive emotions (Lorr and Wunderlich, 1988; Zuckerman et al., 1983). Even if some verbal instruments have begun to address this limitation (e.g. Scherer, 2005; Watson and

Clark, 1994), it is still acute when it comes to non-verbal self-report. This problem seems all the more pressing as various lines of research recently provided new evidence for the existence of distinct categories of positive affect (Sauter, 2010). This particular issue is also obviously particularly important for design, which primarily targets positive experiences (Yoon et al., 2014).

Some instruments addressing this limitation have been described in the literature, often with an eye toward specific applications like clinical psychology (Manassis et al., 2009; Stone, 2006), product appearance (PrEmo1; Desmet, 2002) or web user experience (Huisman and Van Hout, 2010). These disparate efforts demonstrate that there is a large interest in this approach to the measurement of feelings, but at the same time underlines the lack of a general-purpose fully non-verbal instrument with sufficient granularity.

3. Animated characters for 14 categorical emotions

PrEmo2, the measurement instrument presented in this paper, was inspired by an earlier effort to use non-verbal self-report to assess subjective experiences associated with the appearance of everyday products (Desmet, 2002). In design research, PrEmo1 was used to measure emotions evoked by a wide variety of products and other designed stimuli, such as wheelchairs (Desmet and Dijkhuis, 2003), automotive design (Desmet, 2003), mobile phones (Desmet et al., 2007), airplane meals and fabric conditioner fragrances (Desmet and Schifferstein, 2012), serving both as a means for generating insights for new product conceptualization and as a means for evaluating the emotional impact of new design concepts.

PrEmo2 comes in the form of a web application with several cartoon characters, one per emotion, analogous to the items of a traditional verbal questionnaire (Figure 1). Research participants have to click each of these characters in turn to see the corresponding animations. They can then report the degree to which the expressions of the animations correspond with their feelings with (three, five or seven-point) scales.

Figure 1. Stills of the 14 animations of the PrEmo2 self-report instrument (top row, from left to right: joy, hope, pride, admiration, satisfaction, fascination and attraction; bottom row: sadness, fear, shame, contempt, dissatisfaction, boredom and disgust).

Because industrial products are generally designed to provide a pleasant experience, the original instrument included several distinct categories of positive affect from the start. At the same time, users are able to provide separate ratings for each categorical emotion and do not need to place their current state

at a single location in a pleasure-arousal-dominance space. This allows participants to report mixed emotional experiences, which have recently been identified as a key aspect of emotion and emerged as an exciting new area of research, especially in design (Fokkinga and Desmet, 2012).

Using animations and body movement to represent these emotions also allows more granularity than other, simpler, representations of affect. This technique makes it plausible to completely avoid using emotion names in the measurement process, as the animations are much more expressive than static schematic faces and were specifically developed to convey rather specific meanings without extended instructions or explanations.

The format of the cartoons was thus designed to maximize the number of affective cues displayed in a limited space. Drawing the character from the waist up allowed the cartoonist to include a larger head and upper body movements, which are missing in many tools and have been found indispensable to identify self-conscious emotion like pride or shame (Tracy and Robins, 2007).

A major advantage of animations compared to static drawings like those used by Stone (2006) or Huisman and Van Hout (2010) is that they can represent dynamic displays of emotion. Previous research showed such displays to be more easily recognizable than still photographs (e.g. Bould et al., 2008; Fujimura and Suzuki, 2010). A second advantage is that these animations also include non-verbal vocalizations, a decision recently vindicated by independent research results that indicated cross-cultural recognition of emotions through such vocalizations (Sauter et al., 2010). A third advantage of cartoon animations is that they can amplify relevant expression signals, contributing to emotion recognition (see Rosset et al., 2008).

The PrEmo2 animations were developed using the procedure introduced by Desmet (2002). Six professional actors were taped as they independently enacted each emotion. A researcher then discussed the general patterns emerging from the actors' rendering of the emotion with the cartoonist who created the animations (Figure 2). Vocalizations were recorded later by another actor and synchronized to the animations.

Figure 2. Stills from the PrEmo2 "Contempt" animation.

Given the project's objectives (creating the most recognizable representations rather than testing a specific hypothesis about previously identified "basic" emotion expressions) and the lack of relevant data for dynamic emotional expressions, the actors were not trained to reproduce published facial configurations or body movements but simply to enact each emotion freely. Reassuringly, when a comparison was possible, the actors' depictions often matched closely those described in the literature. The actors' anger expressions, for example, showed the fixed stare, contracted eyebrows, and compressed lips that are typically reported in emotion expression studies (Ekman and Friesen, 1975),

The main difference between PrEmo2 and other non-verbal self-report tools is the set of emotions targeted. In Desmet (2002)'s original work, the items were selected empirically to cover the most frequent feelings that research participants associated with the appearance of everyday products. A more systematic

set of target emotions would make this measurement approach easier to apply in other settings including other subfields of design, general psychological research and clinical or industrial applications.

We therefore chose to base the development of PrEmo2 on the appraisal model proposed by Ortony, Clore and Collins (1988). It was selected as a good compromise between a purely bottom-up empirical approach focusing solely on a specific aspect of product experience and generic psychological approaches with a heavier focus on negative emotions and major life events. The Ortony, Clore and Collins model is particularly attractive for our purpose as it offers a theoretically sound basis for our work, grounded in modern appraisal theory, while still retaining an even balance between positive and negative emotions so often missing from appraisal theories. Finally, it has been found to be a fecund framework for design research (e.g. Desmet, 2008) and the availability of an easy-to-use measurement tool based on the same principle would make it easier to connect this work with empirical research.

Specifically, the new set of emotions is divided into four domains: general well-being emotions (joy, sadness, hope and fear), expectation-based emotions (satisfaction and dissatisfaction), social context emotions (admiration, contempt, shame and pride), and material context emotions (attraction, aversion, fascination and boredom).

4. Validating Non-Verbal Items for Self-Report Tools

While the voluminous literature on non-verbal expression of emotion provides much evidence for the link between various types of behaviours and affect, the validation of non-verbal items used for self-report tools requires a different standard of evidence than traditional facial expression research. Thus, reliable self-report requires items to convey the intended meaning to most research participants and not simply showing that some expressions are recognized more often than would be expected by chance. It is also important that each representation is understood after a single presentation, as opposed to multiple trials with several different exemplars of the target expression. However, recognition in isolated or exotic cultures, a popular criterion to identify “basic” emotions, is not important as such. Valid self-report does not require any claim about universality or innateness, merely good recognition by potential research participants.

Compared to other domains of applied psychometrics, the validation of emotion self-report questionnaires – whether verbal or non-verbal – has received relatively little attention. For example, while extensive studies are conducted to ensure that items in each subscale of personality questionnaires really do measure the same thing and are related to the psychological construct they are meant to represent, only a handful of published studies address the same question for categorical affective self-report questionnaires, mostly focused on negative emotions, clinical contexts and trait affect rather than ordinary feelings (Izard, 1972; Watson et al., 1988). Most often, researchers in applied fields put a list of words together, assuming that each of them has the same meaning to all potential participants as it has to the researchers and that each of the items thus constructed does therefore measure a distinct emotion.

There is nonetheless ample ground to doubt that the kinds of words used in emotion self-report questionnaires are as transparent as they might seem to researchers working with them constantly. For example, analysing a large set of studies using the trait version of the Differential Emotion Scale, Youngstrom and Green (2003) found that the internal consistency of several subscales differed depending on education, socio-economic status or age of the participants. This result suggests that the meaning of

emotion words selected to be synonyms appear more closely related for some (groups of) participants than others. Youngstrom and Green also hypothesise that research participants with greater education and higher social status could have greater familiarity with the nuances of emotion words because "in the United Sates, lower SES [social-economic status] groups tend to have social mores deemphasizing emotion expression, particularly in men, whereas higher SES and college education are associated with increased sensitivity and emotional expressiveness".

This underlines the need for additional empirical evidence regarding the validity of categorical emotion self-report scales. An important part of this validation effort is to establish that research participants interpret the representations supporting self-report – emotion names or non-verbal displays – as intended by the researchers. The methods used in facial expression research, and in particular the judgement study in which several participants or "judges" are asked to indicate how they understand a non-verbal display by matching it with the name of an emotion are very relevant for this purpose.

This paper presents data from eight separate such studies in the order in which they were conducted. To improve clarity, they are also grouped according to their objectives. The first two studies – conducted in China and the United Kingdom – were the first large sample test of the recognizability of the new animations (and, incidentally, the first judgement studies to assess categorical non-verbal representations of emotion intended for self-report with hundreds of participants across several cultures). The results of these studies inspired several revisions to the animations and were followed by two pilot studies to evaluate the results of these modifications. Two subsequent studies used different procedures and response formats to specifically investigate the nature of miscategorizations in the regular judgement studies. Finally, two extra studies tested the same animations against a new set of slightly adjusted interpretations.

5. Studies 1 and 2: Large Sample Judgement Studies

Besides providing a test of a new set of animations, studies 1 and 2 also benefited from an external collaboration that allowed us to address some of the limitations of the judgement studies used in the development of non-verbal emotion self-report tools. The financial and logistical support of a private partner thus enabled us to set up two large sample studies in different countries, with a broader population than university students. The composition of the sample, however, was in part limited by the industrial partner's interest and the participants were all women.

Study 1: China

The first large sample study with the new animation set was conducted in China. In this study, 411 women (mean age 33 ± 11) participated individually on a computer. The animations were presented one-by-one, asking first to categorize them as positive or negative and then to choose an emotion label from a list of 14 (plus "don't know"). Each label consisted of three words describing closely related feelings like "joy/happiness/pleasure". Instructions and labels were translated into Mandarin and the labels were also back-translated into English to check for any potential translation problem.

The intended emotional valence (i.e. the pleasantness dimension) was well recognized (participants chose the intended valence in 82% of all trials) except for the hope animation (52% of the participants categorized the emotion as "positive", 27% chose a negative label to describe this animation). Categorical judgement was less effective, with 52% of correct identification across all animations and all participants.

This headline figure hides a very contrasted picture, with good or very good recognition for some animations like satisfaction, joy, and most animations representing negative emotions while some other emotions were not recognized at all (see Table 1).

Table 1 Results from study 1 (China fixed-choice study, N = 411)

Animation	*Hit rate*	*Other choices above 10%*
Attraction	15%	Hope (46%), Fascination (14%)
Satisfaction	60%	Pride (19%), Joy (11%)
Pride	54%	Satisfaction (30%)
Hope	22%	Attraction (15%), Fear (12%)
Joy	59%	Satisfaction (14%), Pride (12%)
Fascination	15%	Attraction (40%)
Admiration	16%	Hope (20%), Attraction (19%), Fascination (17%), Joy (10%)
Disgust	81%	NA
Dissatisfaction	59%	Contempt (17%), Boredom (11%)
Shame	69%	Sadness (14%)
Fear	73%	Dissatisfaction (11%)
Sadness	89%	NA
Boredom	67%	NA
Contempt	66%	Dissatisfaction (23%)

Study 2: United Kingdom

Another large sample study was conducted shortly afterwards in the United Kingdom. In this study, 206 women (mean age 40 ± 10) participated individually on a computer using a slightly altered procedure. In this study, animations were presented two-by-two, with only seven possible labels (i.e. only positive emotions for animations intended to represent a positive emotion and vice versa) for each animation, plus a "don't know" choice. Each label consisted of a pair of words like "sadness, grief".

Because of the response format, it was not possible to assess valence recognition but the intended emotion was recognized in 63% of all trials. As in the China study, this overall hit rate hides large differences between animations.

Table 2 Results from study 2 (United Kingdom fixed-choice study, N = 206)

Animation	*Hit rate*	*Other choices above 10%*
Attraction	50%	Hope (21%), Joy (10%)
Satisfaction	61%	Pride (26%)
Pride	51%	Satisfaction (34%)
Hope	48%	Fascination (16%)
Joy	61%	Satisfaction (13%)
Fascination	72%	NA
Admiration	10%	Hope (29%), Joy (25%), Fascination (18%)
Disgust	88%	NA
Dissatisfaction	59%	Disgust (18%), Contempt (17%)
Shame	75%	Sadness (15%)
Fear	70%	Disgust (17%)

Sadness	93%	NA
Boredom	96%	NA
Contempt	50%	Disgust (34%)

Discussion of studies 1 and 2

The results of studies 1 and 2 paint a contrasted picture. On the one hand, the data clearly established that the animations were broadly effective in communicating affective meaning in two very different cultures and beyond student samples. On the other hand, the overall rate of recognition is clearly lower than one might have wished and the results point to several important weaknesses. These mixed findings also underline the value of stringent validation studies for the improvement of non-verbal self-report measurement.

Two more specific findings are also encouraging for the future of the technique. Firstly, some animations (e.g. disgust, sadness, fear or boredom) were very well recognized across these two challenging samples. Their interpretation and use for non-verbal self-report therefore already seems totally unproblematic. Secondly, even when an animation failed to convey its intended meaning to all participants, the answers were far from random. At the very least, an overwhelming majority of the participants could recognize the valence of the emotion and unintended interpretations were very often limited to one or two alternative meanings.

Thus, for example, nearly all participants who failed to recognize the pride animation as “pride” selected “satisfaction” instead. Obviously, “satisfaction” was not the label intended for this animation but “pride” and “satisfaction” seem quite similar conceptually and quite possibly share several appraisals. These data therefore suggest that the animation did convey a widely recognizable affective meaning that goes beyond a generic positive or negative feeling.

6. Studies 3 and 4: Revised Animations

Studies 1 and 2 pointed to several problems in the new set of animations. One way to address this problem is to improve the animations so as to clarify their meaning. Two animations (hope and admiration) were targeted for redesign based on their particularly low recognition scores.

The changes to the animations were based on the original actors’ takes collected during the initial development of the animations. One of the researchers looked at these videos again to identify alternative expressions or variants that could inspire new animations. In both cases, the changes concerned arm movements (Figure 3). Two small-scale pilot studies were then conducted to evaluate the effectiveness of the changes.

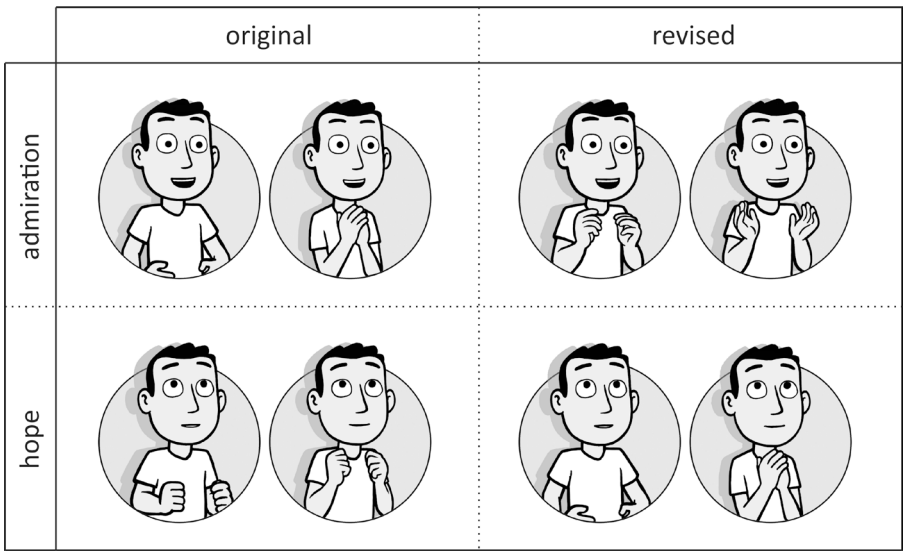

Figure 3. Two stills from the original (left) and revised (right) versions of the admiration (top) and hope (bottom) animations.

Study 3: Netherlands Revised Animations Study

Study 3 followed the same procedure as study 1. In this study, 22 students in Industrial Design Engineering (mean age 23 ± 3, 16 men) participated individually on a computer. The animations were presented one-by-one and the participants were first asked to categorize each animation as positive or negative and then to select a label from a list of 14 Dutch-language emotion names. A "don't know" choice was available for both questions.

Valence recognition was very good (95% across all trials) and the intended emotion was recognized in 75% of all trials. The hit rate for hope, one of the revised animations, was 82% (95% CI: [61%, 93%]) while admiration, the other revised animation, was recognized by 41% of the sample (95% CI: [23%, 61%]).

Table 3 Results from study 3 (Netherlands pilot judgement study, N = 22)

Animation	*Hit rate*	*Other choices above 10%*
Attraction	41%	Admiration (27%), Hope (14%), Fascination (14%)
Satisfaction	82%	Pride (14%)
Pride	57%	Satisfaction (33%)
Hope	82%	NA
Joy	73%	Admiration (14%)
Fascination	41%	Admiration (45%)
Admiration	41%	Fascination (41%), Hope (14%)
Disgust	100%	NA
Dissatisfaction	77%	Contempt (18%)
Shame	100%	NA
Fear	86%	NA
Sadness	95%	NA
Boredom	95%	NA
Contempt	82%	NA

Study 4: United States Revised Animation Study

Study 4 was conducted using Amazon's Mechanical Turk (AMT) crowdsourcing platform. The platform is increasingly used in social psychology or decision research and enabled us to reach non-student participants in another country. In study 4, 20 US-based users of AMT (mean age 32 ± 13, 11 men) took part on their own computers through the same web-based software used in previous studies. Animations were shown one-by-one and, for each one, participants were asked to choose a label from a list of 14 emotion names ("attraction", "satisfaction", "pride", "hope", "joy", "fascination", "admiration", "disgust", "dissatisfaction", "shame", "fear", "sadness", "boredom", "contempt") or to select "none of the above".

Participants were not asked directly about the valence of the emotion represented by each animation but valence recognition can be assessed through the categorical labels attributed to each animation. Valence recognition was very high as the label selected was the name of an emotion of the same valence as the intended emotion in 95% of all trials. It was precisely the name of the intended emotion in 66% of all trials. The hit rates for the revised animations were 90% (95% CI: 58%, 92%]) for hope and 10% for admiration (95% CI: [3%, 30%]).

Table 4 Results from Study 4 (United States pilot judgement study, N = 20)

Animation	*Hit rate*	*Other choices above 10%*
Attraction	50%	Admiration (20%), Fascination (15%), Hope (10%)
Satisfaction	80%	Pride (15%)
Pride	60%	Satisfaction (30%)
Hope	90%	Shame (10%)
Joy	80%	Admiration (10%)
Fascination	55%	Attraction (10%), Admiration (10%)
Admiration	10%	Joy (30%), Fascination (30%), Fear (15%), Hope (10%)
Disgust	90%	NA
Dissatisfaction	50%	Disgust (20%), Contempt (20%)
Shame	65%	Sadness (15%)
Fear	70%	Disgust (10%), Dissatisfaction (10%), Contempt (10%)
Sadness	95%	NA
Boredom	100%	NA
Contempt	25%	Dissatisfaction (55%), Disgust (15%)

Discussion of studies 3 and 4

Studies 3 and 4 show that the animation redesign was successful in one case (the hope animation) but not the other (admiration). The limited size of the sample limits the precision of the estimate but it was sufficient to establish that the revised hope admiration would be interpreted as intended by at least 60% of potential participants, a noticeable improvement over the results of the original version. Recognition of the new admiration animation seemed somewhat better in a Dutch student sample but it remained dismal in the online general population US sample.

In other respects, these two studies largely confirmed the findings of previous judgement studies, including the very good recognition of valence and the good performance of many animations. At the same time, they also documented the persistence of interpretation problems for several animations, underlining the need for further investigation.

7. Studies 5 and 6: Exploring Unintended Interpretations

All four judgement studies revealed persistent interpretation problems for several of the animations. In particular, fascination, and admiration failed to convey the intended meaning to a sizable majority of the participants in several of the samples. While not as poorly recognized, attraction and pride were also often confused for other emotions.

Interestingly however, confusions were far from random, suggesting that even low performing animations do convey a rather specific affective meaning. This observation also prompts new questions about the interpretation process: Are the meanings of the animations properly captured by the intended labels? Does the lack of agreement between participants reflect stable differences of interpretation between observers or transient errors stemming from contextual factors?

While fixed-choice judgement studies provide a good way to quantify agreement on the meaning of non-verbal representations across different cultures and languages, they provide only limited information about these questions. Judgement studies with slightly different procedures or response formats can provide additional insight into interpretation differences and the ways to address them.

Study 5: Paired Forced-Choice Study

In the paired forced-choice study, animations were presented in pairs with only two possible labels and participants were explicitly instructed that each label must match one and only one of the animations. Additionally, the labels were not limited to a single word and included both emotion names and a short definition of their meaning.

The choice for this particular format stems from the observation that, for several animations, most unintended responses were limited to only one or two alternative labels and that these misinterpretations were often reciprocal. For example, a significant number of participants interpreted the pride animation as a representation of “satisfaction” or the satisfaction animation as a representation of “pride” but very few interpreted either animation as conveying any other meaning from our list of 14 potential labels.

These “clusters” of confusion are open to several competing interpretations. If the confusions reflect the absence of meaningful non-verbal cues for some of the participants, the response format should have little or no effect on the recognition rate. However, if the confusions result from a failure to grasp the nuances of the some emotion words or perhaps a “satisficing” strategy (picking the first acceptable label rather than the best one), the modified format would be expected to promote a better recognition of the intended meaning of each animation.

By presenting two animations side-by-side and playing them several times, the modified procedure should focus the attention of the participants on the nuances of the expression and allow them to make a direct comparison. The use of several words and a definition to describe each emotion should also minimize label interpretation issues, ensuring that the error rate only reflects some irresolvable ambiguity of the animations themselves.

The paired forced-choice study was administered in a classroom setting. The class was taught in English and the sample included both Dutch and international students. In study 5, 79 Master-level Industrial Design Engineering students (mean age 23 ± 2, 36 men) participated concurrently using a pen-and-paper answer booklet while the animations were shown using the room’s presentation system. Each page in the booklet

corresponded to a pair of animations and included a still from each animation (the apex of the expression, typically the last frame in the animation), two words and a description for each emotion (e.g. “Fear/Anxiety – Simon feels threatened by something dangerous”).

Table 5 Results from study 5 (Netherlands paired forced-choice study, N = 79)

Emotion pair	*Hit rate*
Fear/Anxiety – Boredom/Dullness	100%
Pride/Self-esteem – Satisfaction/Approval	94%
Attraction/Desire – Hope/Optimism	97%
Dissatisfaction/Anger – Contempt/Disrespect	97%
Fascination/Curiosity – Admiration/Respect	82%
Dissatisfaction/Anger – Disgust/Aversion	97%
Attraction/Desire – Admiration/Respect	89%
Contempt/Disrespect – Disgust/Aversion	100%

The paired forced-choice response format resulted in near-perfect discrimination between animations, with the possible exception of the admiration – fascination pair (Table 5). Even pairs of animations like pride and satisfaction, or contempt and dissatisfaction that were consistently mislabelled by as many as 20% or 30% of the participants in previous studies were very well discriminated from each other using the modified procedure.

This result shows that far from reflecting a fundamental ambiguity or a complete lack of relevant cues in the animations, unintended interpretations mostly result from the response strategy induced by the judgement procedure or from transient factors like a lack of attention or the ambiguity of the labels.

At the same time, good discrimination between two animations only means that one of the labels is a better fit for one of the animations than the other and that all participants do not perceive the animations in each pair as equivalent in their meaning. Such a result does not imply that any particular label is a perfect description of all the meanings that can be attributed to the animations. It does, however, show that low hit rates in fixed-choice judgement studies do not imply that an animation is completely uninterpretable for some participants, and that even participants who might select another label in some conditions are able to recognize the intended meaning of the animation under other conditions.

Study 6: Free Labelling

Another approach to understand people’s interpretations of the animations is to simply ask for open-ended descriptions of the feelings expressed by the character. While open-ended answers make cross-cultural studies and quantification more difficult, a free labelling task can provide more insight into errors and misinterpretations. It is particularly useful to investigate animations like attraction, which have a poor recognition rate but also no interpretable pattern of confusion. Furthermore, open-ended data can suggest alternative interpretations that were not part of the original emotion list and ensure that agreement between participants on the meaning of the animations is not merely an artefact of the fixed-choice response format.

For this study, 29 US participants were recruited on AMT (mean age 35 ± 13, 11 men) and participated on their own computer. Animations were shown one-by-one, asking participants “what emotion or mood” the character in the cartoon is feeling. Answers were coded according to the list of intended emotions to

compute a hit rate for each animation (Table 6). Other answers were also tallied to identify unintended interpretations.

Table 6 Results from study 6 (United States free labelling study, N = 29)

Animation	*Labels coded as correct*	*Hit rate*	*Other common labels*
Attraction	Wanting, Yearning, Longing	46%	Happy
Satisfaction	Satisfied, Content	59%	Happy
Pride	Proud, Smug, Accomplished	72%	Satisfied
Hope	Hopeful, Anticipation	34%	Praying, Begging
Joy	Happy, Pleased, Delighted	59%	Surprised, Greeting
Fascination	Curious, Inquisitive, Interested	62%	Surprised
Admiration	Wonder, Impressed	3%	Surprised, Excited, Shocked
Disgust	Disgusted, Grossed out, Put off	100%	NA
Dissatisfaction	Angry, Mad, Discontent	79%	*Only idiosyncratic answers*
Shame	Embarrassed, Shy, Aloof, Bashful	76%	Sad
Fear	Scared, Fearful, Afraid	62%	Disgusted, Shocked, Surprised
Sadness	Sad	97%	*Only idiosyncratic answers*
Boredom	Bored	38%	Tired, Sleepy
Contempt	Contempt, Disdainful, Righteous	7%	Angry, Upset, Mad

Recognition rate for some animations was just as good as in the fixed-choice judgement studies, with nearly all participants using exactly the intended name to describe the animation (case of sadness or disgust). Other animations also attained a high level of agreement among participants but not exactly on the intended meaning (as in the case of dissatisfaction described as “anger” or shame described as “embarrassment”).

Of the animations exhibiting the most serious problems in previous studies, fascination, was described with words like “curious” or “interested”, which seem related to the intended emotion, albeit somewhat milder and broader in meaning. Attraction was described with many different words, none of them endorsed by more than four participants, but many broadly related to the idea of desire (e.g. “wanting”, “yearning”, “longing” or even “loving” or “caring”). These answers suggest that the problems encountered with this animation reflect not so much an inability to interpret its affective meaning as difficulties to articulate this meaning and agree on a specific word to describe it. Admiration was not understood at all but participants did in fact all agree on a limited number of alternative interpretations with only three words (“surprised”, “excited” and “shocked”) used by more than one participant. Together, they represent 86% of all answers.

Two animations (hope and boredom) also received a number of descriptions that point not so much to other emotions as to largely non-affective states or behaviours like “tired” or “praying”. The recognition rate for the contempt animation was very low with many idiosyncratic answers (but see Matsumoto and Ekman, 2004 on the difficulties English speakers have with the word “contempt”). Finally, looking at all answers irrespective of the animation that elicited them also reveals two emotion names that were absent from the list of labels used in our previous studies: “surprise” and “interest”.

Discussion of studies 5 and 6

Studies 5 and 6 established that even some animations that did not perform so well in fixed-choice judgement studies are not necessarily as ambiguous as those studies would appear to suggest. The free

labelling task also revealed a high level of agreement on the meaning of most animations including some interpretations that were not included in the original emotion list.

Interestingly, participants frequently mentioned surprise and interest, two states featured in some lists of basic or fundamental emotions even though we did not set out to represent these emotions or to reproduce prototypical expressions described in the literature. Since the relevant animations do appear to convey a clear and recognizable meaning, they could still be used in a non-verbal self-report instrument, albeit not with the interpretation originally ascribed to them at the beginning of the tool's development.

Somewhat more surprisingly, the shame animation was frequently described as "embarrassment" even though it resembles shame expressions described in the literature (Izard, 1971; Tracy and Robins, 2007) but does not feature the smile that was found to be a key element of embarrassment displays and one of the main differences between shame and embarrassment expressions (Keltner, 1995). Other labels for this animation like "shy" suggest that its meaning is very close to the shame/shyness scale included in Izard's Differential Emotions Scale.

8. Studies 7 and 8: Testing Alternative Interpretations

The data collected so far suggest that at least some of the unintended interpretations in previous judgement studies could be the result of the inadequacies of the list of labels. Two further fixed-choice judgement studies were set up to test this hypothesis by altering the label list.

Study 7: Pilot Reinterpretation Study

The pilot reinterpretation study followed the same procedure as study 4 but returned to the double labels used in study 2. The list of potential labels was adjusted based on the open-ended data from study 6 and included in particular "interest/curiosity" (for the fascination animation) and "surprise/shock" (for admiration).

For this study, 20 US participants were recruited on AMT (mean age 32 ± 13, 11 men), who participated on their own computer. For each animation, the participants were asked to pick a label from a list of 14 emotion pairs or "none of the above".

Table 7 Results from study 7 (United States, $N = 20$)

Animation	*Intended label*	*Hit rate*	*Other choices above 10%*
Attraction	Desire/Love	90%	Interest (10%)
Satisfaction	Satisfaction/Approval	65%	Pride (15%), Joy (10%)
Pride	Pride/Self-esteem	65%	Satisfaction (20%), Joy (15%)
Hope	Hope/Optimism	75%	Desire (10%)
Joy	Joy/Happiness	65%	Surprise (20%), Hope (10%)
Fascination	Interest/Curiosity	90%	NA
Admiration	Surprise/Shock	70%	Joy (15%), Interest (10%)
Disgust	Disgust/Aversion	100%	NA
Dissatisfaction	Anger/Dissatisfaction	90%	NA
Shame	Embarrassment/Shyness	95%	NA
Fear	Fear/Anxiety	35%	Surprise (35%), Disgust (15%)
Sadness	Sadness/Grief	100%	NA

Boredom	Boredom/Dullness	100%	NA
Contempt	Contempt/Disrespect	80%	Anger (10%)

Recognition was very good with an 80% hit rate across all trials (detailed results are presented in Table 7). The main source of unexpected answers was the fear animation, which was frequently interpreted as "surprise", possibly because the "surprise" label also included the word "shock", suggesting a negative valence. Arguably, "shock" is in fact closer to fear than to surprise and would better fit the former animation even though it was occasionally used to describe both in the free labelling study. One finding notably at odds with previous results (especially with US participants) is the good performance of the contempt animation but given the limited sample size, it should not be given too much weight.

Study 8: Main Reinterpretation Study

Study 8 followed the same procedure as study 7, with two small changes to the list of labels – "surprise/excitement" instead of "surprise/shock" and "fear/shock" instead of "fear/anxiety" – to better reflect the positive tone intended for the admiration animation.

39 US participants were recruited on AMT (mean age 37 ± 13, 17 men) and participated on their own computer. For each animation, the participants were asked to pick a label from a list of 14 emotion pairs or "none of the above".

Table 8 Results from study 8 (United States, $N = 39$)

Animation	*Intended label*	*Hit rate*	*Other choices above 10%*
Attraction	Desire/Love	72%	NA
Satisfaction	Satisfaction/Approval	87%	Pride (10%)
Pride	Pride/Self-esteem	74%	Satisfaction (13%)
Hope	Hope/Optimism	64%	Desire (13%)
Joy	Joy/Happiness	54%	Surprise (41%)
Fascination	Interest/Curiosity	77%	Surprise (10%)
Admiration	Surprise/Excitement	85%	Joy (10%)
Disgust	Disgust/Aversion	92%	NA
Dissatisfaction	Anger/Dissatisfaction	74%	Contempt (15%)
Shame	Embarrassment/Shyness	77%	Sadness (18%)
Fear	Fear/Shock	90%	NA
Sadness	Sadness/Grief	100%	NA
Boredom	Boredom/Dullness	100%	NA
Contempt	Contempt/Disrespect	51%	Disgust (26%), Anger (18%)

Across all trials, the recognition rate was 78%. The two new labels appeared to better capture the nuances of the admiration and fear animations, as there was no confusion between them. "Surprise/excitement" was, however, also used to describe the joy animation.

9. General Discussion

A series of judgement studies investigated research participants' understanding of a set of animations designed to support non-verbal self-report of feelings. While differing in sample size, these studies found broad agreement across a variety of countries and response formats, especially for negative emotions, but also several deficiencies (hard to interpret animations and confusions). The animations themselves and their descriptions were adjusted based on these findings and the two final studies confirmed that the adjusted labels represent participants' understanding of the animations' affective meaning, at least in a US sample, even if the joy or contempt animations still seem somewhat more difficult to match with the corresponding label.

Such a large international dataset also affords interesting comparisons between cultures and especially between studies 1 and 2. Interestingly, while the overall recognition rate tended to be lower in China, errors were far from random with some animations eliciting essentially identical performance and others being much more challenging for one sample or the other. Thus, the animations for attraction, hope and fascination were difficult to label for Chinese participants while contempt only created difficulties in English-speaking samples (an issue also documented by, e.g. Matsumoto and Ekman, 2004).

Similarly, patterns of confusions between emotions exhibit interesting differences and similarities between the two samples. For example, "Hope" was often the second choice for low-performance animations like attraction and admiration. On the other hand, the Chinese translation for "Attraction" was a frequent choice for several animations including fascination and admiration, which was not the case in the British sample.

However, a major difficulty in interpreting these differences and a serious limitation of judgement studies in general is their reliance on verbal descriptions of the relevant feelings. Whether it is by asking observers to freely produce labels for a set of photographs or prompting them to pick a word from a list, judgement studies of the meaning of non-verbal behaviour are based on the quantification of the match between non-verbal stimuli and words describing the feelings associated with these stimuli.

The ability of research participants in such judgement studies to reliably describe non-verbal stimuli therefore depends on the availability of an appropriate label for each affective state. In effect, these methods work best when all participants perfectly understand all the words used to describe the moods or emotions of interest in the study. If they disagree on the precise meaning or nuances of the labels offered or are unable to understand or to produce the intended (or "correct") word, participants would appear to disagree on the meaning of the non-verbal stimuli even if they are in fact perfectly able to relate it to their own feelings or to specific appraisals, behaviours or eliciting conditions.

This leads to the slightly paradoxical consequence that obtaining perfect recognition scores in judgement studies requires the availability of universally understandable and unambiguous labels. However, if such labels were indeed available, they would also make non-verbal representations unnecessary for measurement purposes, as a verbal scale based on these labels would presumably be entirely sufficient. If, on the other hand, non-verbal expressions are in fact natural representations for affective states and are easier to interpret than words describing these states for many people, one would expect recognition scores in judgement studies to be far from perfect, not because the non-verbal representations are deficient but because the participants in the judgement studies do not interpret the various labels in the same way.

Faced with the complexity of the validation of non-verbal representations of emotion, it could be tempting to eschew non-verbal self-report entirely and stay content with traditional self-report questionnaires. Using verbal scales, however, hardly makes the problem disappear as they also presuppose that research participants are able to understand the emotion names selected by the researchers and agree on their meaning. From this perspective, the development of non-verbal self-report tools does not so much create new problems as reveal fundamental issues that all self-report tools need to overcome.

It follows that neither verbal nor non-verbal representations of emotion can simply be assumed to be well understood by all potential research participants, and the match between non-verbal representations and emotion words in judgement studies can never be expected to be perfect. Furthermore, a failure to match a representation with its intended label can have many causes: lack of clarity in the representation, disagreement regarding the meaning of potential labels, but also differences in "emotional granularity", insufficient attention and a number of other contextual factors.

The fact that we were able to increase recognition rates by using slightly different labels (while staying very close to the original intended appraisal dimension) or to get near-perfect discrimination when pitting two animations against each other (study 5) suggests that very often the words, rather than the animations, create difficulties for the participants.

With this in mind, no single study should be thought of as a final test of the validity of an emotion self-report scale. Rather, the judgement studies presented here are offered as incremental evidence of the meaning of the non-verbal representations used in PrEmo2. Among the various approaches presented here, the use of richer verbal descriptions of affect based on stories/vignettes (study 5) rather than single labels or adjectives seems like a particularly promising type of evidence. Additional studies with such stimuli would be especially valuable.

10. Conclusion

This paper presented the development and empirical evaluation of a new set of animations designed to become part of a non-verbal self-report tool for the measurement of emotion. Several studies established that most of these animations were able to convey a distinctive affective meaning across several different settings and populations. Revision of some of the animations and a small reinterpretation of the meaning of others was also able to address most of the issues identified in the earlier studies.

While we did not succeed in creating a representation for all the emotion words we initially set out to include in the set, we were able to create a set of animations covering a wide range of emotions of interest to design researchers and practitioners but also in psychology and many other related fields. While we have not been able to test the last version in all countries included in the research, these animations certainly form one of the most thoroughly tested and best recognized sets of non-verbal representations of emotion and the only one to differentiate so many positively toned emotions. Future work could extend this validation further to other populations and should compare the animations' usefulness in actual self-report situations.

References

Bould, E., Morris, N. and Wink, B. (2008) 'Recognising subtle emotional expressions: The role of facial movements', Cognition and Emotion, Vol. 22, No. 8, pp. 1569-1587.

Bradley, M.M. and Lang, P.J. (1994) 'Measuring emotion: The self-assessment manikin and the semantic differential', Journal of Behavior Therapy and Experimental Psychiatry, Vol. 25, No. 1, pp. 49-59.

Broekens, J. and Brinkman, W.P. (2013) 'Affect button: a method for reliable and valid affective self-report', International Journal of Human-Computer Studies, Vol. 71, No. 6, pp.641–667.

Desmet, P.M.A. (2002) Designing Emotions, PhD Thesis, Delft University of Technology.

Desmet, P.M.A. (2003) 'Measuring emotion; development and application of an instrument to measure emotional responses to products', in Blythe, M.A., Monk, A.F., Overbeeke, K., and Wright, P.C. (ed.), Funology: from Usability to Enjoyment, pp. 111-123, Kluwer Academic Publishers, Dordrecht.

Desmet, P.M.A. (2008). 'Product emotion', in Hekkert, P. and Schifferstein, H.N.J. (eds.), Product Experience, pp. 379-397, Elsevier, Amsterdam.

Desmet, P.M.A. and Dijkhuis, E.A. (2003) 'Wheelchairs can be fun: a case of emotion-driven design', Proceedings of the International Conference on Designing Pleasurable Products and Interfaces, 23-26 June 2003, Pittsburgh, Pennsylvania, USA, ACM publishing, New York.

Desmet P.M.A., Porcelijn, R. and van Dijk, M. (2007) 'Emotional design; application of a research based design approach', Journal of Knowledge, Technology & Policy, Vol. 20, No. 3, pp. 141-155.

Desmet, P.M.A. and Schifferstein, N.J.H. (2012) 'Emotion research as input for product design', in Beckley, J., Paredes, D. and Lopetcharat, K. (ed.), Product Innovation Toolbox: A Field Guide to Consumer Understanding and Research, pp. 149-175, John Wiley & Sons, Hoboken, NJ.

Desmet, P.M.A., Vastenburg, M.H. and Romero Herrera, N. (2016) 'Mood Measurement with Pick-A-Mood: Review of current methods and design of a pictorial self-report scale', Journal of Design Research, Vol. 14, No. 3, pp. 241-279.

Ekman, P. and Friesen, W.V. (1975) 'Unmasking the face: A guide to recognizing emotions from facial cues', Prentice-Hall, Englewood Cliffs.

Fokkinga, S.F. and Desmet, P.M.A. (2012). 'Darker shades of joy: The role of negative emotion in rich product experiences', Design Issues, Vol. 28, No. 4, pp. 42-56.

Fujimura, T. and Suzuki, N. (2010) 'Effects of dynamic information in recognising facial expressions on dimensional and categorical judgments', Perception, Vol. 39, No. 4, pp. 543-552.

Huisman, G. and Van Hout, M. (2010) 'The development of a graphical emotion measurement instrument using caricatured expressions: the LEMtool', in Peter, C., Crane, E., Fabri, M., Agius, H. and Axelrod, L. (ed.), Emotion in HCI – Designing for People. Proceedings of the 2008 International Workshop, pp. 5-8, Fraunhofer, Rostock, Germany.

Izard, C.E. (1971) The face of emotion, Appleton-Century-Crofts, New York.

Izard, C.E. (1972) Patterns of emotions: A new analysis of anxiety and depression, Academic Press, San Diego, CA.

Keltner, D. (1995) 'Signs of Appeasement: Evidence for the Distinct Displays of Embarrassment, Amusement, and Shame', Journal of Personality and Social Psychology, Vol. 68, No. 3, pp. 441-454.

Lorr, M. and Wunderlich, R.A. (1988) 'A semantic differential mood scale', Journal of Clinical Psychology, Vol. 44, No. 1, pp. 33-36.

Manassis, K., Mendlowitz, S., Kreindler, D., Lumsden, C., Sharpe, J., Simon, M.D., Woolridge, N., Monga, S. and Adler-Nevo, G. (2009) 'Mood assessment via animated characters: A novel instrument to evaluate feelings in young children with anxiety disorders', Journal of Clinical Child and Adolescent Psychology, Vol. 38, No. 3, pp. 380-389.

Matsumoto, D. and Ekman, P. (2004) 'The relationship among expressions, labels, and descriptions of contempt', Journal of Personality and Social Psychology, Vol. 87, No. 4, pp. 529-540.

Ortony, A., Clore, G.L. and Collins, A. (1988) The cognitive structure of emotions, Cambridge University Press, New York.

Rosset, D.B., Rondan, C., Da Fonseca, D., Santos, A., Assouline, B. and Deruelle, C. (2008), 'Typical emotion processing for cartoon but not for real faces in children with autistic spectrum disorders', Journal of Autism and Developmental Disorders, Vol. 38, No. 5, pp. 919-925.

Sauter, D. (2010) 'More than happy: The need for disentangling positive emotions', Current Directions in Psychological Science, Vol. 19, No. 1, pp. 36-40.

Sauter, D.A., Eisner, F., Ekman, P. and Scott, S.K. (2010). 'Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations', Proceedings of the National Academy of Sciences of the United States of America, Vol. 107, No. 6, pp. 2408-2412.

Scherer, K.R. (2005) 'What are emotions? And how can they be measured?', Social Science Information, Vol. 44, No. 4, pp. 695-729.

Stone, B.A.S. (2006) 'A Nonverbal Pictorial Vocabulary of Feelings', 114th Annual Convention of the American Psychological Association, New Orleans, Louisiana, August 2006.

Tracy, J.L., and Robins, R.W. (2007) 'The Prototypical Pride Expression: Development of a Nonverbal Behavior Coding System', Emotion, Vol. 7, No. 4, pp. 789-801.

Watson, D., and Clark, L.A. (1994) Manual for the Positive and Negative Affect Schedule – Expanded Form.

Watson, D., Clark, L.A. and Tellegen, A. (1988) 'Development and validation of brief measures of positive and negative affect: the PANAS scales', Journal of Personality and Social Psychology, Vol. 54, No. 6, pp. 1063-1070.

Yoon, J., Pohlmeyer, A. E. and Desmet, P.M.A. (2014). 'Nuances of emotions in product development: Seven key opportunities identified by design professionals', in DS 77: Proceedings of the DESIGN 2014 13th International Design Conference, 19-22May 2014, Dubrovnik, Croatia, pp. 643-652, The Design Society, Glasgow.

Youngstrom, E.A., and Green, K.W. (2003) 'Reliability generalization of self-report of emotions when using the differential emotions scale', Educational and Psychological Measurement, Vol. 63, No. 2, pp. 279-295.

Zuckerman, M. (1960) 'The development of an affect adjective check list for the measurement of anxiety', Journal of Consulting Psychology, Vol. 24, No. 5, pp. 457-462.

Zuckerman, M., Lubin, B., and Rinck, C.M. (1983) 'Construction of new scales for the multiple affect adjective check list', Journal of Behavioral Assessment, Vol. 5, No. 2, pp. 119-129.