Computational approaches for emotion research

Jan Niklas Schneider

Dissertation zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

in der Wissenschaftsdisziplin Angewandte Informatik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät

Institut für Informatik und Computational Science

der Universität Potsdam

Tag der Disputation: 02.03.2020

Betreuer:
Prof. Dr. Ulrike Lucke
Prof. Dr. Tim Landgraf

Gutachter:
Prof. Dr. Ulrike Lucke
Dr. Timothy Brick
Prof. Dr. Mathias Weymar

Published online at the
Institutional Repository of the University of Potsdam:
https://doi.org/10.25932/publishup-45927
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-459275

Acknowledgements

I would like to express my sincere gratitude to my first supervisor Prof. Dr. Ulrike Lucke, who provided me with a great deal of support and assistance, who was always available in times of need and who reliably helped me solve countless problems in my time as a PhD student. I thank Prof. Dr. Tim Landgraf for taking on the role of second supervisor and Prof. Dr. Isabel Dziobek for being my mentor.

My sincere thanks goes to Dr. Timothy Brick, who has always provided me with excellent advice, with whom I had many great discussions and work sessions and who has taught me innumerable things about research and much else. Thank you! You have been a tremendous scientific inspiration to me.

I thank my colleagues for the pleasant work environment and the good times we had in- and outside of the office.

I would like to thank my parents who have always supported me throughout my academic pursuits and throughout my life. Last but not least, I thank my friends (you know who you are!) for great company, good talks and fun times.

Zusammenfassung

Emotionen sind ein zentrales Element menschlichen Erlebens und spielen eine wichtige Rolle bei der Entscheidungsfindung. Diese Dissertation identifiziert drei methodische Probleme der aktuellen Emotionsforschung und zeigt auf, wie diese mittels computergestützter Methoden gelöst werden können. Dieser Ansatz wird in drei Forschungsprojekten demonstriert, die die Entwicklung solcher Methoden sowie deren Anwendung auf konkrete Forschungsfragen beschreiben.

Das erste Projekt beschreibt ein Paradigma welches es ermöglicht, die subjektive und objektive Schwierigkeit der Emotionswahrnehmung zu messen. Darüber hinaus ermöglicht es die Verwendung einer beliebigen Anzahl von Emotionskategorien im Vergleich zu den üblichen sechs Kategorien der Basisemotionen. Die Ergebnisse deuten auf eine Zunahme der Schwierigkeiten bei der Wahrnehmung von Emotionen mit zunehmendem Alter der Darsteller hin und liefern Hinweise darauf, dass junge Erwachsene, ältere Menschen und Männer ihre Schwierigkeit bei der Wahrnehmung von Emotionen unterschätzen. Weitere Analysen zeigten eine geringe Relevanz personenbezogener Variablen und deuteten darauf hin, dass die Schwierigkeit der Emotionswahrnehmung vornehmlich durch die Ausprägung der Wertigkeit des Ausdrucks bestimmt wird.

Das zweite Projekt zeigt am Beispiel von Arousal, einem etablierten, aber vagen Konstrukt der Emotionsforschung, wie Face-Tracking-Daten dazu genutzt werden können solche Konstrukte zu schärfen. Es beschreibt, wie aus Face-Tracking-Daten Maße für die Entfernung, Geschwindigkeit und Beschleunigung von Gesichtsausdrücken berechnet werden können. Das Projekt untersuchte wie diesen Maße mit der Arousal-Wahrnehmung in Menschen mit und ohne Autismus zusammenhängen. Der Abstand zum Neutralgesicht war prädiktiv für die Arousal-Bewertungen in beiden Gruppen. Die Ergebnisse deuten auf eine qualitativ ähnliche Wahrnehmung von Arousal für Menschen mit und ohne Autismus hin.

Im dritten Projekt stellen wir die Partial-Least-Squares-Analyse als allgemeine Methode vor, um eine optimale Repräsentation zur Verknüpfung zweier hochdimensionale Datensätze zu finden. Das Projekt demonstriert die Anwendbarkeit dieser Methode in der Emotionsforschung anhand der Frage nach Unterschieden in der Emotionswahrnehmung zwischen Männern und Frauen. Wir konnten zeigen, dass die emotionale Wahrnehmung von Frauen systematisch mehr Varianz der Gesichtsausdrücke erfasst und dass signifikante Unterschiede in der Art und Weise bestehen, wie Frauen und Männer einige Gesichtsausdrücke wahrnehmen. Diese konnten wir als dynamische Gesichtsausdrücke visualisieren. Um die

Anwendung der entwickelten Methode für die Forschungsgemeinschaft zu erleichtern, wurde ein Software-Paket für die Statistikumgebung R geschrieben. Zudem wurde eine Website entwickelt (thisemotiondoesnotexist.com), die es Besuchern erlaubt, ein Partial-Least-Squares-Modell von Emotionsbewertungen und Face-Tracking-Daten interaktiv zu erkunden, um die entwickelte Methode zu verbreiten und ihren Nutzen für die Emotionsforschung zu illustrieren.

Abstract

Emotions are a central element of human experience. They occur with high frequency in everyday life and play an important role in decision making. However, currently there is no consensus among researchers on what constitutes an emotion and on how emotions should be investigated. This dissertation identifies three problems of current emotion research: the problem of ground truth, the problem of incomplete constructs and the problem of optimal representation. I argue for a focus on the detailed measurement of emotion manifestations with computer-aided methods to solve these problems. This approach is demonstrated in three research projects, which describe the development of methods specific to these problems as well as their application to concrete research questions.

The problem of ground truth describes the practice to presuppose a certain structure of emotions as the a priori ground truth. This determines the range of emotion descriptions and sets a standard for the correct assignment of these descriptions. The first project illustrates how this problem can be circumvented with a multidimensional emotion perception paradigm which stands in contrast to the emotion recognition paradigm typically employed in emotion research. This paradigm allows to calculate an objective difficulty measure and to collect subjective difficulty ratings for the perception of emotional stimuli. Moreover, it enables the use of an arbitrary number of emotion stimuli categories as compared to the commonly used six basic emotion categories. Accordingly, we collected data from 441 participants using dynamic facial expression stimuli from 40 emotion categories. Our findings suggest an increase in emotion perception difficulty with increasing actor age and provide evidence to suggest that young adults, the elderly and men underestimate their emotion perception difficulty. While these effects were predicted from the literature, we also found unexpected and novel results. In particular, the increased difficulty on the objective difficulty measure for female actors and observers stood in contrast to reported findings. Exploratory analyses revealed low relevance of person-specific variables for the prediction of emotion perception difficulty, but highlighted the importance of a general pleasure dimension for the ease of emotion perception.

The second project targets the problem of incomplete constructs which relates to vaguely defined psychological constructs on emotion with insufficient ties to tangible manifestations. The project exemplifies how a modern data collection method such as face tracking data can be used to sharpen these constructs on the example of arousal, a long-standing but fuzzy construct in emotion research. It describes how measures of distance, speed and magnitude of acceleration can be computed from face tracking data and investigates their

intercorrelations. We find moderate to strong correlations among all measures of static information on one hand and all measures of dynamic information on the other. The project then investigates how self-rated arousal is tied to these measures in 401 neurotypical individuals and 19 individuals with autism. Distance to the neutral face was predictive of arousal ratings in both groups. Lower mean arousal ratings were found for the autistic group, but no difference in correlation of the measures and arousal ratings could be found between groups. Results were replicated in a high autistic traits group consisting of 41 participants. The findings suggest a qualitatively similar perception of arousal for individuals with and without autism. No correlations between valence ratings and any of the measures could be found which emphasizes the specificity of our tested measures for the construct of arousal.

The problem of optimal representation refers to the search for the best representation of emotions and the assumption that there is a one-fits-all solution. In the third project we introduce partial least squares analysis as a general method to find an optimal representation to relate two high-dimensional data sets to each other. The project demonstrates its applicability to emotion research on the question of emotion perception differences between men and women. The method was used with emotion rating data from 441 participants and face tracking data computed on 306 videos. We found quantitative as well as qualitative differences in the perception of emotional facial expressions between these groups. We showed that women's emotional perception systematically captured more of the variance in facial expressions. Additionally, we could show that significant differences exist in the way that women and men perceive some facial expressions which could be visualized as concrete facial expression sequences. These expressions suggest differing perceptions of masked and ambiguous facial expressions between the sexes. In order to facilitate use of the developed method by the research community, a package for the statistical environment R was written. Furthermore, to call attention to the method and its usefulness for emotion research, a website was designed that allows users to explore a model of emotion ratings and facial expression data in an interactive fashion.

Contents

1. Introduction 1

1.1 What Are Emotions? 2

1.2 Biological Determinism, Constructivism and the Question of Natural Kinds 3

1.3 Theories of Emotion 5

1.3.1 Basic emotion theory 6

1.3.2 Criticism of basic emotion theory 7

1.3.3 Theory of constructed emotion 8

1.3.4 Criticism of the theory of constructed emotion 10

1.4 Psychological Constructs and their Operationalization 12

1.5 Current Methodological Problems of Emotion Research 14

1.6 A Focus on Measurement 17

1.7 Benefits of Using Computer-aided Methods for Emotion Research 19

1.8 Challenges of Using Computer-aided Methods for Emotion Research 21

1.9 Research Projects 22

2. Project 1: Difficulty of Emotional Expression Perception 24

2.1 Research Motivation 24

2.2 Limitations of Common Basic Emotion Recognition Paradigms 25

2.2.1 Forced-choice and basic emotion framework 25

2.2.2 Stimuli in emotion recognition paradigms 26

2.2.3 The question of ground truth 27

2.3 Difficulty of Emotion Perception 28

2.4 Relationship of Subjective and Objective Difficulty 28

2.5 Measures of Subjective and Objective Difficulty 29

2.6 Person- and Stimuli-specific Predictors of Emotion Perception Difficulty 29

2.6.1 Age and sex of actor and observer 29

2.6.2 Valence and arousal of stimuli 30

2.7 Hypotheses 30

2.8 Methods 31

2.8.1 Participants 31
2.8.2 Materials 32
2.8.3 Measures 33
2.8.4 Procedure 34
2.8.5 Data analysis 34
2.9 Results 37
2.9.1 Assumption check and confirmatory analyses 37
2.9.2 Exploratory analyses 41
2.10 Discussion 44
2.10.1 Age and sex of actor 44
2.10.2 Age and sex of observer 45
2.10.3 Stimulus valence and arousal 47
2.10.4 Effect sizes and feature importance of predictors 48
2.11 General Discussion 49
2.12 Limitations 50
3. Project 2: Arousal Perception from Facial Expressions 51
3.1 Research Motivation 51
3.2 Arousal Cues from Facial Expressions 52
3.3 Face Processing in Autism Spectrum Disorder 53
3.4 Perception of Movement in Autism Spectrum Disorder 54
3.5 Study 1: Measure Selection 54
3.6 Methods 55
3.6.1 Materials 55
3.6.2 Measures 55
3.6.3 Data analysis 59
3.7 Exploratory Results and Discussion 59
3.8 Study 2: Predictors of Arousal 61
3.9 Methods 62
3.9.1 Participants 62
3.9.2 Materials 63
3.9.3 Data analyses 63

3.9.4 Power estimation 64

3.10 Confirmatory Results 65

3.10.1 Within-group results 65

3.10.2 Between-group results 67

3.10.3 Specificity of predictors: analyses of valence 68

3.11 Exploratory Results 68

3.11.1 Individual examination of predictors 68

3.11.2 Replication of results on a sample of individuals with high autistic traits 70

3.12 Discussion 71

3.12.1 Predictors of arousal 71

3.12.2 Group differences in arousal perception 72

3.12.3 Specificity of arousal predictors 73

3.12.4 Outlier videos 73

3.12.5 Design choice and future work 74

3.12.6 Limitations 75

4. Project 3: Relationship Between Facial Expressions and Emotion Perception 76

4.1 Research Motivation 76

4.2 Emotion Representations 77

4.2.1 Representations of facial expressions 77

4.2.2 Representation of emotion impression 79

4.2.3 Finding suitable representations 79

4.3 Introduction to Partial Least Squares Analysis 80

4.3.1 Statistical inference with PLS 81

4.3.2 PLS as a tool for research on emotional facial expression perception 81

4.4 Methods 82

4.4.1 Participants 82

4.4.2 Materials 82

4.4.3 Procedure 83

4.4.4 Data analysis 83

4.5 Results 85

4.5.1 Descriptive statistics 85

4.5.2 Number of significant latent variables 88

4.5.3 Explained variance 90

4.5.4 Loading patterns of within-group models 91

4.5.5 Loading patterns of between-group model 94

4.6 Discussion 97

4.6.1 Hypothesis 1: quantitative differences in emotion perception 97

4.6.2 Hypothesis 2: qualitative differences in emotion perception 98

4.6.3 Limitations and outlook 101

4.7 Dissemination of Methodological Developments 102

4.7.1 Partial least squares statistic library for R 102

4.7.2 Interactive website 102

5. Thesis Conclusion 104

5.1 Contribution to the Methods of Emotion Research 104

5.1.1 Problem of ground truth 104

5.1.2 Problem of incomplete constructs 105

5.1.3 Problem of optimal representation 107

5.1.4 Method reusability 108

5.2 Outlook: Quantification – the Way Forward in Psychology? 109

6. References 111

7. Supplementary Materials 133

List of Figures

Figure 1. Core affect space 10
Figure 2. Six exemplary still frames of the video data set. 33
Figure 3. Plots visualizing the relationship of valence and arousal with both difficulty measures. 42
Figure 4. Feature importance ranking for the prediction of SRD (a) and OD (b). 44
Figure 5. Actor mean face (upper row) and two clip neutral faces are shown for three actors 56
Figure 6. Correlation matrix of measures for potential arousal cues. 61
Figure 7. Added variable plots 67
Figure 8. Correlation matrix for ratings of women (upper triangle) and men (lower triangle). 87
Figure 9. P-Values of the singular values of the latent variables 89
Figure 10. Explained variance of rating (left) and facial expression datasets 90
Figure 11. Effect of the significant latent variables (LV) on the face side. 93
Figure 12. Visualization of the first difference LV. 95
Figure 13. Biplot for the difference PLS model. 96
Figure 14. View of the interactive website thisemotiondoesnotexist.com. 103

List of Tables

Table 1. Mixed effects models for subjective and objective difficulty predicted by actor sex and age. 38

Table 2. Mixed effects models for subjective and objective difficulty predicted by observer sex and age..... 39

Table 3. Mixed effects models for subjective and objective difficulty predicted by valence and arousal ratings. 40

Table 4. Regression models for the NT and ASD group predicting arousal ratings 66

Table 5. Regression models predicting differences in arousal ratings 68

Table 6. Regression models for the NT and ASD group predicting arousal ratings 69

Table 7. Regression models for the NT and ASD group predicting arousal ratings 70

Table 8. Loadings of the PLS models for female and male group on the rating data side. 92

1. Introduction

During my studies I was sitting in a seminar on computational mathematics with a couple of friends and someone had said something along the lines of "that is as subjective as emotions" as a sort of dismissal of a previous statement. One of my friends was puzzled by this and, with a confused look on his face, asked out loud "what are emotions?" To our great amusement he then began to look up "emotions" on the internet. Our group of friends burst into laughter when he started reading the German Wikipedia article on emotions with great interest. It was just too comical; the stereotypical image of the computer scientist who does not know what emotions are, to the point where he needs to look up the meaning in an online encyclopedia. Of course, my friend knew emotions from personal experience, but what he was really interested in (and what we willfully ignored for the sake of amusement) was a scientific definition of the phenomenon. As it turns out, this is complicated. Fehr and Russell (1984, p. 464) have stated: "Everyone knows what an emotion is, until asked to give a definition. Then, it seems, no one knows." So it looks like everyone could have been in the position of my friend when put on the spot. This is peculiar, because emotions seem to be a central element of human experience (Lazarus, 1991). They occur with high frequency (Trampe, Quoidbach, & Taquet, 2015) in everyday life and play an important role in decision making (Bechara, 2000; Schwarz, 2000). Then, how come we cannot easily explain what they are? Another peculiar bit from the start of this episode is the uttered assumption that emotions are subjective and therefore intangible. Is that so? Does that mean that emotions cannot be measured and researched?

This dissertation will object these deliberations and approach the subject of emotion research from a computational perspective. It will show that the question of what emotions are is closely linked with the question of how they can be investigated appropriately. Emotions are transient phenomena, which require specific practices in order to measure and investigate them. However, research on emotions is complicated by disagreement on the appropriate theoretical frameworks. This dissertation identifies underlying methodological issues of the field of emotion research and describes the development of novel methods which provide solutions for these issues. The thesis argues for a focus on detailed measurement through the application of modern, computer-aided methods of data collection and analysis in order to advance emotion research. A major challenge lies in the appropriate analysis of the data generated by these methods. The thesis describes such approaches in detail and demonstrates that these methods can provide solutions to the stated problems by generating novel results in the context of specific psychological research questions.

First, I introduce the phenomenon of emotions and discuss current theories of emotion as well as their limitations. Second, I identify three fundamental problems of current emotion research and outline a general approach of using computer-aided methods to solve these problems. Three research projects are introduced that exemplify solutions to one of the problems each. Third, I present the implementation of the research projects and their results. Finally, the methodological approaches are evaluated in terms of the novel results generated with them and in terms of their utility and potential for reusability in future research.

1.1 What Are Emotions?

Emotions have been a subject of empirical psychological research for more than a century. For example, Wundt, often called the founding father of experimental psychology (Kim, 2016), already formulated theories on emotions (Wundt, 1908). However, to the present day there is no definition for the phenomenon emotion on which emotion researchers can agree (Izard, 2010; Mulligan & Scherer, 2012; Scherer, 2005). Scherer (2005) called the definition of the term "emotion" a "notorious problem". Already more than two decades before, Kleinginna and Kleinginna (1981) had reviewed 92 definitions of emotion and sorted them into 11 categories, each of which emphasized other aspects of the phenomenon, which should illustrate the multiplicity of perspectives on emotions. Finally, Mulligan and Scherer (2012) recently expressed "little hope that there ever will be agreement on a common definition of emotion" attributing it to "sacred traditions" and the "egos of the scholars" (p. 345) involved in the research field.

While a considerable spread in opinions on what defines an emotion is indeed apparent in the literature, commonalities can be observed, too. Two prominent figures of emotion research, Izard and Ekman, conducted two independent surveys among emotion researchers (Ekman, 2016; Izard, 2010). Both came to the conclusion that considerable agreement among researchers about "emotion activation, functions and regulation" (Izard, 2010) and the signals tied to emotions (Ekman, 2016) exist. It seems that most researchers agree on many of the components that should have at least some relevance for the phenomenon. Instead, separate views emerge from the way in which these components and the relation between them are defined as well as from the weight and necessity of individual components. Therefore, here I give a loose definition of emotions for the context of this thesis by naming components of the phenomenon deemed important and mentioned by most authors:

Emotions are psychological and physiological states, which are triggered as a reaction to events or situations, which are coupled with bodily processes, such as neuronal and hormonal activities and with cognitive processes, such as perceptual effects, appraisal and labeling, and which affect behavior, decision making and interaction with the environment often in a goal-directed and adaptive way (Barrett, 2007; Cabanac, 2002; Damasio, 1998; Ekman, 1992a).

Other terms related to emotions and often encountered in the research are: *mood, affect* and *feeling*. Similar to the term "emotion" no clear definition can be given, but I will try to distinguish these terms from emotion. Moods are commonly distinguished from emotions by a slower on- and offset, longer duration and the lack of a triggering event. Whereas emotions follow their trigger immediately and might vanish within seconds to minutes, causes for mood are either not identifiable or not immediately related in a temporal sense. In this sense, moods might just appear and eventually disappear without any apparent reason. Moods therefore seem to be more general and unspecific states (Ekkekakis, 2012). Affect and feeling are both broad umbrella terms that might encompass emotion and mood or refer to their precursors. Both are used to describe general reactive cognitive and physiological states and phenomena (Munezero, Montero, Sutinen, & Pajunen, 2014). However, feeling is often used for the description of subjective experiences. Affect is more widely used in the psychological literature and relates to "anything that is emotional" (Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012, p. 124). The following sections give an overview over the major schools of thoughts in emotion research and their preferred theories of emotion.

1.2 Biological Determinism, Constructivism and the Question of Natural Kinds

Apart from the agreement on the components that make up emotion two broad but conflicting notions can be identified in the literature on emotion. One is founded on the belief that emotions are primarily biologically determined phenomena, called biological determinism, whereas the other assumes emotions to be mainly socially constructed, cultural phenomena and thus primarily learned, called constructivism. In a way, this is the nature-nurture debate of emotion psychology. While the biological and the constructivist viewpoint are often pitted against each other as polar opposites, it has to be mentioned, that most proponents of either view acknowledge at least some contribution to the phenomenon of the respective other side.

At the core of the conflict between biological determinism and constructivism lies the question of whether emotions are so-called *natural kinds* or not (Barrett, 2006). Natural kinds is a term from philosophy and describes concepts that relate to real-world phenomena that have objective qualities independent of human attribution or knowledge (Quine, 1969). For example, the word "gorilla" describes a category for a specific animal. A gorilla has certain qualities, for example, a stocky body with four limbs as well as black and silverish fur. These qualities and the entity itself are not subject to human conventions and will persists independent of human attention to and reasoning about the concept. Hence gorilla (and any other species) could be a natural kind. In this sense, the word "gorilla" labels a category that is found in reality. This category is distinct from other categories like, for example, the category with the label "horse". Conversely, kinds, which are not natural kinds are "artificial" concepts defined by human conventions and attribution, whose existence and specific manifestation depends on and changes with their human definitions. Money as a category could be an example for such a non-natural kind. What is accepted as money is completely up to human convention. If humans collectively decided to exclusively allow potatoes as payment, but only on Sundays, it would radically alter the phenomenon itself. If humans collectively decided to stop believing in money, the phenomenon would cease to exist. Because the phenomenon is not independent of its human definition, it cannot be a natural kind. Several definitions of natural kinds exist. Barrett (2006) mentions the following two: natural kinds can either be defined by a common source, cause or mechanism that produces all instances of a kind or by a shared set of properties essential to each instance of the kind.

Relating this back to the subject of emotion, biological determinists believe in some underlying natural source of emotions or in shared properties among all instances of each emotion (or both). Hence, they believe that emotions are natural kinds and that human-made emotion categories align with the true structure, the "essence", of the phenomenon. This viewpoint usually encompasses two nested assumptions: First, that there are emotions that can be distinctly separated from other bodily or cognitive states, such as hunger or sleep, based on their special features and, second, that there are distinguishable subcategories within the supercategory "emotion". In contrast, constructivists believe that emotions are arbitrary labels for general bodily or cognitive states, which cannot be distinguished cleanly from one-another by means of any strict set of criteria. Hence, they believe emotions not to be natural kinds and that there is no essential structure or common cause to be discovered. They believe that the

definition of the supercategory "emotion" is arbitrary as well as the subcategories for specific emotions.

1.3 Theories of Emotion

Despite the lack of a universal definition of emotions and disagreement on the nature of emotion, a tremendous amount of research has been conducted on emotions. Google Scholar (17th July 2019) lists approximately 66,700 publications from the last 100 years that contain the word "emotion" in the title. Based on different views on emotion, several so-called theories of emotion were developed. A theory of emotion determines how emotions are conceptualized, i.e. which attributes are relevant to describe them and might also describe how individual emotions are distinguished from one another. By this, a theory of emotion postulates a certain structure of emotions and will typically propose a way to sort emotions into categories or to organize them alongside certain dimensions. Hence, theories of emotion provide a theoretical framework for experimental research on emotions.

In the field of psychology two theories of emotion are especially prevalent in the empirical research on emotions. These are Ekman's *basic emotion theory* (Ekman, 1992a) and the *theory of constructed emotion* developed by Russel and Barrett (Russell & Barrett, 1999). They are prominent representatives of the biological and constructivist view on emotions respectively and are frequently used in experimental research. These two theories of emotion will be explained in detail in the following. There are other theoretical emotion frameworks, which differ substantially from the biological and constructivist accounts of emotion. There is, for example, Scherer's component process model (Scherer, 2009), which describes emotion as a complex interplay of a myriad of cognitive and bodily processes across different levels of appraisal. There are also functional accounts of emotions (Campos, Mumme, Kermoian, & Campos, 1994; Dacher Keltner & Gross, 1999), which describe emotions as solutions to social and physical problems and focus on their beneficial components for achieving goals. However, these theories are rarely applied in empirical research, potentially because, so far, novel ways of measuring emotions have not been proposed by these theories. This might be due to the complexity of the accounts and the sheer number of emotion components described within these theories, which complicates the development of standardized measurement procedures. Additionally, some accounts explicitly oppose the definition of a fixed set of measurement procedures (Campos et al., 1994) with the reasoning that this would not do justice to the flexible ways of emotion manifestation.

1.3.1 Basic emotion theory

Darwin (1872) postulated universally expressed and understood facial expressions in his work "The Expression of the Emotions in Man and Animals" as a consequence of inherited capabilities for emotion that were shaped by the same evolutionary processes that he previously described in his seminal works on evolution (Darwin, 1859, 1871). Darwin was, thus, one of the first to put forward a biological basis of emotions.

Ekman and Friesen (1971) made a case for the universality of some facial expressions by showing that the same facial expressions traditionally associated with certain emotions in Western and Eastern cultures were also associated with the same emotions in a remote and pre-literate population in New Guinea, which had minimal exposure to other cultures. This research was conducted after previous efforts (Ekman, Sorenson, & Friesen, 1969; Izard, 1968) to show universality by comparing Western and Eastern cultures were rejected as insufficient, because of the frequent intercultural exchange and prevalence of mass-media in these populations. Both factors could have theoretically allowed for the spread of common emotion conceptualizations that were learned as opposed to innate.

Later, Ekman compiled his empirical results into a theory of emotion that he labeled "basic emotions" (Ekman, 1992a, 1992c). It described six discrete emotion categories (happiness, sadness, anger, fear, surprise and disgust) that are universally expressed and recognized in humans. Ekman's definition of "basic" specifically refers to discrete categories that are shaped by evolutionary adaptations (Ekman & Cordaro, 2011) and therefore are grounded in biology. Discreteness refers to properties of emotions that enable a categorical distinction between them in the sense of natural kinds, i.e. these categories relate to separable phenomena of the real world. Each basic emotion category is associated with specific physiological changes, expressive signals and antecedent events that trigger it. For example, anger might be triggered in a person if they observe harm being inflicted onto a person they care about. This might result in an increase in heart-rate and blood pressure as well as elevated blood levels of the hormones adrenaline and noradrenaline (Stemmler, 2004; Stemmler, Aue, & Wacker, 2007). Anger might then be visibly expressed in the face through lowered and drawn-together eye brows and tightened lips (Ekman, 1992c).

According to Ekman (1992a), instead of a single affective state each basic emotion category describes a family of related affective states, where each member of a family shares certain characteristics with all other members of its family and which distinguish it from members of the other emotion families. This can be extended to expressions of emotions so

that, for example, all angry facial expressions share an activation of the same facial muscles, whereas they might differ in the activation of additional muscles. Originally, six basic emotions were postulated by Ekman (Ekman, 1992c). However, other candidates for additional basic emotions with distinct and potentially universal expression patterns have been proposed, for example, contempt (Ekman & Friesen, 1986; Ekman & Heider, 1988; Matsumoto, 1992), amusement (Dacher Keltner, 1995; Dacher Keltner & Bonanno, 1997), shame, embarrassment and guilt (D Keltner & Buswell, 1996).

1.3.2 Criticism of basic emotion theory

Basic emotion theory has often been criticized for its discrete categories. Here, a major criticism is that no corresponding discrete states can be found in measurements of emotion in physiological (Norman, Necka, & Berntson, 2016) or brain imaging data (Lindquist et al., 2012), since induced basic emotions produce overlapping patterns in these data. Instead, these data rather seem to support broad emotional dimensions, such as general pleasantness (Mauss & Robinson, 2009), than discrete categories. Furthermore, research shows that even blends of basic emotion categories, for example, happily-surprised or angrily-disgusted, can be reliably produced and distinguished from one-another in a categorical sense (Du, Tao, & Martinez, 2014) or experienced at the same time (Hemenover & Schimmack, 2007) and that participants describe even prototypical emotion stimuli on multiple continuous dimensions, if given the choice (Hall & Matsumoto, 2004; Phillips & Allen, 2004; Riediger, Voelkle, Ebner, & Lindenberger, 2011).

The universality of the correspondence of facial expressions and experienced basic emotions has been questioned and the role of culture for emotion recognition has been highlighted. Evidence (R. E. Jack, Garrod, Yu, Caldara, & Schyns, 2012; Rachael E. Jack, Blais, Scheepers, Schyns, & Caldara, 2009) indicates that six distinct groups of facial expressions might be specific to the emotion perception of western cultures but do not emerge for other cultures, if these categories are not implied in some form by the research design (Nelson & Russell, 2013). This is directly relevant for empirical research as discrete emotion categories might limit the descriptive possibilities for emotions of researchers and participants alike and in consequence restrict the set of attainable results. Frequently the six basic emotion categories constitute the stimuli set as well as the set of response options within an emotion research paradigm. The categorical nature of the theory results in paradigms where the participant has no other choice than to select a single option from many choices, so-called forced-choice paradigms (Frank & Stennett, 2001). Likewise, stimuli creation as well as stimuli

selection guided by basic emotion theory often results in prototypical or even caricature emotion expressions (Barrett, 2006; Goldstone, Steyvers, & Rogosky, 2003) of basic emotion instances, which result in an artificially high differentiability of the categories. Apart from the apparent low ecological validity, this might also be directly responsible for so-called ceiling effects in emotion recognition often observed in such paradigms, for example described in (Isaacowitz et al., 2007; Palermo & Coltheart, 2004). Ceiling effects refer to observations of near perfect accuracy for the recognition of certain emotions across participants or groups, which limits the sensitivity of these research paradigms. Moreover, this might also artificially reinforce evidence for basic emotion categories.

From observations of language, literature, art, music and entertainment, but also from personal experience alone, it seems likely that the range of emotional experiences is not exhausted by the six basic emotion categories. For example, Cowen & Keltner (2017) extracted 27 distinct emotion categories from categorical, free response and dimensional self-reports, although their universality and innateness is debatable. However, it is not apparent why research should focus exclusively on universally understood emotions as opposed to culture specific emotions or any other subset of emotions. On top of that, emotions and their expression might be influenced by both, culture and biology, to varying degrees. For instance, Cordaro et al. (2018) investigated 22 emotional expressions in five cultures for stable between-culture patterns and found them in all cases but also found systematic cultural variations, which highlights the difficulty in separating influences of culture and biology in emotions. Although the influence of culture is acknowledged by proponents of basic emotion theory, empirical work based on the theory usually does not account for it. Basic emotion theory therefore presents a limited set of emotion categories as an implicit ground truth, which might restrict and bias research endeavors and their outcomes.

1.3.3 Theory of constructed emotion

Russell & Barrett (1999) brought forward the popular (2382 citations, Google Scholar 14th August 2019) account of constructed emotion. It stands in contrast to the biological deterministic view of basic emotion theory. Central to the theory is the hypothetical construct of *core affect*. According to Russel (2003) core affect is the "neurophysiological state consciously accessible as the simplest raw . . . feelings evident in moods and emotions." (p. 148). Such a state exists at all times. In that way, core affect presents the basis for all further affective phenomena, such as mood or specific emotions, which are constructed by a conscious evaluation of core affect. Core affect is described alongside two dimensions, which are assumed

to be largely independent. Therefore, core affect can be represented as a two-dimensional coordinate system with two continuous, orthogonal axes (Figure 1). The first dimensions of this coordinate system is labeled *valence* and the second dimension is called *arousal*. Valence is the level of pleasantness or pleasure and typically measured on a scale ranging from "very negative/unpleasant" to "very positive/pleasant". Arousal captures the level of physical activation or excitement and is typically measured on a scale ranging from "very calm" to "very aroused" or from "deactivation" to "activation" (Figure 1). Although both dimensions have a history as separate constructs in emotion research, they can be found empirically as the first two underlying variables in dimension reduction analyses (such as factor analysis or multidimensional scaling) on self-reported or ascribed affect and ordered emotional words (Russell, 2003b; Russell & Barrett, 1999). Higher-order affective phenomena like mood and emotion states can be mapped back to the specific core affect, which gives rise to them. Mood could simply be viewed as core affect over time. Emotional states like anger or joy are more likely to arise from certain regions of core affect than others. However, similar core affect might potentially be associated with different emotion experiences. The theory of constructed emotion postulates that an emotional experience only originates after core affect has been categorized as a familiar emotion concept by means of an appraisal process. This appraisal process takes into account the context of the situation to match the core affect state to a specific emotion concept, which is learned and potentially shaped through culture.

For example, an upcoming rollercoaster ride might induce a state of neutral valence and heightened arousal in a person queuing for it. If the person had heard that this rollercoaster is a lot of fun they might evaluate their core affect as excitement, whereas they might experience it as anxiety, if they had heard that the ride makes some people sick and they know that they have a weak stomach. Although core affect, according to its theoretical underpinning, does not directly describe emotion states, it is used in empirical research to assess emotional perception of stimuli, for example in (Grühn & Scheibe, 2008).

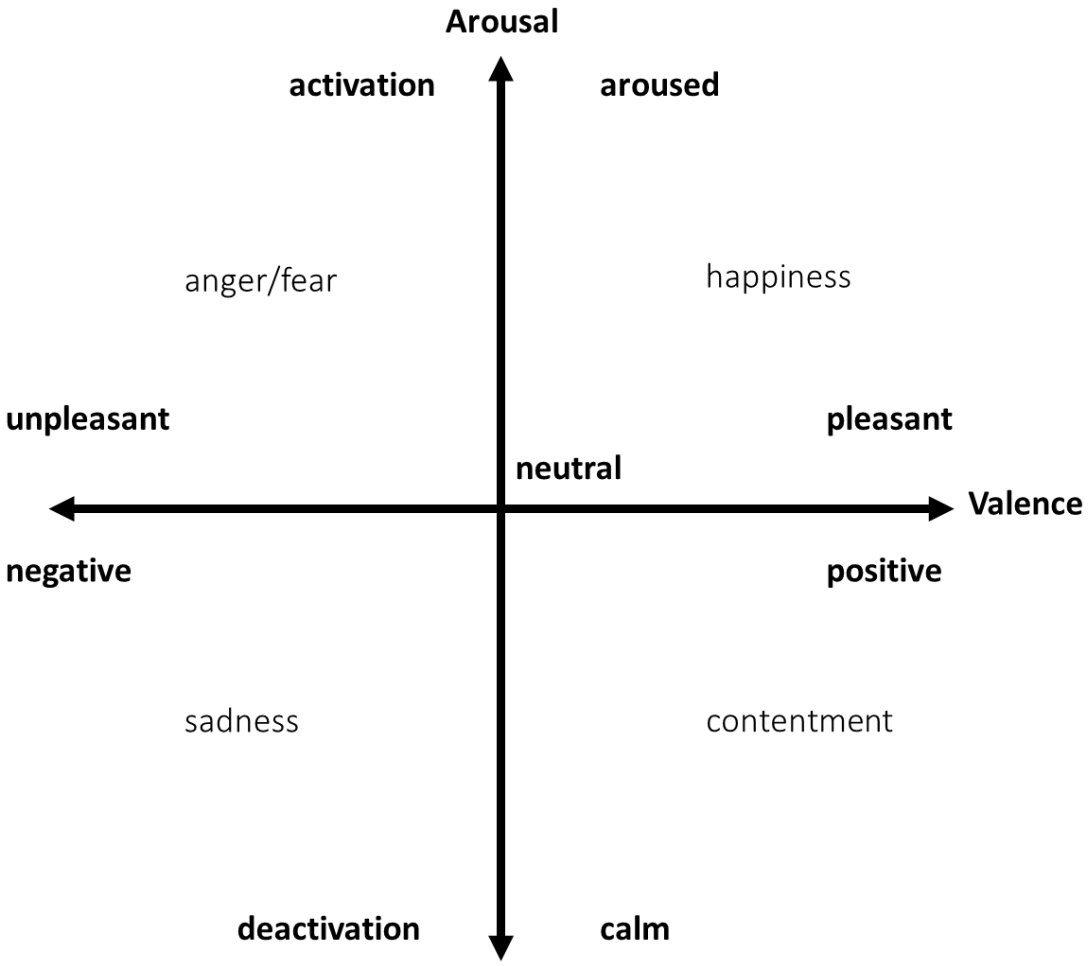


Figure 1. Core affect space according to Russell (2003a). The dimensions valence and arousal span this space in which emotions and moods can be located.

1.3.4 Criticism of the theory of constructed emotion

Interestingly, core affect, the central construct of the theory of constructed emotion, faces some of the same criticism that is voiced towards basic emotions. For example, that experiences of opposite valence, such as disgust and amusement (Hemenover & Schimmack, 2007) or positive and negative affect (Riediger, Schmiedek, Wagner, & Lindenberger, 2009) can be experienced simultaneously, is an argument against the discrete categories of basic emotion theory, but also cannot be resolved by core affect, since these states would reside on opposite sides of the valence dimension, which means they cannot be represented by a single point in core affect space. Scarantino (2009) argues that, based on the criteria brought forward by Barrett herself (common cause and/or common set of properties), core affect is even less likely to be a natural kind than discrete emotion states are. The reason for this is that core affect encompasses a much wider range of affective states than basic emotions so that it is less likely that they are caused

by the same mechanism or share a set of properties. Moreover, Scarantino states that, if scientific usefulness of a representation hinges on being a natural kind, then core affect's usefulness might not be better either than that of basic emotions.

Although its dimensional system allows for infinite affective states, core affect alone is too simplistic to distinguish common emotion categories like fear and anger, which are both high in arousal and low in valence. According to the theory, these emotions should arise from the same core affect by a different appraisal predicated on different context. However, since context is difficult to assess in a standardized way empirically, additional dimensions, such as "Dominance-submissiveness", have been proposed to allow for the differentiation of these emotions within the dimensional coordinate system alone (Fontaine, Scherer, Roesch, & Ellsworth, 2007; Mehrabian, 2007). In general, it is unclear how stable the dimensions of core affect are when they are derived empirically. Depending on the original observations, the ordering of affect dimensions derived by dimension reduction methods can change, so that, for example, the arousal dimension is the third dimension and a dimension labeled "potency-control" is found as the second dimension (Fontaine et al., 2007). This calls the primacy of the two dimensions of core affect into question. The specific components of emotion that are covered by the observations seem to play an important role here. Including observations on components such as action tendencies, bodily phenomena and regulation tendencies might lead to dimensional representations different from core affect, which is usually found via observations of self-reported emotions or ratings of facial expressions and vocal expressions. However, even in self-reported affect the relationship between valence and arousal seems to exhibit highly idiosyncratic patterns across individuals and situations, so that the assumption of orthogonality of these dimensions can also be called into question (Kuppens, Tuerlinckx, Russell, & Barrett, 2013).

Another related point of criticism is that both individual dimensions of core affect refer to theoretical concepts that potentially cannot be assessed in full by a single dimension. For instance, valence has been argued to be a multifaceted phenomenon that is closely intertwined with appraisal processes itself (Shuman, Sander, & Scherer, 2013), which also questions the separation of appraisal processes from core affect itself. Similarly, up to three separate dimension have been proposed to account for all theoretical aspects of the concept of arousal (Schimmack & Grob, 2000). Moreover, potential autonomic, somatic and cortical measures of arousal exhibit low correlations (Barrett, Bliss-Moreau, Quigley, & Aronson, 2004). Hence, a

substantial spread in theoretical concepts and specific measurements exists which could be tied to the arousal concept.

At first glance, the dimensional structure of the core affect construct of the theory of constructed emotion seems to provide more flexibility than the discrete basic emotion categories. However, core affect requires additional processes or dimensions to employ it in the differentiation of fundamental emotion states. Moreover, theoretical concerns and empirical results cast doubt on the validity of the construct in terms of stability and completeness.

1.4 Psychological Constructs and their Operationalization

One of the fundamental questions of emotion research is how emotions can be measured. Emotions, like most psychological phenomena, are not directly observable. Emotions can, however, be observed through their manifestations, for example in facial expressions or in self-reports of felt emotion. These manifestations point towards the phenomenon itself. In order to investigate psychological phenomena, researchers define constructs and propose ways to measure them through their manifestations (Cronbach & Meehl, 1955). Constructs are mental abstractions that serve as a label for a collection of co-occurring features and in this way organize and facilitate scientific discourse and reasoning (G. T. Smith, 2005; Teglasi, Simcox, & Kim, 2007). They are central to the methodology of psychological research and thus important in order to understand the generation and evaluation of empirical results of the field.

Constructs can be postulated arbitrarily for all kinds of things. In order for them to be empirically useful, it has to be determined which measurements relate to them and which do not. A measure or test is said to have *construct validity*, if it measures the construct it is supposed to measure (Cronbach & Meehl, 1955). This has to be determined by the theory of the construct. For instance, Cronbach and Meehl (1955) bring forward the example of a potential measure for the construct “anxiety proneness”, which could be justified, because this measure also correlates with palmar sweating after academic failure as well as with another established measure of anxiety as a personality trait. On the other hand, the potential measure should exhibit little or no correlation with measures theoretically unrelated to the target construct, for example “academic aspiration”, which could offer alternative explanations of the data. This means that although constructs are theoretically conceptualized, they are fleshed out in the manifestations that can be linked to them, because these describe the influence of a construct in a practical sense. In the same sense, theoretical predictions of relationships between constructs can only be confirmed by testing for a relationship between the established

measurements of these constructs. Likewise, measures of the same construct should correlate, because they point to or are generated by the same phenomenon. A construct with clear theory facilitates linkage of measures, whereas ambiguous or conflicting theories about a construct hamper this process.

The process of making a construct measurable, i.e. finding observations that can be linked to a construct, is called *operationalization*. The name alludes to the operations that have to be carried out for the particular measurement (Walsh, 1927). For example, the happiness of a person could be operationalized in a number of ways: by self-reported happiness of that person, by measuring the activity of their facial muscles that raise the corners of the mouth or by counting the number of their smiles in a certain time period. Although the operationalizations differ, the construct to which they point and hence the phenomenon of interest is the same. However, because the operationalizations differ, the part of the actual phenomenon that is measured by them might also differ. For example, smile-counting might overestimate happiness, if the observed subject displays many awkward or embarrassed smiles in order to mask uncomfortable feelings. On the other hand, self-reported happiness is a measure closely tied to the internal state of a person but also depends on the honesty of the subject. Hence, different operationalizations result in different biases. In an ideal case, results and therefore inferences about the concept itself are stable under different operationalizations. However, this might not always hold as exemplified above.

Operationalizations are driven and influenced by several factors. First, operationalizations are driven by the conceptual framework or theory within which the construct was conceived. These define what the construct is and is not and will therefore guide decisions on how the construct should be measured. In emotion research the discussed views on emotion and the theories of emotion associated with them propose specific operationalizations, such as self-report of the valence and arousal dimensions to measure core affect in the case of the theory of constructed emotion or to choose between the six basic emotion categories in the case of basic emotion theory.

Operationalizations are also driven by a need for simplicity and feasibility of the measurement itself and the subsequent analysis of the collected data. An elaborate idea for operationalization is of little use, if it cannot be implemented with the current technology or only with a considerable amount of time and money and thus limiting its applicability. Accordingly, one of the most common ways of operationalization in psychology is self-report (Haeffel & Howard, 2010) and it has been argued that it has progressively replaced direct

observations of behavior (Baumeister, Vohs, & Funder, 2007). This operationalization is easy to understand for the researcher and the participant and requires a minimal amount of technology, because it can be administered verbally or with the help of pen and paper. On the other hand, an example for an operationalization on the extreme end of technological sophistication would be measuring threat perception by monitoring blood flow to certain areas of the brain with functional magnetic resonance imaging. This certainly would be a more effortful and expensive endeavor than merely asking a participant, whether they felt threatened.

One of the goals of empirical research in psychology is to find generalizable statements about constructs and the relationships between them. This can be accomplished by drawing inferences about constructs from a sample to the population from which it was taken. Ultimately, the operationalization of a construct is just one step in this process. Inferences are made by using statistical modeling and inferential statistical methods. The prevalent models for statistical analyses in psychology are usually derivations of the generalized linear model (GLM) (Skidmore & Thompson, 2010). Widely used examples for such model types are t-tests, ANOVA and (multiple) linear regression. Usually a low number of variables enter these models, which forces researchers to select only the most relevant measurements. If the goal of the scientific endeavor is inference of population characteristics, then operationalizations are certainly also influenced by considerations of the possibilities of using them for inferential statistical analysis with the common models. This might also be the reason for a clear preference of single-value operationalizations in psychology.

In summary, psychological constructs are mental abstractions that facilitate scientific discourse and which are defined by theory as well as by specific manifestations that can be tied to them. The operationalizations of constructs are driven by the theory of the construct, but also by practical considerations, such as feasibility of measurement and subsequent analysis of the gathered data.

1.5 Current Methodological Problems of Emotion Research

The previous sections have introduced prominent schools of thoughts of emotion research as well as constructs and operationalization as central concepts of empirical psychology. In the following I will summarize general issues of contemporary emotion research and identify three underlying methodological problems of the field.

Empirical research on emotions has been conducted for more than hundred years (Wundt, 1908). Yet, many questions that concern the methodology of the field remain open.

Prominent scholars of the field have acknowledged that the present theories of emotion and their methods of measuring emotions face fundamental problems that are unlikely to be solved in the near future by improvement of either theory or by a unifying framework (Mulligan & Scherer, 2012). Disagreement between theories of emotion results in uncertainty about the adequate measurement of emotions and complicates research, since different views on emotions influence crucial research decisions, such as the measurement of the phenomenon, the experimental design, the analysis of collected data and the interpretation of results. Different research efforts might come to different conclusions, solely because they start with a different premise on emotions (Gross & Feldman Barrett, 2011). This is related to the fact that the prevalent theories of emotion are limited in the way in which they can describe and measure emotion. Mauss and Robinson (2009) review the literature on emotion measurement and come to the conclusion that emotion is likely a multidimensional phenomenon and hence no single gold standard to measure emotion can be found. If emotion is a multi-dimensional phenomenon, it is easy to see that reduction onto a low-dimensional space or onto discrete categories will inevitably bias its measurement. Here it should be noted that each reduction onto a subspace introduces a bias specific to the reduction. Different reductions will result in different biases. Using these reductions as emotion representations in empirical research might impose a ground truth that is arbitrary and will lead to results that seem to reinforce the evidence for the chosen emotion representation. In general, when reducing a space in the mathematical sense to a subspace of lower dimensionality, the loss of information and discriminability grows with the reduction in dimensionality (apart from a few exceptions unlikely to occur in empirical data). Therefore, reductions of the original space of emotions, which is unknown, to a representation of lower dimensionality should be carefully conducted.

Many debates in the field of emotion research, such as the one about natural kinds, can be reduced to a search for the "true structure" of emotion and the attempt to find a representation that aligns with it. Although the question of natural kinds is the fundamental division between views on emotion, one can remain skeptical whether trying to answer it might be of particular help for finding the "correct" emotion representation or for advancing the scientific knowledge on emotions in general. First, similar to the problem of defining emotion, there seems to be no agreement on a definition of what constitutes a natural kind and different views and theoretical frameworks exist that are difficult to unify. It seems quite ironic that it is debatable, whether the class of natural kinds is itself a natural kind or not (Dupré, 2002). Therefore, the question of natural kinds might not be any more useful than the question of biological or constructed

emotions and seems to be a mere shift of the problem into the philosophical domain. Kinds can be researched whether they are natural or not, as is the case with many socially constructed phenomena.

Fried (2017) has remarked that being a natural kind might not add anything relevant to a construct. The only concern seems to be the possibility of a change of a non-natural kind construct as an effect of the discourse about it, as has been noted by Hacking (1999). However, such changes are probably occurring on a long-term basis and might all the more call for a proper and repeated examination of the phenomena. In this view, scientific usefulness of a kind should take precedence over the origin of a kind in the philosophical sense. Hence, a good kind is one that is useful. Likewise, it should be possible to research emotions, whether they are biological or social phenomena or a combination of both. This is not to say that the question of biological and cultural influences on emotions is not an interesting one. It certainly is. Rather, it is to say that there are many interesting scientific questions on emotions to be answered which do not depend on the nature of the phenomenon. Similarly, it should be considered that there might not be a one-fits-all representation for emotion or that we are not able to define one at this point in time. Instead, research should find and use appropriate representations for each specific research question or context.

Important for the scientific usefulness of constructs is a clear definition and, as described in the previous section, the definition of constructs can be achieved in a practical sense by linking them to concrete manifestations. However, many constructs with a long-standing history of use in emotion research seem vague and incomplete, because multiple, sometimes conflicting theoretical accounts for the same constructs exist and because the extent of the concrete manifestations which can be ascribed to them is unknown. One reason for the latter might be the reliance on self-report measures to quantify these manifestations. Self-report measures might only capture certain manifestations, which are accessible through introspection. In addition, self-report measures suffer from a number of biases, such as social-desirability bias and acquiescence bias (Anusic, Schimmack, Pinkus, & Lockwood, 2009; Mortel, 2008). Social-desirability bias, for example, is well known, however, studies using self-report measures rarely correct for it (Mortel, 2008). Schimmack (2010) has stated that proper construct validation has to rely on multi-method data and therefore psychological constructs might benefit from methods other than self-report to identify their respective manifestations. However, currently not many other methods to quantify emotion manifestations are available to emotion researchers.

Many of the discussed problems of the field of emotion can be reduced to underlying general problems. This thesis will approach the following three:

1. The *problem of ground truth* refers to the practice in empirical emotion research to presuppose a certain theory of emotion and with it a definite structure of emotions. This is paralleled with a restrictive way in which emotions can be described and labeled. Certain labels or descriptions are denoted as correct, which sets an a priori ground truth. Ground truth is useful when abilities tied to emotions are estimated. However, a fixed a priori ground truth is problematic because it is predicated on theories of emotion, which are themselves debatable and limited in their descriptive potential. As such, a priori ground truth might bias results or prevent certain research endeavors altogether.
2. The *problem of incomplete constructs* refers to psychological constructs which are vaguely defined in theory and lack ties to specific manifestations. Since these constructs are neither sufficiently defined in theory nor practice, they are incomplete. Although many constructs of emotion psychology have a long-standing tradition, it is still unclear what they mean and how they are expressed in or perceived from concrete displays of emotion to which they directly relate. This is problematic because it restricts the applicability of these constructs and adds to their fuzziness. Conversely, understanding the practical implications of a construct adds to its theoretical value and its usability in scientific discourse, because the areas in the real world it affects and does not affect become evident.
3. The *problem of optimal representation* refers to the search for the best representation of emotion. One of the central endeavors of the field of emotion has been the search for the true structure of emotion and a representation that depicts it accurately. However, the current theories of emotion propose emotion representations that are difficult to unify. Current representations are simplistic one-fits-all solutions, meaning the representation stays the same independent of the context of its use. A more attainable goal might be to find a detailed representation, which is optimal to a specific context or task.

1.6 A Focus on Measurement

This thesis argues that a reduction of measurement bias is possible with a methodology that operates largely agnostic of any particular theory of emotion. I argue for a focus on the concrete manifestations of the phenomenon itself and hence for a detailed quantification of emotional

expression and experience. As I will explain in the following, this is possible with computer-aided and data-driven approaches.

Emotion research appears to be trapped in a vicious cycle of mutual uncertainty on the side of theory and practical measurement. Because it is unclear what emotions are, it is unclear how they should be measured and because it is unclear how they should be measured emotion theory cannot be advanced by new insights gained with measurement. An analogy to this situation might be found in physics. Chang (2005) describes in great detail the history of human attempts to understand and measure temperature. This account is a fascinating example of the interplay of improvements in measurement of a phenomenon and increases in theory about it and bears many similarities to current emotion research. Similar to emotions, temperature was merely a construct without a clear definition in the beginning, which was primarily described by bodily sensations ("ice feels cold, fire feels hot") and simple sensory observations ("water can change temperature when exposed to ice or fire"). Similar to emotions, temperature research seemed to be trapped in a vicious cycle of uncertainty and went through a period of widespread theoretical disagreement. Similarly, many things seemed to have an influence on temperature and therefore complicate research on it. Many problems had to be solved until the first reliable thermometers could be produced and each question seemed to be immediately related to others. For example, researchers had to find reliable fix points to anchor the temperature scale, which begets the problem of how to know whether a fix point is really fixed. They needed to find a way to annotate the scale between the fix points, which is directly related to questions about the linearity of expansions of substances under temperature changes and so on.

In what Chang calls "epistemic iteration", every improvement of the measurement instruments sparked new theoretical contributions, which in turn improved the instruments and so forth. This process was not locally restricted to the field of temperature research but also heavily influenced by progress of other fields, such as the knowledge and measurement developments on the phenomenon of pressure. Although the ancient Greeks were already concerned with rough categorizations of different temperature, it took until the late 16th century for the first instruments for temperature measurement to appear and yet about another hundred years to arrive at a reliable, calibrated and standardized measurement instrument for temperature (Wisniak, 2000). Given the short history of empirical research in psychology then, it is not surprising that uncertainty and disagreement about fundamental concepts of emotion are still widespread. In comparison to the science of temperature, the science of emotion is in

the early stages of measurement instrument design and yet far away from calibration and standardization.

Although there seem to be many similarities, some differences between the measurement of temperature and emotion might be immediately obvious. Whereas physical phenomena, as we know now, are stable across time and can be measured on a single scale, psychological phenomena appear to have a much greater variance and it is not clear whether they can be broken down to single dimensions. However, one useful insight can be extracted from the analogy: better measurement can lead to improvements in theory and because theory in turn influences measurement this can lead to progress of the whole field. Consequently, I first propose to focus on improving the measurement of emotions. There is a simple reason why at this stage measurement should be improved instead of theory. The capacity for human reasoning has remained largely unchanged for the last few thousand years, whereas technology has advanced rapidly. Our advantage in this day and age is the access to technology, whereas the tools that our own mind provides have been at the disposal of generations before us and are likely exhausted at this point. This might be apparent in the similarity of theoretical accounts on emotion which were drafted at different times in human history. For example, Aristotle, famous Ancient Greek philosopher and polymath, already formulated quite sophisticated thoughts about emotions (Leighton, 1982). Hence, I believe that the way forward in emotion research is better measurement through technical solutions, specifically computer-aided approaches.

1.7 Benefits of Using Computer-aided Methods for Emotion Research

Computer-aided methods have several properties that render them ideal for emotion research. These methods offer the potential for stable and standardized measurement as well as for novel ways of gathering data in rapid and extensive ways. In the following I will describe how these methods are beneficial for psychological research in general as well as for emotion research in particular.

Emotions can be expected to have a lot of variation in the way and extent that they manifest independent of the method employed for their measurement. This is because emotions are expressed, observed and felt by people and therefore will be influenced by their idiosyncracies. For example, facial muscles vary in number and symmetry across people as does the general appearance of the face (Hess, Adams, & Kleck, 2009; Waller, Cray, & Burrows, 2008). This means that even if two people felt the exact same happiness, they might

not be able to produce the exact same smile and the emotion they express might therefore be perceived differently. In comparison to measuring something that is directly tied to the laws of nature like temperature, this seems to be adding more difficulty to the task. However, unlike the researchers who attempted to measure temperature with instruments they could not know to be comparable or stable across time and environmental conditions, computer science offers tools that are by design deterministic. This is ideal, because no additional variation (those of the measurement instruments) is introduced.

Another beneficial property of computer-aided methods is the general potential to acquire large amounts of data rapidly (Kraut et al., 2004) through automation of data collection. This is of great use, since emotions seem to be a multivariate phenomenon (Mauss & Robinson, 2009) and thus could be difficult to understand with individual and sparse measurements. Moreover, comprehensive ways of measuring emotion manifestations could reduce bias due to the assumptions and restrictions of a particular theory of emotion because no a priori reduction of the phenomenon has to take place. Simplicity in measurement might have been necessary and inevitable in the past because of a lack of alternatives.

However, in the recent past computer-aided practices of data collection have become more prevalent. For example, data can now be collected in experimental tasks that are completed entirely in front of a computer by the participant, which reduces the need of supervision and guidance by a researcher (Kraut et al., 2004). This also allows for complex experimental sequences with a high degree of standardization compared to tasks that are primarily researcher guided. Moreover, multi-media elements ranging from video clips to interactive virtual reality experiences can be incorporated with traditional survey elements seamlessly. Emotion research in particular might benefit from these elements as emotional responses and perceptions tend to be elicited especially in lifelike and naturalistic settings (Riva et al., 2007).

Computerized experiments also allow to gather data online without the physical presence of participants. This has resulted in a surge in Internet-based research across disciplines in psychology (Evans & Mathur, 2005, 2018). Online studies allow for low-cost, rapid data gathering and potentially bigger sample sizes through the possibility of online advertisement coupled with online dissemination of the study. Conventional psychology experiments conducted in the laboratory have been accused of oversampling from the population of specific groups such as students found on university campuses (Henrich, Heine, & Norenzayan, 2010; D. Jones, 2010). These groups might possess traits, which are rarely

found in the general population and thus generate results, which cannot be generalized to the wider population. Here, online studies might provide access to a more diverse population and hence generate more generalizable results (Gosling, Sandy, John, & Potter, 2010; Nosek, Banaji, & Greenwald, 2002). Online studies have been found to produce similar results as conventional laboratory studies (Paolacci & Chandler, 2014; Paolacci, Chandler, & Ipeirotis, 2010), which indicates that their beneficial properties are not necessarily offset by a reduced accuracy or trustworthiness of the results.

Some advancements in computer science allow to capture novel types of data. Many of these advancements come from the field of computer vision. In particular, relevant for experimental psychology are technologies such as eye tracking (Duchowski, 2017), automatic emotion recognition (D'Mello & Kory, 2015) and face tracking (Chrysos, Antonakos, Snape, Asthana, & Zafeiriou, 2018). Not only do these technologies allow to collect vast amounts of spatial and temporal data automatically, but they also provide novel possibilities for operationalizations of emotion phenomena that are potentially more objective than self-reports. These methods might also offer a way to produce the data for multi-method construct validation (Schimmack, 2010). In this way, long-standing constructs of psychology can be validated and extended. Concrete and reproducible measurements of emotion manifestations might allow to verify predictions derived from the theory of certain emotion constructs. Moreover, it might also allow to quantify the extent to which measurements and constructs are related between and among themselves. In this way, these methods could help to clarify existing constructs and foster the development of novel ones.

1.8 Challenges of Using Computer-aided Methods for Emotion Research

Computer-aided methods provide new challenges, too. These lie on the side of data collection as well as subsequent data processing. Caution has to be taken to ensure adequate data quality when collecting data in online studies (Kraut et al., 2004). Because of the unsupervised approach of online studies there is a risk of gathering data from untrustworthy participants who give deceptive answers intentionally. This might be because they only participate because of a given incentive, such as a monetary reward for the study, and therefore seek to complete it as quickly as possible or to participate multiple times. Additionally, participants could misunderstand instructions or could have technical difficulties in displaying the study, and especially multimedia elements, correctly on their particular device and thus provide faulty data without intention. Therefore, clear and unambiguous instructions have to be given and technical

requirements should be kept low as well as explicitly stated. However, additional measures to ensure the adequate quality of the collected data, such as checks for faulty and nonsensical answers, should also be implemented.

Data from technical methods such as face or eye tracking is often high-dimensional because data is recorded on multiple locations across time. A challenge lies in either fitting these data into the statistical models commonly used for analysis in psychological research or in finding novel possibilities to analyze these data. Because these techniques have seen little application in emotion research, standardized protocols for data collection, processing and analysis do not exist and have yet to be established. This is a crucial step in ensuring comparability of collected data and in enabling applicability to a wide range of researchers. This thesis presents work that shows how data can be collected with a combination of computer-aided methods, such as online studies and face tracking software. The thesis shows how such data can be processed so that it can be analyzed with the conventional statistical models of the field. Furthermore, this thesis demonstrates a novel statistical method which can utilize high-dimensional spatiotemporal data directly.

1.9 Research Projects

Three research projects were conducted as part of this thesis. Each of them tackled one of the underlying problems of the current field of emotion research (the problem of ground truth, the problem of incomplete constructs and the problem of optimal representation; Section 1.5). Common to all projects is a focus on emotion perception as a dyadic process between the person expressing emotion and the person perceiving emotion. This view allows research independent of prevailing emotion theories and their proposed ways of measuring emotion. Instead, computational methods are used to provide a detailed quantification of emotion.

Project 1 illustrates how the problem of ground truth can be circumvented with a novel emotion perception paradigm that allows the use of an arbitrary number of emotion categories. The method was applied to research the difficulty of the perception of emotional facial expressions and the factors that constitute it across age and sex. The project shows how the objective difficulty of emotion perception can be computed without a priori ground truth and measures subjective difficulty of emotion perception as well. Person-specific variables as well as emotion specific variables are investigated for their influence on these two difficulty measures. This project exemplifies how research on emotion perception abilities can be conducted without reliance on the common emotion recognition paradigm and its shortcomings.

Project 2 targets the problem of incomplete constructs and exemplifies how face tracking data can be used to extend the understanding of arousal, a long-standing psychological construct. It describes how measures of distance, speed and magnitude of acceleration can be computed from face tracking data and investigates their intercorrelations. The project then investigates how arousal is perceived from facial expressions in neurotypical individuals and individuals with autism. The project tests whether these groups use static and dynamic features of the face to assess emotional arousal and if they do so to the same extent. This project thereby illustrates how computer-aided methods add to the knowledge on psychological constructs and exemplifies how they can be employed for specific research questions.

Project 3 describes a solution to the problem of optimal representation. It outlines a general method that allows to find the optimal representation to investigate the relationship between two high-dimensional data sets. The method enables detailed research on emotion perception and perception differences between groups without the a priori assumption of any particular emotion theory. This is illustrated by using it on emotion rating data and face tracking data to find differences in the perception of facial expressions between men and women. The project thereby provides a solution to the fundamental question of appropriate emotion representation. All projects focused on an application of the proposed methodologies to generate novel results to demonstrate usefulness and usability for psychological research. The three research projects are presented in the following in the outlined order.

2. Project 1: Difficulty of Emotional Expression Perception

The following study was conducted in collaboration with Timothy R. Brick, Anne Weigand, Isabel Dziobek and Ulrike Lucke. I designed the study, collected and analyzed the data as well as described and visualized the results under their supervision and with the help of their advice. As this was a joint effort, I use the first person plural ("we") in the following.

2.1 Research Motivation

Facial expressions are an essential part of human interaction and have been a research topic for more than a century (Darwin, 1872). Facial expressions transmit information between the individual displaying the expression (called actor in the following) and the individual perceiving the expression (called observer in the following). For facial expressions to function as an effective communication channel the actor must produce expressions that will be readily understood by the observer. The observer, in turn, must possess the necessary abilities to interpret what they see in the intended way. It can be assumed that the difficulty that the observer experiences in trying to understand the portrayed emotion might vary depending on several factors pertaining to the actor, the observer and the displayed emotions. An understanding of the constituents of the difficulty of facial expressions is valuable for human communication in general, but in particular for situations in which the focus lies predominantly on the correct interpretation of facial expressions. Such situations might arise in training-based therapeutic approaches for individuals with impaired emotion recognition as is the case in certain clinical populations. An example for such a population is autism spectrum disorder (Harms, Martin, & Wallace, 2010). Computerized training programs provide a convenient and highly-standardized way to train socio-cognitive skills and have been shown to improve affect recognition in individuals with autism (Bölte et al., 2002).

Knowing the factors contributing to the difficulty of emotion perception might allow a fine-grained compilation of tasks of appropriate difficulty for the respective ability of the user. Such an adaptive difficulty algorithm was conceptualized for a training system targeted at autistic individuals (Moebert & Lucke, 2019), which was developed within the EMOTISK Project funded by the German Federal Ministry of Education and Research (German: Bundesministerium für Bildung und Forschung). The basis for an adaptive algorithm is data about the variables that affect the adaptive variable. However, influences of variables such as the age or sex of the actor and observer on the understanding of emotions have been typically studied with emotion recognition paradigms (e.g., (Hoffmann, Kessler, Eppel, Rukavina, &

Traue, 2010; Isaacowitz et al., 2007)), which exhibit several limitations for the in-depth study of emotion perception difficulty. *Recognition* means matching a stimulus to some concept recalled from memory, which, in the case of emotions, implies a distinct set of predefined emotion categories and a formally correct answer to recognize. In this study, we challenge limitations of emotion recognition paradigms arising from the use of discrete basic emotion categories, the calculation of accuracy predicated upon a priori ground truth and static stimuli. Instead we propose a framework that allows to study emotion *perception* without a priori ground truth. By emotion perception we mean all perceptual processes that occur when an observer views an expression.

We here present a framework that enables the estimation of ground truth in a dimensional space of emotion from the population consensus and from a wide array of stimuli categories. Using this approach, we investigated predictors of subjective and objective difficulty of emotion perception in a sample of 441 participants who rated dynamic emotion stimuli from 40 categories. We reproduced results known from the emotion recognition literature but also present unexpected relationships and novel results. Furthermore, we calculated the importance of each predictor for the difficulty of emotion recognition using a machine learning model. A part of the results of this study were published in (Moebert, Schneider, Zoerner, Tscherejkina, & Lucke, 2019).

2.2 Limitations of Common Basic Emotion Recognition Paradigms

2.2.1 Forced-choice and basic emotion framework

The majority of facial expression recognition studies of the past used a forced-choice paradigm in which only a single discrete emotion category was predefined as the correct answer (for example (Kessels, Montagne, Hendriks, Perrett, & de Haan, 2014; Montagne, Kessels, Frigerio, De Haan, & Perrett, 2005)), namely the one the actor intended or was instructed to display. The discrete categories of the stimuli as well as the response options were commonly selected to align with Ekman's so-called "basic" emotions (Ekman, 1992a), a set of six prototypically displayed emotions. The choice to provide only basic emotion labels as response options implicitly assumes that emotion perception is discrete and exclusive.

However, studies (Hemenover & Schimmack, 2007; Riediger et al., 2009) have repeatedly shown that even affective states of opposite valence (disgust and amusement, positive and negative affect) can be experienced at the same time. Other studies showed that participants readily use multiple emotions in a dimensional way to describe even prototypical

target emotions (Hall & Matsumoto, 2004; Phillips & Allen, 2004; Riediger et al., 2011). This evidence supports a multifaceted and dimensional rather than a categorical perception of emotion. In a study by Hall & Matsumo (2004; Study 2), increased accuracy of female participants was only found when dimensional response options instead of discrete emotion labels were provided. This indicates that some group differences may lie in the pattern of multi-dimensional emotion judgments and thus cannot be observed when participants provide only a single emotion label.

It is also possible that the reduction of emotional perception into discrete labels produces artifacts. For example, in a study on the commonly used forced-choice basic emotion paradigm Frank and Stennett (2001) found wide agreement on an incorrect label when the correct one was omitted and even for a nonsensical expression for which a response label was not conceivable. They showed that they could remedy this effect by providing a "none of these terms are correct" option and investigated multiple other variations of the basic emotion paradigm. Taken together, the research by Frank and Stennett (2001) questions the generalizability of effects found in a basic-emotions paradigm. Moreover, these results indicate that the accuracy in emotion recognition might also change substantially depending on the number and nature of the provided categories.

2.2.2 Stimuli in emotion recognition paradigms

2.2.2.1 Prototypical expressions

A well-known problem with common basic emotion stimuli sets is the often prototypical and over-exaggerated facial expressions that produce ceiling effects and lack ecological validity (Kessels et al., 2014). This prompts the question of how comparable these stimuli are to authentic real-life expressions and what impact this difference from naturally-occurring expressions has on research outcomes. In fact, some effects, such as improved emotional facial expression perception among females could only be detected if stimuli were used in subtle and "toned-down" versions (Hoffmann, Kessler, et al., 2010; Montagne et al., 2005).

2.2.2.2 Static vs dynamic stimuli

Past studies of emotional facial expressions traditionally employed static stimuli in the form of photographs. In recent years, research has investigated whether dynamic stimuli offer additional benefits or might differ systematically in the perception that they evoke. One consistent finding is that emotional facial expressions have a specific dynamic time signature which is perceived as most naturalistic (Hoffmann, Traue, Bachmayr, & Kessler, 2010; Sato & Yoshikawa, 2004) and which results in higher recognition accuracy compared to deviant time signatures (Kamachi

et al., 2001). Calvo, Avero, Fernández-Martín and Recio (2016) concluded that dynamic facial expressions were recognized faster and with higher accuracy than static ones. Research by Kilts, Egan, Gideon, Ely and Hoffman (2003) found distinct neural pathways for the processing of static versus dynamic stimuli, while Perdikis et al. (2017) found systematic differences in EEG signals from viewing natural dynamic stimuli even when dynamics were perturbed but not removed. A review by Krumhuber, Kappas and Manstead (2013) concluded that dynamic information increases emotion recognition accuracy, ratings of intensity and arousal and detection rates of genuine and fake expressions, a finding backed by automated expression analysis (Brick, Hunter, & Cohn, 2009). They argue that dynamics are an important part in understanding the phenomenon of facial expressions and urge researchers to overcome the use of static facial expression stimuli. In agreement with this, we expect that dynamics provide essential information for the perception of facial expressions and hence cannot be ignored when researching its difficulty.

2.2.3 The question of ground truth

Traditional studies tend to assume that the intention of the actor determines the meaning of an emotional expression. Consequently, posed expression stimuli are often labeled with the emotion that was requested from the actor (Lucey et al., 2010). Accuracy in an emotion recognition paradigm then refers to the proportion of labels assigned by the participant that match the predetermined stimulus labels. However, a facial expression might not always convey its intended meaning. For example, a person smiling at a camera might appear awkward or uncomfortable rather than showing the intended emotion of happiness. From the observer's perspective, then, a potentially more appropriate ground truth for an emotional label is the consensus among culturally-similar observers.

This raises a new question in the context of group differences. Traditionally, if two groups differ systematically in their interpretations of an expression, one group is considered to be incorrect. If the ground truth is consensus, however, it is difficult to argue that the consensus among male observers, for example, is more or less correct than the consensus among female observers. Instead, these should be interpreted to be reflective of different biases in ground truth. A research paradigm that investigates the difficulties of emotion perception should take these potential differences in perception into account.

2.3 Difficulty of Emotion Perception

It is evident that some facial expressions are generally easier to understand than others. For example, within Ekman's basic emotions (Ekman, 1992a) expressions of happiness show a higher recognition accuracy across cultures than expressions of disgust or fear (Elfenbein & Ambady, 2002). However, it is unclear what drives the difficulty in perceiving an emotional facial expression. Since an emotional perception emerges from the interplay of actor and observer, are features of both equally important for a successful perception? The present study identifies predictors of the difficulty of emotional facial expression perception. Some aspects of this difficulty might be salient to the observer while others might remain more opaque and difficult to report. These two facets of difficulty should therefore be examined separately: First as the subjectively experienced difficulty of the observer and second as an objective measure that captures the observer's abilities in relation to the rest of the population.

2.4 Relationship of Subjective and Objective Difficulty

Kelly and Metcalfe (2011) found that participants' trial-by-trial confidence of providing a correct answer in an emotion recognition task correlated with their performance, whereas participants' self-assessments of their general emotion recognition abilities did not predict task performance. Significant rank-order correlations between trial-by-trial performance and confidence self-reports of this study were in the range between .07 and .45. Importantly, the higher correlations were found for the Ekman Emotional Expression Multimorph Task (Blair, Colledge, Murray, & Mitchell, 2001), which features whole-face stimuli, compared to the Reading the Mind in the Eyes task (Simon Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), which features eye region stimuli; correlations were also higher when response options were revealed next to the stimuli.

Clearly, the amount of information given about the emotion task to be solved is a crucial factor in accurately predicting whether a "correct" answer will be provided. Overall, these findings indicate that only a weak to moderate correlation between the subjectively rated difficulty of an observer and their objective performance can be expected. We therefore expected that they capture influences from different aspects of the human emotion perception process. Studying their similarities and differences may provide insight into how these aspects are interrelated.

2.5 Measures of Subjective and Objective Difficulty

In this study, we quantified the difficulty of emotion perception in two different ways. The first uses observers' self-reported difficulty (SRD) and captures their subjective experience of difficulty in evaluating a given emotional stimulus. As described above, the ground truth for *perception* of an emotional stimulus can be derived from the cultural consensus among perceivers. We estimate this consensus point in a dimensional emotion space with the mean ratings given by a sample of observers. Our second measure of difficulty captures the deviation of an individual's perception from this estimated ground truth; it measures how "correct" the observer is relative to this consensus, and will therefore be called objective difficulty (OD). It is estimated from the study sample (see Section 2.8.3 for details on calculation) and therefore free of biases stemming from researcher decisions on stimulus labels.

2.6 Person- and Stimuli-specific Predictors of Emotion Perception Difficulty

We examined person-specific (age and sex of observer and actor) and stimulus-specific (valence and arousal of the displayed expression) variables as predictors for both proposed difficulty measures. Here, we briefly summarize known effects of these variables from the emotion recognition literature.

2.6.1 Age and sex of actor and observer

A review by Fölster, Hess and Werheid (2014) concluded that emotional expressions were more difficult to read from old faces than from young ones. These differences might stem from decreased intentional muscle control in the elderly affecting posed emotions and negative implicit attitudes towards old faces. Freudenberg, Adams, Kleck, and Hess (2015) could furthermore show that morphological changes of the face, such as folds and wrinkles, interfere with the emotional display in the elderly.

Women are more expressive in their facial expressions (Fischer & Lafrance, 2015; Kring & Gordon, 1998) showing, e.g., more smiles and more frequent head movements than men (Boker et al., 2011; Hess & Bourgeois, 2010; LaFrance, Hecht, & Paluck, 2003). A recent study by McDuff, Kodra, El Kaliouby and LaFrance (2017) used automatic facial expression recognition technology on over 2000 participants from five countries who displayed emotional expressions while watching television advertisements and showed significant sex differences in emotional expressivity. In particular, they found that women displayed most investigated

facial actions more frequently than men. However, they also found that men showed a greater frequency of brow furrowing compared to women.

There is agreement that decoding abilities decline with increasing age in adulthood (Isaacowitz & Stanley, 2011; Ruffman, Henry, Livingstone, & Phillips, 2008). By contrast, developmental effects of emotion recognition capabilities can also be observed until the end of adolescence (Herba, Landau, Russell, Ecker, & Phillips, 2006; Montirosso, Peverelli, Frigerio, Crespi, & Borgatti, 2010; Thomas, De Bellis, Graham, & LaBar, 2007).

Previous studies have found a sex difference in recognition abilities only when employing subtle emotional stimuli (Hoffmann, Kessler, et al., 2010; Montagne et al., 2005). A recent meta-analysis concluded that women have a small advantage over men in the recognition of non-verbal displays of emotion (mean Cohen's d= 0.19) (A. E. Thompson & Voyer, 2014).

2.6.2 Valence and arousal of stimuli

The core affect representation (Section 1.3.3) describes the space of emotions in terms of valence, which captures the level of positivity or pleasure of a stimulus, and arousal, which describes the level of alertness or physical activation (Russell, 2003b; Russell & Barrett, 1999). These dimensions frequently appear as the underlying variables in dimension reduction analyses (such as factor analysis or multidimensional scaling) on self-reported or ascribed affect and ordered emotional words. In this study we investigated the relationship of valence and arousal respectively with difficulty of emotion perception.

2.7 Hypotheses

Based on the discussed literature we formulated the following hypotheses:

Hypothesis 1: The age and sex of the actor have an effect on the objective and subjective difficulty experienced by the observer.

Hypothesis 2: The age and sex of the observer have an effect on the observer's objective and subjective difficulty in labeling a stimulus. Difficulty follows a quadratic function ("u-shape") across the lifespan because of developmental and deterioration effects at the respective ends of the age spectrum.

Hypothesis 3: The subjective and objective difficulty of stimuli are associated with the observer's perception of their valence and arousal.

We tested these hypotheses with a dimensional rating paradigm on video clips of emotional facial expressions from 40 different categories. Observers reported their perceived difficulty for each stimuli judgement and in addition we computed an OD measure based on the individual rating distance to the consensus of the whole sample. We then explored the relative importance of all the aforementioned predictors in predicting difficulty.

2.8 Methods

2.8.1 Participants

In total 658 took part in an online survey in German. Participants were included if they were either female or male native German speakers and not undergoing psychotherapy or taking psychoactive medication at the time of study. We included only participants who were 60 years or younger to avoid spurious age effects due to the high leverage of data points from a small number of very old participants (oldest 79).

Participants were only included, if they reported working video playback and if they passed a simple test of their reliability in emotion reporting. The data set was furthermore filtered for participants completing valence-arousal rating pages in an average time lower than 10 seconds and emotion rating pages in an average time lower than 20 seconds. The cutoff values were estimated by considering that the playback of each video clip takes 4 seconds and the usage of the rating scales takes at least two seconds per scale. The exact cutoff values were determined by visual inspection of the distribution of average times per page across all subjects.

Additionally, participants whose responses showed a low correlation between valence and happiness were excluded. Happiness is frequently used as a direct measure of positive affect (e.g., in the PANAS (Watson, Clark, & Tellegen, 1988)) and the two should therefore exhibit a high correlation. This assumption was further confirmed by visual inspection of the distribution of within-participant valence happiness correlations which showed a clear division around 0.4. Thus, this value was used as an exclusion threshold.

The mean correlation between happiness and valence was 0.868 before and 0.875 after this filtering step. In total 217 participants (33%) were excluded and 441 remained (129 males). Mean age was 28.08±8.17 for women and 29.82±8.88 for men. 423 participants reported a German cultural background. 380 were of Caucasian ethnicity, 23 of other ethnicity and 38 chose not to provide an answer to this question. The minimum sample size was set to 400 participants, which would ensure that on average each one of the 480 video clips was rated by 10 participants in the case of 12 ratings per participant. We collected more participants initially

because we made an estimate of having to exclude at least 30% due to the aforementioned criteria.

2.8.2 Materials

The present study used a set of 480 videos taken from a larger data set of over 2000 facial expression video clips (Kliemann, Rosenblau, Bölte, Heekeren, & Dziobek, 2013). This selected set of videos consisted of 12 actors (6 male, 6 female, age range: 21-64) displaying emotional facial expressions from 40 different emotion categories (Supplementary Table 1), including Ekman's six basic emotions (Ekman, 1992a) and 34 complex emotions. These categories were selected based on their frequency in everyday life and on their even distribution across valence and arousal (Hepach, Kliemann, Grüneisen, Heekeren, & Dziobek, 2011). The videos were recorded at the film studio of the Humboldt University, Berlin, Germany in cooperation with its Computer and Media Service. The actors received specific emotion induction instructions that included situations in which the emotion to be portrayed occurs and information about physiological changes associated with the emotion. Video clips were validated by experts and showed high emotion recognition rates and good believability. Each video clip was cut to a length of 4 seconds and featured the actor's head facing directly towards the camera in front of a grey background (Figure 2). The actor first displays a neutral facial expression, moves to display an emotional facial expression of the selected emotion and subsequently returns to a neutral expression.

Figure 2. Six exemplary still frames of the video data set. The original labels for the displayed expressions from top left to bottom-right are: curious, bored, contemptuous, disgusted, enthusiastic and fearful.

2.8.3 Measures

The study data was analyzed using mixed effects models with random intercepts for observers and videos. In cases where including both random effects overly reduced power and interfered with estimation of fixed effects, one or the other was omitted, as described below. Dependent variables were self-rated difficulty (SRD) and objective difficulty (OD). Furthermore, ratings for basic emotions and interest (BEI; described below) served as dependent variables in models checking assumptions of the OD measure. Independent variables were actor age and sex, observer age and sex, ratings for BEI as well as valence and arousal.

The OD measure represents the Euclidean distance from an observer's rating to the population consensus on that video. The population consensus was calculated as the centroid in the seven-dimensional BEI rating space of each video. This distance therefore reflects how similar an individual observer's perception of a particular video item was to the average perception of that video. As expected statistical tests revealed an influence of the observer's sex on the rating dimensions (see Results). Such an effect was not present for age. Therefore, all analyses including observer sex as a predictor use an objective difficulty measure, which

calculated the distance from a separate centroid for each sex to account for the difference in rating behavior. In these analyses, 171 video clips rated by less than 3 male observers were excluded.

2.8.4 Procedure

Testing was carried out in German on the soscisurvey.de platform (Leiner, 2014). Each participant rated 12 videos randomly chosen from the pool of 480 videos. First, participants rated valence and arousal on scales anchored by pictures of the respective Self-Assessment-Manikin (Bradley & Lang, 1994). Second, participants provided ratings on the Basic Emotions and Interest (BEI) scales: happiness, sadness, fear, anger, surprise, disgust, interest, and then rated the subjective difficulty of making the BEI ratings. All ratings were given on continuous visual sliding scales ranging from "not X at all" to "completely X" where X was the rated emotion word. For the difficulty rating the extreme ends of the scale were "very easy" to "very difficult". All rating responses were encoded between 1 and 101.

2.8.5 Data analysis

Linear mixed effects models and ordinary least squares models were fit to test the previously stated assumptions and hypotheses. Across all models, *i* is an index iterating over participants, *j* an index iterating over actors and *k* an index iterating over videos. To check for a potential systematic influence of observer age and sex a model for each BEI rating was set up with the rating as the dependent variable and observer age and observer sex as independent variables. Equation (1) shows this exemplarily for the happiness rating dimension. The happiness rating for a video *k* rated by observer *i* is estimated by an intercept β_0 , equal to the mean happiness rating across videos and observers, plus a term β_1 for observer sex and a term β_2 for observer age. Additionally, the model contains a random effect β_{video_k} , which adjusts the intercept β_0 for each video *k* and therefore accounts for the repeated occurrence of the videos.

$$Happiness\ rating_{ik} = \beta_0 + \beta_{video_k} + \beta_1 * (observer\ sex_i) + \beta_2 * (observer\ age_i) + \epsilon_{ik} \quad (1)$$

Hypothesis 1 was examined with two mixed-effects models (Equation (2) and (3)), one for each difficulty measure (OD and SRD) as the dependent variable. Fixed effect variables were actor sex and actor age. A random intercept for observer was specified ($\beta_{observer_i}$) to account for the repeated measurements of each observer *i*. Since the variables of interest were actor-related and

each video stimulus was nested within actor, no random effect for the multiple appearances of video stimuli was included in this model. We instead assume that the fixed effects of actor sex and actor age already account for the common variance of a given video.

$$OD_{ij} = \beta_0 + \beta_{observer_i} + \beta_1 * \left(actor\ sex_j\right) + \beta_2 * \left(actor\ age_j\right) + \epsilon_{ij} \quad (2)$$

$$SRD_{ij} = \beta_0 + \beta_{observer_i} + \beta_1 * \left(actor\ sex_j\right) + \beta_2 * \left(actor\ age_j\right) + \epsilon_{ij} \quad (3)$$

In a similar way, two mixed-effects models (Equation (4) and (5)) were specified for Hypothesis 2 with observer sex, observer age and squared observer age as fixed effects and a random intercept term for the rated video. The simple and the squared observer age variables were mean centered to reduce collinearity of these predictors. Again, a random intercept term for observer could not be easily specified, because of nesting of observer age and sex under that variable. While this way of modeling does leave repeated measurements of the observers in the data, their influence should be accounted for by the fixed effects.

$$OD_{ik} = \beta_0 + \beta_{video_k} + \beta_1 * (observer\ sex_i) + \beta_2 * (observer\ age_i) + \beta_3 * (observer\ age_i^2) + \epsilon_{ik} \quad (4)$$

$$SRD_{ik} = \beta_0 + \beta_{video_k} + \beta_1 * (observer\ sex_i) + \beta_2 * (observer\ age_i) + \beta_3 * (observer\ age_i^2) + \epsilon_{ik} \quad (5)$$

Hypothesis 3 was examined with two mixed effects models (Equation (6) and (7)) using valence and arousal as independent variables. Random intercepts for video (β_{video_k}) and observer ($\beta_{observer_i}$) were included in the model to account for the repeated measurements of observers and the repeated occurrence of rated videos. After the first model fit, models with additional squared terms for valence and arousal were tested. Equation (6) and (7) show these final models. Valence and arousal entered the model as Z-standardized variables to reduce collinearity between the linear and squared terms and to ensure comparability of estimated coefficients. An additional multiple testing correction was applied for the subsequent inclusion and testing of the quadratic term (see next section).

$$OD_{ik} = \beta_0 + \beta_{rater_i} + \beta_{video_k} + \beta_1 * (valence_i) + \beta_2 * (valence_i^2) + \beta_3 * (arousal_i) + \beta_4 * (arousal_i^2) + \epsilon_{ik} \quad (6)$$

$$SRD_{ik} = \beta_0 + \beta_{rater_i} + \beta_{video_k} + \beta_1 * (valence_i) + \beta_2 * (valence_i^2) + \beta_3 * (arousal_i) + \beta_4 * (arousal_i^2) + \epsilon_{ik} \quad (7)$$

Additional exploratory models (Supplementary Equation (1)-(8)) were run to investigate results from the previous models.

2.8.5.1 Multiple testing correction

Multiple testing correction was applied to ensure a consistent alpha level of 0.05. Models built on Equation (1) were subjected to a 7-fold multiple testing correction to account for the repeated testing (separately for each rating dimensions) of the hypothesis that ratings were correlated with observer sex and/or age. Models of Hypotheses 1 and 2 were subjected to 2-fold multiple testing correction each to account for the common structure of OD and SRD models and the correlation in the outcome values. Models of Hypothesis 3 were subjected to a 4-fold testing correction for the aforementioned reason of common structure and because first models with linear predictors were fitted and additional models with squared terms were fitted subsequently. For all corrections the Bonferroni-Holm method was used.

2.8.5.2 Software

All analyses were conducted in RStudio (RStudio Team, 2015) using R 3.3.3 (Core Team R, 2017) under Windows 7. Mixed effects models were built using the "lme4" package (Bates D, Maechler M, Bolker B, & Walker S, 2015). Ordinary linear regression was conducted with the functionality of the base R package. P-values for mixed effects models were provided by "lmerTest" package version 2.0-30 (Kuznetsova, Brockhoff, & Christensen, 2016). Model output was arranged in tables using the R "stargazer" package (Hlavac, 2015).

2.8.5.3 Feature importance

Feature importance estimates the relative impact of a given predictor on the ability of a model to predict either difficulty measure. Feature Importance was calculated by the increase in node purity measured in Gini in a random forest model similar to work by Brick, Koffer, Gerstorf, & Ram (2017), and used the "caret" package (Kuhn, 2008).

2.9 Results

2.9.1 Assumption check and confirmatory analyses

For the calculation of our objective difficulty measure (OD) it was important to know if observers' sex and age significantly influenced their rating behavior on the basic emotions and interest (BEI). A significant influence of observer sex was found for all emotion dimensions except "happy" (Supplementary Table 2). The coefficients of the observer sex term are negative, indicating that women rate these dimensions systematically lower than males do. No significant effect for observer age could be shown. As a consequence, separate centroids for each sex were used to calculate the OD measure.

Table 1 shows models examining the relationship between the two difficulty measures and actor sex and age according to Hypothesis 1. For SRD a significant positive effect was found for actor age, indicating that increased actor age increases also the perceived difficulty of observers. For the OD measure a positive effect was found for actor sex and age, showing that OD increases with increasing actor age and is higher for female actors.

Hypothesis 2 was tested with the models depicted in Table 2. A positive association of SRD with female observer sex and a negative quadratic effect for observer age (Table 2.1) were found. The latter indicates that the youngest and oldest observers in the sample tended to rate items as less difficult than the middle age group. OD likewise showed a positive association with female observer sex but no significant association with observer age (Table 2.3).

Table 3 shows models estimating the influence of valence and arousal ratings on SRD and OD according to Hypothesis 3. For SRD (Table 3.1) a negative squared effect of valence and arousal was observed. This means that on average stimuli rated on the low or high ends of the valence or arousal scales are reported as easier than stimuli with ratings towards the middle on these scales. Comparison of the coefficients for the squared valence and arousal terms indicate a steeper slope for valence than for arousal. In the OD model (Table 3.3) a negative coefficient for the squared valence term was found, which indicates lower OD for extreme ratings on the valence scale in contrast to higher OD for ratings towards the center of the scale. However, the squared arousal predictor has a positive coefficient. This indicates a lower OD for stimuli rated as medium-level in arousal as opposed to stimuli rated towards the extreme ends of the scale.

Table 1. Mixed effects models for subjective and objective difficulty predicted by actor sex and age. Random effect for observer.

	Dependent variable:			
	Self-rated Difficulty		Objective Difficulty	
	Unstandardized	Standardized	Unstandardized	Standardized
Fixed effects	(1)	(2)	(3)	(4)
Actor female	0.36	0.01	**1.84***	**0.09***
	[-0.89, 1.61]	[-0.03, 0.06]	[0.57, 3.11]	[0.03, 0.16]
Actor Age	**0.06***	**0.03***	**0.06***	**0.04***
	[0.01, 0.11]	[0.01, 0.05]	[0.01, 0.11]	[0.01, 0.07]
Intercept	38.03	-0.01	49.54	-0.04
	[35.66, 40.41]	[-0.07, 0.05]	[47.47, 51.62]	[-0.10, 0.01]
Random effects	SD	SD	SD	SD
Observer	14.37	0.54	5.95	0.30
Residual	22.33	0.84	18.73	0.95
Observations	5,292	5,292	3,537	3,537
Log Likelihood	-24,338.78	-6,993.32	-15,513.14	-4,981.63
Akaike Inf. Crit.	48,687.57	13,996.65	31,036.29	9,973.25
Bayesian Inf. Crit.	48,720.44	14,029.52	31,067.14	10,004.11
Marginal R^2	0.001	0.001	0.004	0.004
Conditional R^2	0.294	0.294	0.095	0.095
Note:				*p<0.05

Table 2. Mixed effects models for subjective and objective difficulty predicted by observer sex and age. Random effect for video.

	Dependent variable:			
	Self-rated Difficulty		Objective Difficulty	
	Unstandardized	Standardized	Unstandardized	Standardized
Fixed effects	(1)	(2)	(3)	(4)
Observer female	**3.51*****	**0.13*****	**3.58*****	**0.18*****
	[1.96, 5.05]	[0.07, 0.19]	[2.33, 4.83]	[0.12, 0.25]
Observer Age	0.14	0.01	0.05	0.002
(mean centered)	[0.02, 0.27]	[0.001, 0.01]	[-0.06, 0.15]	[-0.003, 0.01]
Observer Age²	**-0.01*****	**-0.001*****	0.002	0.0001
(mean centered)	[-0.02, -0.01]	[-0.001, -0.0003]	[-0.004, 0.01]	[-0.0002, 0.0005]
Intercept	38.94	-0.06	49.91	-0.13
	[37.402, 40.472]	[-0.11, 0.002]	[48.51, 51.32]	[-0.20, -0.06]
Random effects	SD	SD	SD	SD
Video	7.43	0.28	8.02	0.41
Residual	25.42	0.96	17.90	0.91
Observations	5,292	5,292	3,537	3,537
Log Likelihood	-24,793.15	-7,453.49	-15,409.75	-4,883.73
Akaike Inf. Crit.	49,598.29	14,918.99	30,831.49	9,779.47
Bayesian Inf. Crit.	49,637.73	14,958.43	30,868.52	9,816.49
Marginal R^2	0.01	0.006	0.01	0.01
Conditional R^2	0.08	0.084	0.17	0.17
Note:				***p<0.001

Table 3. Mixed effects models for subjective and objective difficulty predicted by valence and arousal ratings. Random effects for video and observer.

	Dependent variable:			
	Self-rated Difficulty		Objective Difficulty	
	Unstandardized	Standardized	Unstandardized	Standardized
Fixed effects	(1)	(2)	(3)	(4)
Valence (standardized)	**2.44*****	**0.09*****	**-6.04*****	**-0.30*****
	[1.67, 3.20]	[0.06, 0.12]	[-6.73, -5.36]	[-0.33, -0.26]
Valence (standardized)²	**-9.29*****	**-0.35*****	**-1.48*****	**-0.07*****
	[-9.99, -8.58]	[-0.38, -0.32]	[-2.09, -0.88]	[-0.10, -0.04]
Arousal (standardized)	0.36	0.01	0.71	0.03
	[-0.28, 1.01]	[-0.01, 0.04]	[0.14, 1.27]	[0.01, 0.06]
Arousal (standardized)²	**-2.30*****	**-0.09*****	**1.14*****	**0.06*****
	[-2.97, -1.62]	[-0.11, -0.06]	[0.57, 1.72]	[0.03, 0.08]
Intercept	51.98	0.44	55.75	0.01
	[50.28, 53.68]	[0.37, 0.50]	[54.56, 56.94]	[-0.05, 0.07]
Random effects	SD	SD	SD	SD
Video	4.45	0.17	5.18	0.25
Observer	13.96	0.53	7.22	0.35
Residual	19.84	0.75	16.83	0.82
Observations	5,292	5,292	5,292	5,292
Log Likelihood	-23,842.71	-6,506.34	-22,868.14	-6,913.66
Akaike Inf. Crit.	47,701.43	13,028.68	45,752.28	13,843.31
Bayesian Inf. Crit.	47,754.02	13,081.27	45,804.88	13,895.90
Marginal R^2	0.13	0.13	0.12	0.12
Conditional R^2	0.44	0.44	0.32	0.32

Note: ***p<0.001

2.9.2 Exploratory analyses

Exploratory analyses were performed to complement the confirmatory results and provide further grounds for interpretation of the data. Because of the exploratory nature no p-values are provided. The relationship of both difficulty measures with valence and arousal was plotted for mean values for each video clip (Figure 3) which provides a simple way of visualization that the mixed effects models do not allow for easily.

We fit linear models to all relationships (Supplementary Equation (5)-(8)). Models including both a simple and a quadratic term for valence showed a good fit, indicated by adjusted R^2 values of 0.37 and 0.51 for dependent variables SRD and OD respectively (Figure 3 a and b). We also fit models containing a simple and a quadratic term for arousal and SRD and OD respectively as the dependent variable to the data. However, the fit was poor as indicated by the very low adjusted R^2 values of 0.05 and 0.01. Graphs c and d in Figure 1 show these much less obvious patterns in the data.

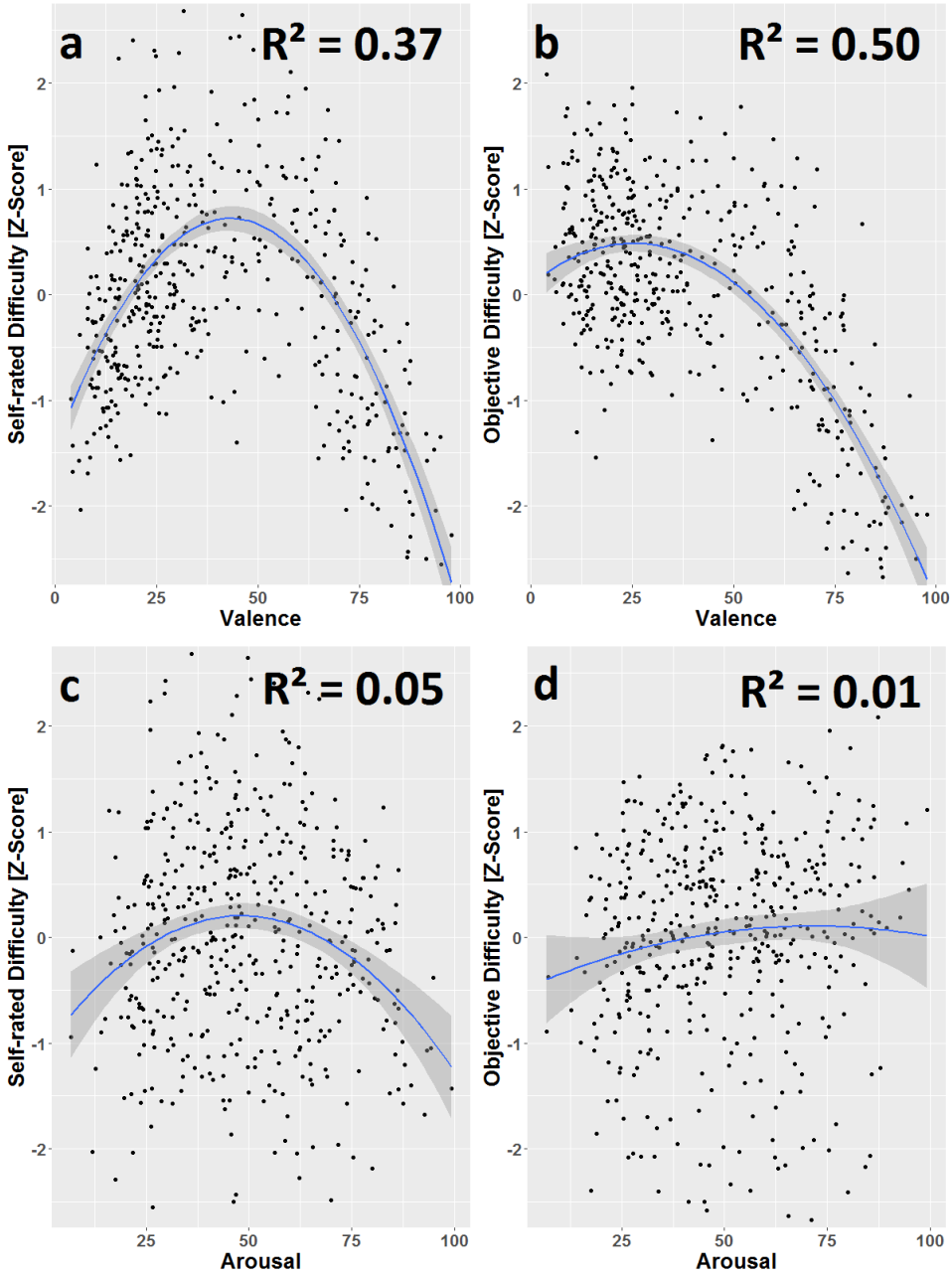


Figure 3. Plots visualizing the relationship of valence and arousal with both difficulty measures. Data points are averages over videos. The ribbon depicts the 95% confidence interval. A clear negative curvilinear relationship can be seen between valence and SRD (a). A similar relationship exists for valence and OD (b), although, the curve is less symmetrical with the high valence region featuring the lowest OD values. For the arousal measure the data does not exhibit such a visually clear relationship with the OD or SRD measure; this is also indicated by the low adjusted R^2 values (c,d).

2.9.2.1 Feature importance comparisons of all variables

Feature importance scores are difficult to interpret in raw form, but can be used as an intuitive guide to the relative impact of different predictors. Here, we present importance relative to the most important predictor (Figure 4), such that a score of 1 indicates the most powerful predictor, and a score of .5 indicates that a given predictor carries half the predictive power of that most powerful predictor. Feature importance calculations show that the valence rating is the strongest predictor for SRD (Figure 4a), followed by the happy rating with 74% of the importance of valence. Most other predictors improve accuracy by 56% (interested rating) to 42% (disgusted rating) as much as valence does. The exceptions are the predictors actor age (30%), observer gender (10%) and actor gender (7%) which seem of rather low importance for the SRD prediction.

For the OD measure the best predictor is the happy rating (Figure 4b), followed closely by the interest rating (93% of happy rating) and the valence rating (79% of happy rating). Most other predictors improve accuracy by about 62% (fearful rating) to 50% (arousal rating) of the happy rating predictor. However, the predictors of lowest importance are all person-specific variables: observer age (36%), actor age (31%), actor gender and observer gender (6%).

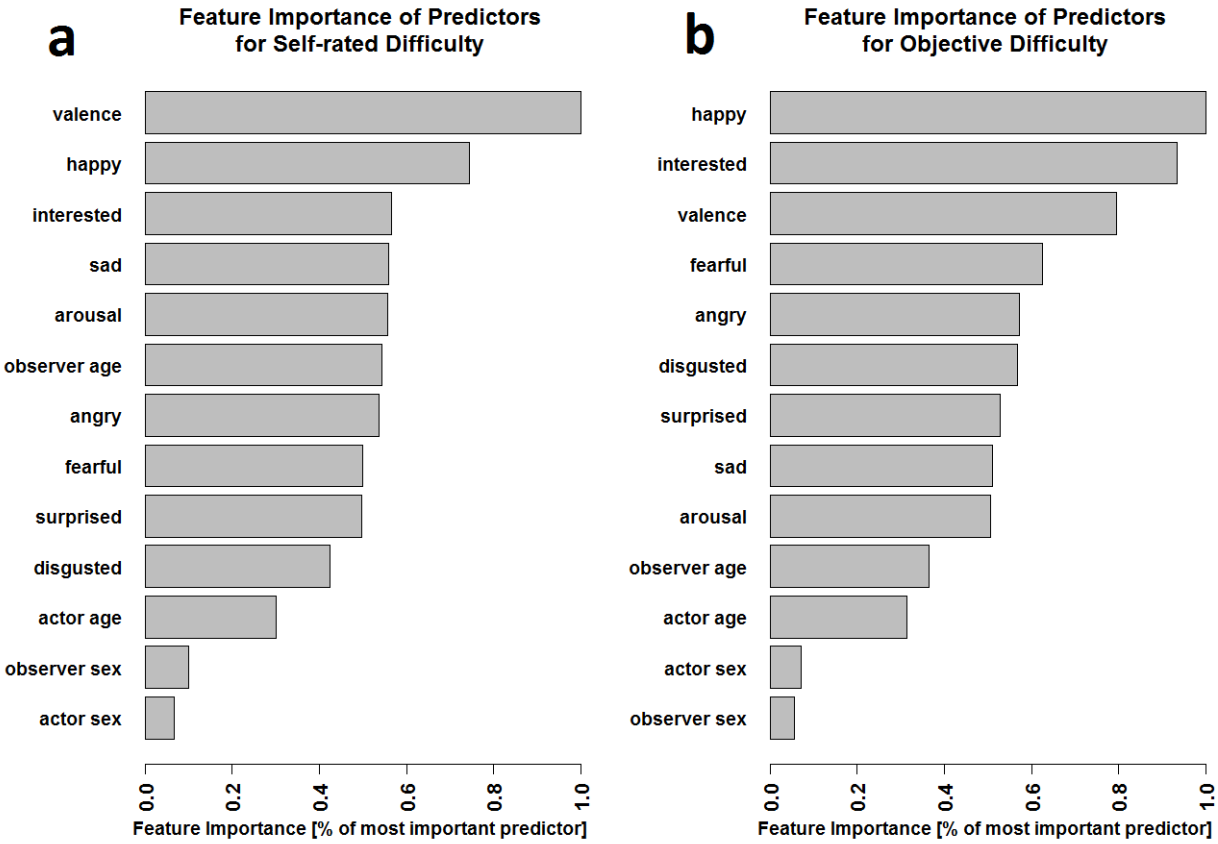


Figure 4. Feature importance ranking for the prediction of SRD (a) and OD (b). The importance is expressed as a proportion of the importance of the strongest predictor and displayed in decreasing order from top to bottom. For the SRD measure the valence and happy ratings show a high importance, while all of the observer- and actor-specific variables except observer age show a particular low importance. For the OD measure the happy rating is the most important predictor with the interest and valence ratings following up close in importance.

2.10 Discussion

The aim of our study was to investigate which person-specific (age and sex of actors and observers) and stimuli-specific variables (valence and arousal) constitute the subjective and objective difficulty of the perception of emotional facial expressions.

2.10.1 Age and sex of actor

Hypothesis 1 stated that the age and sex of the actor have an effect on the subjective and objective difficulty experienced by the observer. In line with previous research (Fölster, Hess, & Werheid, 2014), we found that emotions displayed by older actors were rated on average as

more difficult to judge and that ratings for older actors were more dispersed, which indicates more disagreement between observers.

An effect for actor sex was only found in the OD model. Interestingly, it described a higher OD for the judgment of expressions from female actors. While the literature describes greater facial expressivity in women (Fischer & Lafrance, 2015; Kring & Gordon, 1998), it is unclear how this relates to the observers difficulty of emotion decoding. In a conceptualization of categorical emotion, greater expressivity should result in the stronger expression of the primary emotion, and therefore a clearer signal that can be more easily interpreted. In a multidimensional emotion perception framework greater expressiveness could be evident in either of two ways. Similar to the categorical case, it is possible that greater expressiveness results in a stronger expression of the primary emotional signal, which should result in lower difficulty for the observers. However, if greater expressiveness results in the stronger expression of all present emotional signals or in the additional expression of further secondary signals, it might instead render the resulting expression more complex and therefore more difficult or ambiguous to decode.

Results from additional exploratory models (Supplementary Equation (1), (2)) are in fact consistent with this possibility. These show that on average 23.93 more rating points (Supplementary Table 3.1) were spent describing the expressions of female actors and that the standard deviation across ratings was 2.23 units higher for female actors (Supplementary Table 3.2). This indicates that female actors elicited a stronger but also more complex emotional response in the observers.

2.10.2 Age and sex of observer

According to Hypothesis 2 the observer's age and sex should influence both difficulty measures. We reasoned that difficulty should follow a quadratic function to account for developmental (Herba et al., 2006; Montirosso et al., 2010; Thomas et al., 2007) and old-age effects (Isaacowitz & Stanley, 2011; Ruffman et al., 2008) on both sides of the age spectrum. In fact, we found no quadratic effect of age on the OD measure and a *negative* quadratic effect for the SRD measure. Younger and older observers within the age range of our sample rated themselves as having on average less difficulty than the middle-aged observers, when in the OD model there was no effect for the same age coefficients. Although a difference in significance patterns between the models does not mean a significant difference between the results, this is consistent with the idea that younger and older people might overestimate their

perceptual abilities, which has been discussed before in the metacognition literature (Palmer, David, & Fleming, 2014).

Although our study features a broad range of ages (observer age range: 16-56) the majority of our sample consisted of young adults (mean observer age: 28.6, SD: 8.4). Twenty-five people older than 60 were deliberately excluded as they were thinly distributed between 60 and 80 and thus posed a considerable risk of biasing our results due to their high leverage. The limitation of range, however, should if anything attenuate the strength of the result—a future study with a wider age range may show a stronger age effect.

Female observers both rated themselves to have more difficulty and exhibited more difficulty according to our OD measure. This first falls in line with studies that show that males exhibit overconfidence, i.e. more confidence than justified by their own abilities, in stock trading (Barber & Odean, 2001), test taking (Lundeberg, Fox, & Punćochaŕ, 1994) and the usage of technology (Hargittai & Shafer, 2006). It is therefore likely that the same effect carries over to the domain of emotion recognition. By contrast, it is unexpected that women also score lower on the objective difficulty measure, as the literature agrees on a female advantage in emotion recognition (A. E. Thompson & Voyer, 2014). One explanation may lie in the dispersion of ratings by female observers. It might be that female observers see more subtle signals in facial expressions but that these perceived secondary signals may not be generally agreed upon, thus resulting in more deviation from the general consensus.

Further exploratory analyses (Supplementary Equation (3), (4)) revealed that the standard deviation across rating scales was more than one unit higher for female observers (Supplementary Table 4.2) and that women assigned on average 2.6 more rating points on the respective target emotion for the BEI stimuli set (Supplementary Table 4.3), which is the only subset of our stimuli where a target emotion can be found in the rating scales. This is in line with the literature that suggests a subtle advantage of women for emotion recognition (Hoffmann, Kessler, et al., 2010; Montagne et al., 2005). This exact same rating behavior of greater mean response to target emotions with a higher standard deviation was also already observed by Hall and Matsumoto (2004). The greater standard distribution across ratings together with the higher OD values that we found hint at greater disagreement among female observers. It is unclear, however, what underlying mechanism is responsible for this.

Our study has therefore in part found effects known from the literature (greater target emotion attribution for women) and in part resulted in novel unexpected findings (greater objective difficulty for women). This should be an indicator that the traditional emotion

recognition paradigms might not be sufficient for the detection of some of the perceptual differences between the sexes and that further research should concentrate on novel approaches to uncover these.

2.10.3 Stimulus valence and arousal

Hypothesis 3 predicted an influence of valence and arousal ratings on both difficulty measures. For the SRD measure as the dependent variable both valence and arousal exhibited a negative quadratic relationship, indicating that stimuli rated on the extreme ends of the valence and arousal scales were also perceived as less difficult. The standardized coefficients reveal that the influence of valence is more than four times stronger than that of arousal. For the OD model valence showed again a negative quadratic relationship. Arousal, on the other hand, exhibited a positive quadratic relationship, meaning that high and low arousal expressions lead to higher values on the OD measure, i.e., more disagreement between observers. The magnitude of the quadratic effects for arousal and valence are similar, however.

In related research presented below (Section 3), we found that arousal ratings are correlated with the displacement from a neutral face. It has also been shown before that facial expressions of highly intense emotional states are often misclassified in terms of valence (an expression arising from a positive experience as one of negative valence and vice-versa) in the absence of further information, such as body posture (Aviezer, Trope, & Todorov, 2012; Aviezer et al., 2015). What we observe in the mixed effect model for the OD measure could be similar. There might be a certain level of intensity above which discrimination of facial expressions worsens again. On the other hand, too little deviation of the face from a neutral expression will result in very low arousal ratings, but potentially also provide an observer with little information on the emotional meaning, hence make it difficult to recognize the displayed expression.

Further data exploration on video clip averages revealed negative quadratic relationships between valence, arousal and both difficulty measures (Figure 3). The sign of the quadratic effect is the same in these linear models except for the relationship between arousal and objective difficulty (Figure 3d) which is now fitted as a positive quadratic function. This change in sign is likely due to the different grouping of the data. As the fit for both difficulty measures and arousal is bad, not much concern has to be given to this. Here, valence seemed to be more predictive in terms of explained variance of SRD (37%) and OD (51%) than arousal (5% and <1% respectively). These results are interesting, because they show that valence is the predominant predictor for both difficulty measures.

The asymmetry of the curves in Figure 1a and Figure 1b shows that high valence stimuli are the easiest, even easier than stimuli from the very low valence spectrum. This might be an artefact of our questionnaire emotion dimensions where the happy scale represents the only definite positive scale (angry, sad, fearful, disgusted being negative and interested and surprised neutral/ambiguous). In this limited set, the range of choice for positive emotions is decreased which could be reflected in the observers' difficulty perception (SRD) and also in the dispersion of overall ratings (OD). It is known that within basic emotion paradigms stimuli of the happiness category show the highest recognition rates and even often produce ceiling effects (Hess, Blairy, & Kleck, 1997; Hoffmann, Kessler, et al., 2010; Rump, Giovannelli, Minshew, & Strauss, 2009). This effect could also be partly responsible for the lower right-hand side of both discussed curves.

We believe, however, that the general shape of the curves reflects a true relationship of valence and difficulty, as it proved to be highly stable across various subsets of the data. For example, if the data are split into the basic emotions and the remaining "complex" emotions (thus removing "pure happiness" here), the effect of valence holds in both subsets of the data. Even when only looking at stimuli of a single emotion category the negative quadratic relationship of valence and difficulty still holds. For example, for the "surprised" category, which is not easily classified as either positive or negative, the individual stimuli are still distributed along the same curves. This strongly indicates that this effect is not driven by the influence of individual stimulus categories, our questionnaire design or an interaction of those. It seems to rather reflect an underlying phenomenon. Further research may be needed to understand the meaning and causes of this effect in more detail, but it indicates an interesting sort of nonlinearity in the detection of emotions with extreme valence.

2.10.4 Effect sizes and feature importance of predictors

Marginal R^2 values, which express the variance explained by the fixed effects, were close to zero and thus represent very weak effects (Cohen, 1992) in all models testing person-specific variables for their influence on both difficulty measures (Table 1 and 2), which implies a limited influence of these person-level effects. Conversely, marginal R^2 values for models containing valence and arousal (Table 3) were in the range of moderate sized effects. Accordingly, the feature importance analyses (Figure 2) showed that the best predictors for SRD were ratings of valence and "happy" whereas for OD "happy", "interested" and valence dominated, emphasizing once again the importance of a general pleasure dimension for emotion perception. For both difficulty measures, the person-specific measures showed low importance with the

exception of observer age in the SRD model. This implies that the impact of these person-specific features may be swamped by the difficulty differences between emotions, but that observer age may still have an important impact on how difficult emotion ratings are perceived.

2.11 General Discussion

Taken together, these results fall in line with the functional account of emotion (Dacher Keltner & Gross, 1999), which describes emotions as signals that carry survival-relevant information. This account would predict that emotional displays can be extreme in terms of movement and therefore might be rated high in arousal but if no clear value judgment can be made an emotional expression would remain difficult to decode. Thus, the high predictability of the difficulty of emotional decoding by the valence dimension alone might be because emotions are in essence valence signals and humans are inherently tuned to them.

The interested rating was the third strongest predictor for SRD and second strongest predictor for OD. The interest dimension may be selectively indicating expressions of "social emotions", i.e., emotions that are directed at another person, from those which do not require the presence of an interaction partner. These emotions are usually more subtle in their display and more dependent on context, which may make them more difficult to evaluate. Future work should investigate whether "social emotions" are in general more difficult and whether they can be separated by ratings of interest from "non-social emotions" to confirm this. Another explanation might be that interest ratings also separate between negative and positive emotions and thus act as a proxy for emotion valence. In fact, in our data valence and interest ratings are moderately (r=.4) and strongly (r=.63) correlated on the individual rating and video level, respectively (Supplementary Table 5 and Supplementary Table 6).

Overall, variables specific to the observer or the actor are of relatively low predictive value for the predictions of both difficulty measures with the exception of the age of the observer influencing SRD. This effect is very likely a byproduct of overconfidence behavior in the young and elderly as discussed before (Discussion: Age and Sex of Observer). It can be argued that the relatively low importance of person-specific variables results from the function of emotional expressions. That is, if individual characteristics were of great importance, expressions would likely be much less useful to either the actor or the observer.

2.12 Limitations

Although the relationship of valence and difficulty proved to be quite robust across multiple subsets of the data further studies are needed to confirm this relationship. Our OD measure demonstrates how the difficulty of emotion recognition can be captured in a multidimensional emotion framework and without assigning an *a priori* ground truth. We calculated the video clip ground truth from our sample of observers by averaging all ratings for a clip within a group (male or female). This should give a reasonable estimate of the population ground truth. However, the precision of the estimate is dependent on the sample size. Due to our study design each individual video clip was only rated by a small group of the total observer pool. On average 11 observers rated a video clip (min: 5, max: 18), which was further broken down for some analyses into male and female observers. However, because we were only interested in effects across videos, a low observer count on individual clips should not systematically bias our results. In fact, the decision to have observers randomly distributed over many video clips instead of letting all observers rate the same few video clips adds to the generalizability of our results as they were observed over a wide array of emotions and actors.

3. Project 2: Arousal Perception from Facial Expressions

The following two studies were conducted in collaboration with Timothy R. Brick and Isabel Dziobek. I designed the study, collected and analyzed the data as well as described and visualized the results under their supervision and with the help of their advice. As this was a joint effort, I use the first person plural ("we") in the following.

3.1 Research Motivation

Arousal is a frequently used and long-standing (Duffy, 1957) construct in psychology. For example, the core affect framework (Russell, 2003b; Russell & Barrett, 1999) describes affective states only along the dimensions valence and arousal. In emotion research this framework is often used to quantify subjective affect experiences or observed affective states, for example from facial expressions (Britton, Taylor, Sudheimer, & Liberzon, 2006). The valence axis of core affect space shows systematic patterns with facial expression. For example, we found that valence ratings have a strong correlation (r = .87) with happiness ratings in Project 1. The happiness of a facial expression in turn is mainly estimated from the mouth area in western cultures (Eisenbarth & Alpers, 2011; R. E. Jack et al., 2012; M. L. Smith, Cottrell, Gosselin, & Schyns, 2005), thus permitting an estimation of valence from facial expressions via estimates of happiness.

The question as to how arousal can be characterized seems more complex. Emotional states such as surprise, anger or fear are consistently located in the high arousal range of core affect space (Russell, 1980; Russell & Barrett, 1999). These discrete emotions have been linked to expressive facial affect, for example in the works of Ekman (1992b, 1992a). However, to the best of our knowledge no literature exists, that examines the qualities of a facial expression, which can be used to directly determine levels of arousal. This is intriguing, since the construct of arousal relates closely to emotional states and should therefore be visible in facial expressions. Currently, self-rated or ascribed arousal seem to be the de facto gold-standard to determine arousal (Mauss & Robinson, 2009). Schimmack and Grob (2000) stated that the arousal dimension is poorly defined and highlighted the spread of theoretical accounts of arousal. The construct of arousal thus seems to be an incomplete construct despite its prevalence in the recent and past literature of emotion psychology. It is therefore a suitable candidate to illustrate an approach for solving the problem of incomplete constructs.

This project shows how the construct can be extended by identifying features of facial expressions consistently related to observer ratings of arousal with the use of face tracking data.

Two studies were carried out as part of this project. In Study 1 we first identified several likely candidate features; we here describe how these features can be computed efficiently from face tracking data and explore their intercorrelations. Based on the exploratory results, we selected two features for confirmatory testing, and examined them in Study 2 within and between neurotypical (NT) individuals and individuals diagnosed with autism spectrum disorder (ASD) as a demonstration of the applicability and use of the method.

3.2 Arousal Cues from Facial Expressions

Observers can readily judge the arousal of a person displaying a facial expression, even in static images (e.g. Sato & Yoshikawa, 2007). The concept of arousal seems to be intuitively linked to physical activation. In fact, Russel and Feldmann Barrett (1999) used the term activation in their seminal paper on core affect to describe this concept and state that other names of the concept have been "energy, tension [and] activity" in various theories of emotion. Two of these terms, energy and activity, seem also to be tightly related to the idea of movement and dynamics, and indeed the literature shows that dynamic stimuli produce increased arousal ratings by observers compared to static stimuli (Detenber, Simons, & Bennett, 1998; Sato, Fujimura, & Suzuki, 2008; Sato & Yoshikawa, 2007). Thus motion information seems to provide important cues for arousal perception.

Several features of the dynamics, but also of the static aspects, of facial expressions might be used as cues for arousal by an observer, such as the distance from the neutral face. For example, a happy expression is typically characterized by an upwards pull of the corners of the lip and a rise of the cheeks instantiated predominantly by the zygomaticus major muscle of the face. The resulting shape changes of the lips and cheeks differ in the amount of displacement among different happy expressions, such as between small and large smiles. The distance of a facial expression to the neutral face is the total quantity of shape change across the face. Displacement in the context of facial expressions can be considered a static feature, because it can be assessed from a still image, given that the observer has sufficient knowledge about human faces to approximate what the actor's neutral face might look like.

Another potential cue for arousal could be the velocity of a facial expression. In the example of a happy expression, this is a measure of how fast the lips and cheeks move as the expression is made and relaxed. The velocity of a facial expression can be computed as the combined velocity of all parts of the face. Yet another arousal cue might be the acceleration of a facial expression, which corresponds to the change in velocity over time—a measure of how

suddenly the expression appears or disappears. Velocity and acceleration are dynamic features, because they measure characteristics of movement—an observer would need a sequence of images or a video to estimate them. Mathematically, the velocity is the first derivative of displacement with respect to time, and acceleration its second derivative.

All of these measures can be calculated at every time point during a facial expression, e.g. for every frame of a video (with minimal loss of information for the dynamic features in a calculation with discrete time steps). An observer of a facial expression, however, has the ability to provide only a single arousal rating for the whole expression, which hints at an aggregation of facial expression information across time. This may be similar to gist representations in memory (Thompson, 2014), which capture essential information of complex phenomena and guide decision making. It is unclear, however, which qualities of a facial expression remain in its gist representation and how these are utilized to give accurate arousal ratings. The aggregation process for facial movement could take any of several forms. For example, an observer could be most sensitive to the average movement over the entire expression, or only keep track of the fastest or furthest extent that the expression reaches. In terms of aggregating measures of the frames of a video this would correspond to averaging measures across frames or taking the maximum across frames respectively.

3.3 Face Processing in Autism Spectrum Disorder

Autism Spectrum disorder (ASD) is a developmental condition characterized in the DSM-5 by pervasive social dysfunction, stereotyped and repetitive behaviors and interests (American Psychiatric Association, 2013). A considerable amount of research has focused on face perception and emotional facial expression recognition in autism (Harms et al., 2010; Lozier, Vanmeter, & Marsh, 2014). Eye tracking studies have found less attention towards faces (Kirchner, Hatri, Heekeren, & Dziobek, 2011; Riby & Hancock, 2009) and increased attention towards bodies and objects in individuals with autism (Klin, Jones, Schultz, Volkmar, & Cohen, 2002). Individuals with ASD looking at faces have been found to show patterns of avoidance of the eye region and to predominantly focus on the mouth area (Jones, Carr, & Klin, 2008; Kliemann, Dziobek, Hatri, Steimke, & Heekeren, 2010; Klin et al., 2002), although these findings have also been challenged in the recent literature (Guillon, Hadjikhani, Baduel, & Rogé, 2014).

ASD is known to be accompanied by difficulties in emotion recognition, for example from facial expressions (Rump et al., 2009). Uljarevic and Hamilton (2013), found a mean

effect size of d=0.41 for the difficulty of emotion recognition in autism in a meta-analysis. Not much is known about the underlying mechanism of emotion understanding deficits. One possible explanation for the difficulties may lie in abnormalities in lower level sensory processing such as in motion perception.

3.4 Perception of Movement in Autism Spectrum Disorder

Although differences in movement perception between NT and ASD individuals are frequently reported, results about the exact nature of those differences are often contradictory. Some studies found a reduced sensitivity for motion detection in autistic individuals (Bertone, Mottron, Jelenic, & Faubert, 2003; Milne et al., 2002; Robertson et al., 2014), but others found an enhancement under certain conditions. For example, Foss-Feig, Tadin, Schauder, & Cascio (2013) found superior motion perception in children with ASD compared to NT children for small and large stimuli in a high contrast condition, but not in a low contrast condition. Some authors reasoned that an impairment might only exist for biological motion, and indeed some (Blake, Turner, Smoski, Pozdol, & Stone, 2003; Nackaerts et al., 2012), but not all, studies showed an impairment of biological motion perception in autism (Rutherford & Troje, 2012; Saygin, Cook, & Blakemore, 2010). Freitag et al. (2008) found a decreased activity in response to biological motion in temporal and parietal areas as well as in the anterior cingulate gyrus for ASD individuals. Additionally, they found an increase in reaction time for ASD individuals when viewing stimuli of biological motion. However, they attributed these differences to difficulties in higher-order motion perception or the integration of complex motion information in ASD, and not to the biological nature of the motion per se. In light of the presented evidence for abnormal motion, face and emotion processing associated with ASD, we expect that the perception of arousal from facial expressions also differs between the NT and ASD population.

3.5 Study 1: Measure Selection

Recent advancements in computer vision software allow frame-by-frame tracking of the entire face, and therefore provide an efficient and automatic way to measure facial movements. In the following such face tracking data are used to determine which features of facial expressions are predictive of raters' arousal judgments. First, we examined the correlation between various measures that could serve as arousal cues in an exploratory way and then tested two selected measures (average distance from the clip neutral face and average speed) on a separate data set with confirmatory analyses. The following describes the initial measures that seemed to be

likely candidates for arousal cues. The section provides detailed information on the calculation of these candidate measures. It then describes the process used to select the two measures that were later tested in the confirmatory analyses in Study 2.

3.6 Methods

3.6.1 Materials

3.6.1.1 Video data set

A data set of 120 video clips showing facial expressions from 40 different emotional categories (Supplementary Table 1) produced by three actors was used for measure selection. These video clips were part of the same large set of videos (Dorit Kliemann et al., 2013) described in Section 2.8.2 and are therefore equal in their properties.

3.6.1.2 Face tracking data

Tracking data for all videos was acquired using the software OpenFace (Baltrusaitis, 2018). OpenFace provides the x- and y-coordinates of 68 landmarks that are placed on the face for each frame of a facial expression video. All expressions were then normalized to a common frame of reference using Generalized Procrustes Analysis to remove differences due to the location on the frame or the overall size of the face in the video.

3.6.2 Measures

The measures that we investigated in this study are based on a measure of displacement and its time-derivatives. This initial measure of displacement must be relative to a "home base", here called the *baseline face*, from which the displacement is calculated. In our exploratory analysis, we examined two different baseline faces: the clip neutral face and the actor's mean face, which are described in the following.

3.6.2.1 Clip neutral face

For each clip the locations of points representing the neutral face of the actor featured in the clip were extracted from the first frame of the raw video clip. Even after the normalization procedure the neutral faces of an actor extracted from different video clips differed slightly in expression and angle towards the camera as is shown in Figure 5 (lower rows).

3.6.2.2 Actor mean face

The tracking data of all frames of an actor were averaged for each coordinate of each point to calculate a mean expression for each actor. Figure 5 (upper row) shows the mean actor face for 3 different actors. Differences in the shape of facial features, for example the mouth, and differences in the general expression are subtle but clearly visible. Importantly, the mean

expression differs from the neutral expression in that the mouth is often slightly open, and the lips turned either upwards or downwards depending on the particular actor.

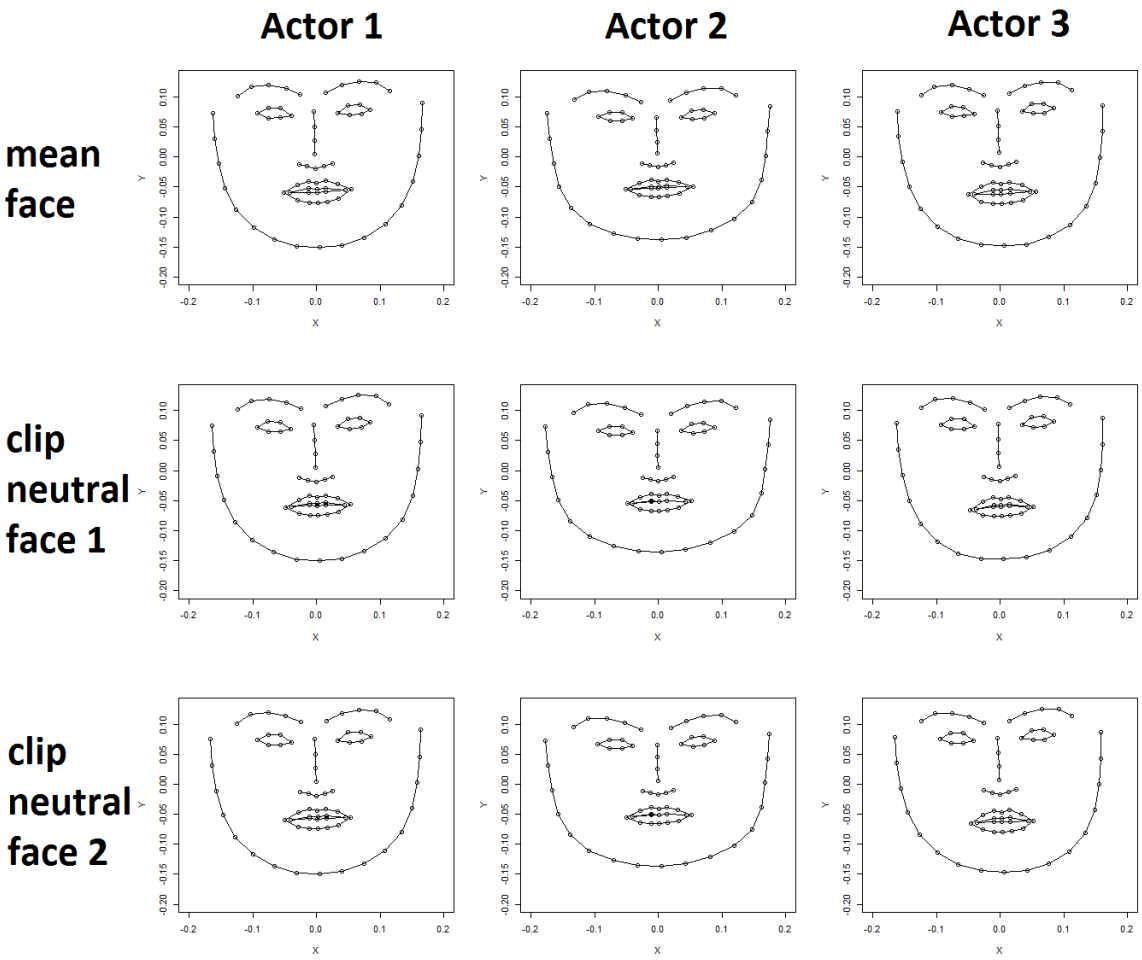


Figure 5. Actor mean face (upper row) and two clip neutral faces are shown for three actors (female, male, female). Clear differences in face shape can be seen between the actors. Differences between an actor's mean face (upper row) and their clip neutral faces (bottom rows) are less apparent but nevertheless visible, for example, in the shape of the mouth.

3.6.2.3 An example for derivatives

One common means of understanding displacement and its derivatives is using the example of a vehicle. Consider a person who rides a bicycle from point A to B. We can calculate their displacement by subtracting the location (in e.g. latitude & longitude) of point A from the location of point B. This displacement is a vector quantity (latitude, longitude) that has both a direction (e.g. north) and a length. The length of this displacement vector is called the (linear)

distance between point A and B and a scalar quantity. Assuming the cyclist biked in a straight line from A to B without needing to turn, this distance would be the value shown on the mileage counter on their bike.

$$\mathrm{d}(\mathrm{p}_{f,}\mathrm{p}_{f+a}) = \sqrt{\left(\mathrm{x}_{p_{f+a}} - \mathrm{x}_{pf}\right)^2 + \left(\mathrm{y}_{p_{f+a}} - \mathrm{y}_{pf}\right)^2} \tag{8}$$

3.6.2.4 Distance measure: Root mean squared deviation

Similarly as for the cyclist, the distance from a baseline can be computed for each face tracking point, now in terms of the distance along the x and y axes of the screen. Equation (8) shows this calculation, called the Euclidean Distance, for a single face tracking point *p* in two frames *f* and *f*+*a*. The result here is a single scalar value, representing the mileage counter for that point of the face. If the cyclist was just one of 68 cyclists traveling the roads at a given time, we would need to accumulate the amount of movement covered by all of them. A simple mean is insufficient because they might travel in different directions and it would be undesirable to have a cyclist traveling south to "cancel out" the efforts of one traveling north. To capture the simultaneous deviation of all face tracking points of a facial expression from a baseline face (the clip neutral face or the actor mean face), we therefore computed the root mean square deviation (RMSD; Equation (9)). The RMSD of a facial expression is useful in this context, because it expresses the total distance in structural movements between two automatically-tracked facial expressions, a good approximation of the total amount of movement required to change one expression into the other.

$$\mathrm{RMSD}(\mathrm{f}, \mathrm{f}+\mathrm{a}) = \sqrt{\frac{\sum_{\mathrm{p}=0}^{\mathrm{N}} \mathrm{d}(\mathrm{p_f}, \mathrm{p_{f+a}})^2}{\mathrm{N}}} \tag{9}$$

3.6.2.5 Speed measure: Root mean squared speed

If we knew how long it took the cyclist to get from point A to point B we could calculate their velocity simply by dividing their displacement by that time. Velocity is the change in displacement over time and has a directionality that indicates in which direction this change occurs. In our example, it occurs in the direction from A to B. The length of this velocity vector pointing from A towards B is called speed.

Similarly, we can compute the speed of each face tracking point. Equation (10) shows how the velocity $\overrightarrow{\mathbf{v_p}}$ of a point *p* can be approximated over a time interval from frame *f* to Frame *f+a*. For subsequent frames *a* corresponds to 1. It is easy to see then, that the speed of a point between subsequent frames is equal to the Euclidean distance and can therefore also be calculated by Equation (8).

$$\overrightarrow{\mathrm{v}_{p_{f,f+a}}} = \frac{\mathrm{p}_{f+a} - \mathrm{p}_f}{(\mathrm{f} + \mathrm{a}) - \mathrm{f}} \tag{10}$$

To quantify the overall speed between two frames the RMSD (Equation (9)) can now be used again. The resulting value is the root mean squared speed of all tracking points between those two frames. It follows from the Euclidean Distance, that resulting speed values cannot be negative and hence cannot cancel out, if tracking points are moving in opposite directions.

3.6.2.6 Acceleration measure: Root mean squared acceleration magnitude

Acceleration is the change in velocity over time. To calculate the acceleration for our cyclist we would need extra information, for example information about the location of some point S along the way of the cyclist and when they reached it. Then we could calculate the velocity of the cyclist between point A and S and between S and B. Then, we could use these two velocity vectors to calculate the acceleration of the cyclist between A and B by subtracting the first from the second and dividing the result by the time it took the cyclist to get from A to B.

Similarly, for each face tracking point *p* acceleration vectors between a frame *f* and the frame two frames after, *f+2*, were approximated by the velocities between the frames *f* and *f+1*, and *f+1* and *f+2*, as shown in Equation (11). This equation therefore describes the acceleration of a tracking point between a frame and the frame two frames after. To calculate the magnitude of the acceleration vector, the Euclidean distance (Equation (8)) of the respective velocity vectors can be used again. Root mean squared acceleration magnitude is therefore computed by computing RMSD(v_f, v_{f+2}) following Equation (9). Once again, the squaring of acceleration magnitudes means that positive and negative acceleration accumulate rather than cancelling each other out.

$$\overrightarrow{\mathrm{a}_{p_f,p_{f+2}}} = \frac{\overrightarrow{\mathrm{v}_{\mathrm{p}_{f+1},p_{f+2}}} - \overrightarrow{\mathrm{v}_{p_f,p_{f+1}}}}{(\mathrm{f} + 2) - \mathrm{f}} \tag{11}$$

3.6.2.7 Aggregation of measures

The measures described above were calculated frame-wise or between subsequent frames in the case of the speed and acceleration magnitude measures. This results in a sequence of values for each video clip. It is possible that people remember the average distance, speed or magnitude of acceleration over the course of the whole clip, implying that the mean would be an appropriate aggregation. Alternatively, it may be that the peak of these is the most memorable part, and therefore has the strongest influence. Accordingly, we used the mean and maximum function to aggregate the frame-wise values into a single value.

3.6.3 Data analysis

We first examined the correlation between the candidate measures of facial displacement and motion. The measures distance to the actor's mean face, distance to the actor's clip neutral face, speed and acceleration were calculated for all videos. These four types of measures were aggregated by the mean and maximum function each. This resulted in 8 variables with 120 observations, i.e. one for each video clip. Based on that, we calculated the Pearson correlation matrix of these variables.

3.7 Exploratory Results and Discussion

Figure 6 shows the Pearson correlations between potential arousal cue measures. From the correlation matrix it is evident that all distance measures are highly correlated with each other, with correlations ranging from $r = .72$ (average distance to neutral face and maximal distance to mean face) to as high as $r = .94$ (maximal distance to mean face and maximal distance to neutral face). Likewise, all measures of speed and acceleration are moderately to strongly correlated with values ranging from $r = .53$ (maximal speed to average acceleration) to $r = .95$ (average acceleration to average speed).

This clear divide between measures of static information, i.e. all distance measures and measures of dynamic information, i.e. measures of velocity and acceleration is of particular importance. High collinearity between predictors in a model is to be avoided, as it leads to unstable estimates of regression coefficients and complicates the interpretation of coefficients, since it is not possible to assign the explained variance in the dependent variable to one of the predictors. Therefore, our strategy was to select one measure from each block of correlations to use them in confirmatory analyses.

The correlation between the average acceleration measure and the average speed measure ($r = .95$) is especially noticeable, because such a strong correlation is not expected

from the general relationship of speed and acceleration. In order to avoid problems with collinearity, we chose not to include any acceleration measure into our analyses. For the same reason we also only picked one variable for distance and speed each. We chose the average distance to the clip neutral face as the measure to quantify the distance to a ground face and the average speed to quantify the speed information.

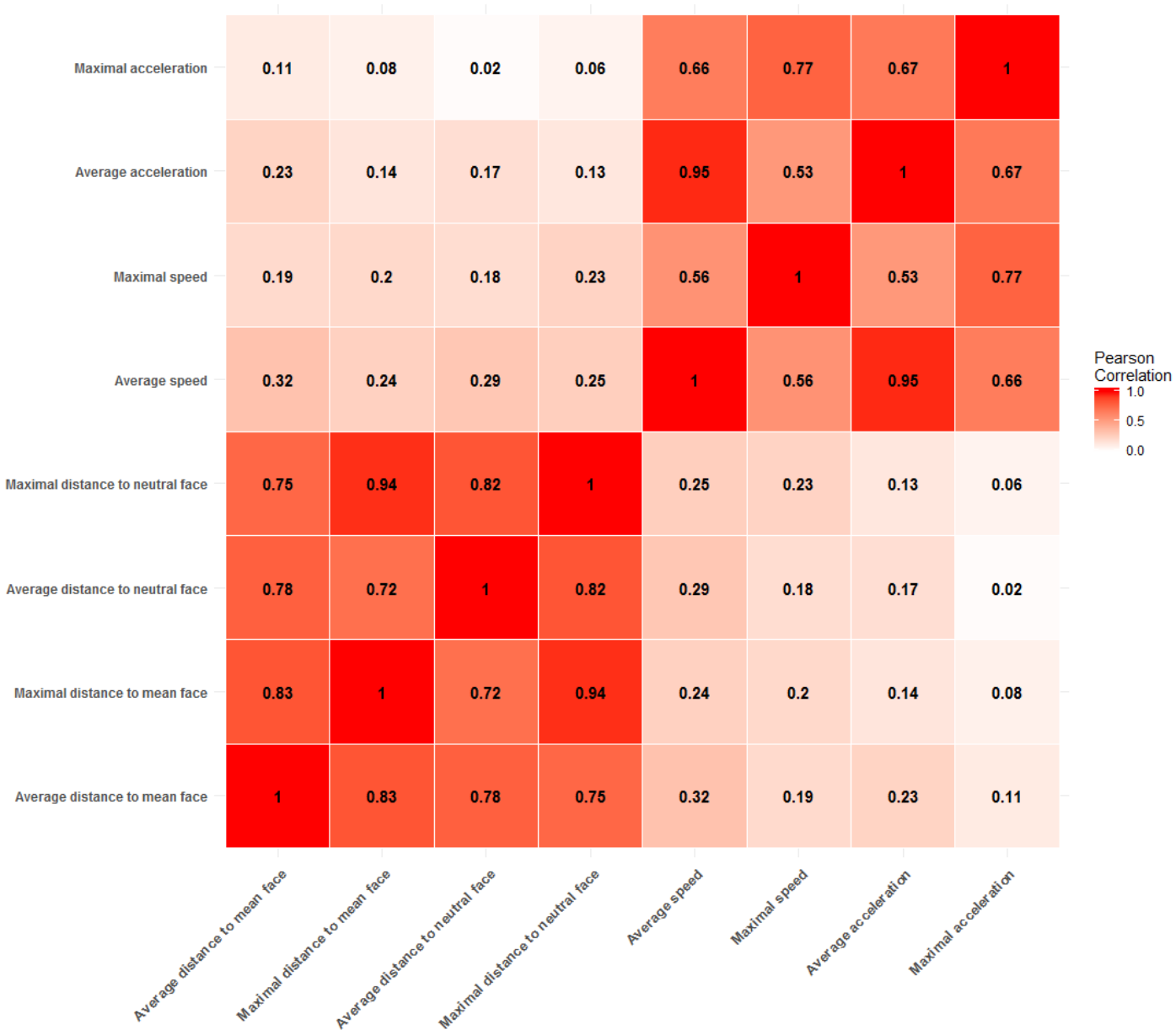


Figure 6. Correlation matrix of measures for potential arousal cues. Two blocks of moderate to strong correlations are clearly visible. One for all distance measures and the other for all speed and acceleration measures.

3.8 Study 2: Predictors of Arousal

The presented exploratory investigations of Study 1 were conducted to pick measures that could act as cues for arousal prediction. We examined a number of measures and their intercorrelation structure, and selected a subset of predictors that we deemed to be representative, meaning they

cover much of the explanatory power of other candidates, and distinct, meaning they do not overlap with the explanatory power of other chosen candidates. Our final feature set included only the average root mean square deviation to the clip neutral face as a measure of displacement and the average root mean square speed as a measure of speed. In Study 2 these measures were tested on their correlation with arousal ratings in NT and ASD individuals in confirmatory models.

We expected the distance to the neutral face and the speed of a facial expression to correlate with the arousal ratings from NT and ASD individuals. We furthermore expected the distance to the neutral face and the speed of a facial expression to correlate with the differences in arousal ratings between the groups. Additionally, we were interested in the arousal-specificity of the predictors. We deemed it possible that they could be general markers that provide information for a multitude of emotional judgments, for example also for the valence perception of the subjects. Therefore, we also tested for a correlation of the predictors with valence ratings within and between groups.

3.9 Methods

3.9.1 Participants

Four hundred and one neurotypical (NT) participants (115 males, mean age: 28.30 ± 8.44) were recruited in an online study on emotion perception through online advertisements. NT participants were included if they were either female or male native German speakers. Nineteen participants (10 males, mean age: 35.26 ± 10.42) with a diagnosis of autism spectrum disorder were recruited through the collaborating autism outpatient clinic of Charité – Universitätsmedizin Berlin. All of the participants were diagnosed according to ICD-10 criteria for Asperger syndrome and Autism (World Health Organization, 1993). The diagnostic procedure included the Autism Diagnostic Observation Schedule (Lord et al., 2000) and the Autism Diagnostic Interview – Revised (Lord et al., 1994), if parental informants were available (n = 11).

An additional sample of 41 (11 male, 2 transgender) German speakers with high autistic traits (HAT) were collected from an autism online forum (https://aspies.de/selbsthilfeforum). All 41 individuals self-reported that they had an autism diagnosis. They also scored above the cutoff of 32 in the Autism Spectrum Quotient (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) (M = 39.58, SD = 3.84). Given, however, that we could neither verify their diagnosis nor assess the diagnostic protocols or instruments that were used, this participant

group was only used in follow-up analyses to replicate our initial results. All of the participants gave informed consent via an online form before their participation, and the study was approved by the ethics committee of the Charité – Universitätsmedizin Berlin.

3.9.2 Materials

3.9.2.1 Video data sets

A data set of 80 video clips taken from the large video data set described in Section 2.8.2 (Project 1) was used. The video clips were cut to be 4 seconds long each, so that they contained the actor's peak expression. 12 actors (6 male, 6 female, age range: 21-64) which showed expressions from 21 emotion categories (Supplementary Table 1, third column) appeared in this subset of the data. Videos were selected to capture emotion categories that equally spanned emotion space and difficulty space, based on emotion and difficulty ratings collected for Project 1.

3.9.2.2 Face tracking data

Tracking data for all facial expression videos was acquired and processed as described in the measure selection section above. Videos used in the present study consisted of 100 frames each, resulting in 6800 coordinates per video and consequently 13600 data points per video.

3.9.2.3 Procedure

Data from NT and ASD participants were collected in separate data collection efforts. Initial data from NT participants were collected as part of Project 1 (Section 2.8.1), which contained a wider range of videos. Eighty representative videos were selected to be rated by ASD individuals. As a result, the number of videos rated by each individual in the NT group varied from 1 to 6 rated video clips, while each individual from the ASD group rated 14 video clips. All data collection was carried out in German on the soscisurvey.de platform (Leiner, 2014). Participants filled out a demographics questionnaire. Then each participant was presented with a sequence of videos randomly chosen from the pool of 80 videos. After each video, participants were asked to rate the valence (from "unpleasant" to "pleasant") and arousal (from "very calm" to "very aroused/excited"). Ratings were to be given on visual analog scales that encoded locations on the scale as integers between 1 and 101. After rating each video, participants were shown the next video.

3.9.3 Data analyses

Distance to the neutral face and facial expression speed were computed as explained in the measurement selection section and averaged across video clips. Valence and arousal ratings

were averaged per clip for the neurotypical and autistic group separately. Even though each participant only rated 14 video clips at most and as a result a considerable amount of missing data is present, this missingness is by definition *missing completely at random* because participants were assigned randomly to videos. Therefore, this missingness pattern is by definition not related to any variables included in the study, and the missingness pattern induces no bias in our results. The data were investigated with three linear regression models: one for each group (neurotypical individuals and individuals with ASD) and a separate model to specifically capture the differences between groups. Equation (12) shows the regression equation used for the separate models for each group. Here, arousal ratings from each group (NT or ASD) for a given video are predicted from the distance and speed measures computed for that video. Respective regression coefficients are estimated for each group as indicated by the group index *G*. Equation (13) shows the linear regression model that was used to investigate differences between the groups. The dependent variable is the difference in arousal ratings between the NT and ASD group. Again, distance and speed coefficients are estimated.

$$Arousal_G = \beta_{0_G} + \beta_{1_G} * Distance + \beta_{2_G} * Speed + \epsilon_G \quad (12)$$

$$Arousal_{NT} - Arousal_{ASD} = \beta_0 + \beta_1 * Distance + \beta_2 * Speed + \epsilon \quad (13)$$

3.9.4 Power estimation

To the best of our knowledge research similar both in approach and topic has not been conducted. Therefore, the estimation of expected effect sizes proved to be difficult. However, given that this study aimed to confirm what seemed to be the most likely, and therefore central, predictors of arousal from facial expressions, the assumption of an at least medium effect seems reasonable. Given a medium effect (e.g. $f^2 = .15$, Cohen J., 1988), a significance threshold of $\propto = .05$ and a desired power of 90%, a sample of 73 data points would be required for two-tailed tests of regression coefficients. The 80 data points of the video clip averages on which we conduct our analyses thus should guarantee power greater than 90% for each individual effect estimation that we perform. Confidence intervals of the estimated regression coefficients are reported in all cases to provide a measure of precision.

3.10 Confirmatory Results

3.10.1 Within-group results

Table 4 displays the results for the separate models for the NT and the ASD groups which were derived from Equation (12). It shows a statistically significant effect of distance in the NT model (p = .0272) and the ASD model (p = .0324). The speed coefficient is not significant in both models (NT: p = .3393, ASD: p = .1709). The estimated standardized effect size for the distance coefficient is 0.28 in the NT model (Table 4.3) and 0.27 in the ASD model (Table 4.4). This corresponds to an average change in 5.93 and 5.79 points on the arousal scale respectively for each standard deviation of distance under a constant speed term (Table 4.1 and Table 4.2). The NT model explains 13% and the ASD model explains 15% of the variance in arousal ratings of the respective groups. The correlation between the distance measure and the speed measure (r=.51) was higher than we expected from the analyses in Study 1, which raised the possibility that predictive variance is shared between the two measures.

Figure 7 shows the added-variable plots for the distance and speed predictors in the NT and ASD models. The plots for the distance predictor (Figure 7a and b) show three data points (blue) which are clearly separated from the main distribution and could thus be considered outliers. Calculating the mean without including them would put them at a distance of 4.9, 6.4 and 10 standard deviations from this new mean, which highlights the extremeness of these data points. These points also have high leverage scores (0.11, 0.17, 0.44; mean leverage: 0.038) resulting from their extreme values on the distance measure (x-axis) but not on the arousal measure (y-axis). Because they deviate from the trend apparent in the data they will have a strong influence on the slope of the regression coefficient for distance from the neutral face. The slopes of the estimated distance coefficients including (solid lines) and excluding (dotted lines) these outliers are shown for the NT and ASD model. One can see that without these three data points the slope would be more extreme, and hence the coefficients would be even larger than estimated by models on all data points. Specifically, without the outliers, the distance coefficient increases from 5.93 to 7.78 in the NT model (p = .00334) and from 5.79 to 7.73 in the ASD model (p = .00393). Predictors for speed remain non-significant in both models. Reasons for the extreme distance values are debated in the Discussion. There do not seem to be any outliers in regard to the speed measure.

Table 4. Regression models for the NT and ASD group predicting arousal ratings from distance to the neutral face and speed.

	Dependent variable:			
	Arousal rating NT	Arousal rating ASD	Arousal rating NT standardized	Arousal rating ASD standardized
	(1)	(2)	(3)	(4)
Distance [Z-Score]	**5.93***	**5.79***	**0.28***	**0.27***
	[0.77, 11.09]	[0.58, 10.99]	[0.04, 0.52]	[0.03, 0.51]
Speed [Z-Score]	2.53	3.67	0.12	0.17
	[-2.63, 7.69]	[-1.53, 8.88]	[-0.12, 0.36]	[-0.07, 0.41]
Intercept	47.60***	43.31***	-0.00	-0.00
	[43.19, 52.01]	[38.85, 47.76]	[-0.21, 0.21]	[-0.21, 0.21]
Observations	80	80	80	80
R^2	0.13	0.15	0.13	0.15
Adjusted R^2	0.10	0.12	0.10	0.12
Residual Std. Error (df = 77)	20.14	20.32	0.95	0.94
F Statistic (df = 2; 77)	5.53**	6.56**	5.53**	6.56**
Note:				*p<0.05; **p<0.01; ***p<0.001

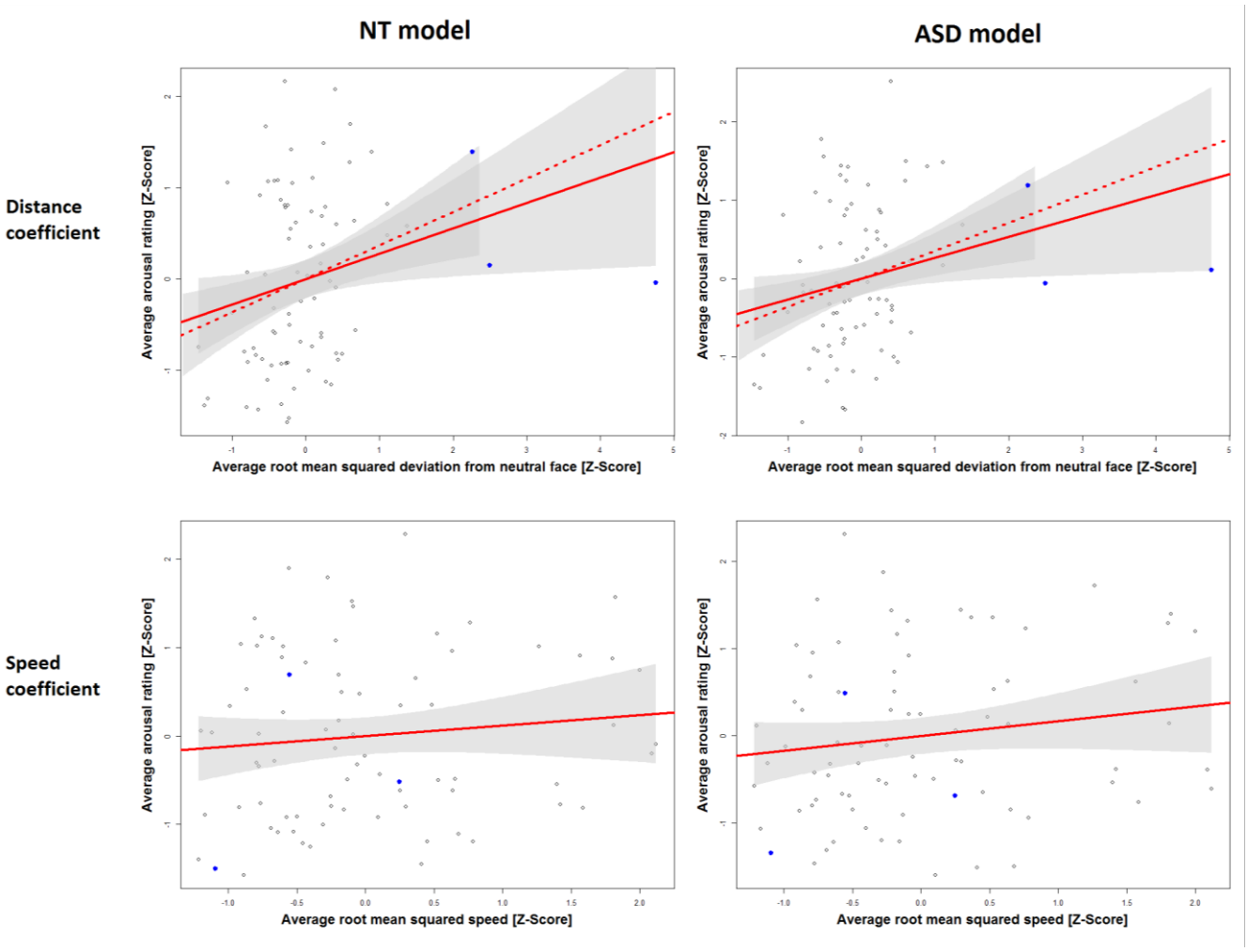


Figure 7. Added variable plots for the slope of the distance coefficient in the NT model (a), the distance coefficient in the ASD model (b), the speed coefficient in the NT model (c) and the speed coefficient in the ASD model (d).Three outliers on the RMSD scale are marked in blue. Plots a and b show the regression slope for the distance coefficient including these outliers (solid line) and excluding them (dashed line). Ribbons (grey) show 95% confidence intervals.

3.10.2 Between-group results

Table 5 shows the model of predicted mean arousal rating differences between the NT and ASD group as specified in Equation (13). The only significant term in the model is the intercept of 4.29 (p = .00219). Because the dependent variable of this model is a difference score this indicates that there is a significant difference between the average arousal ratings of the NT and the ASD group: individuals with autism rated videos on average as showing 4 points (out of 101) lower in arousal than neurotypicals. Coefficients for distance and speed of the stimuli are not significant, however, indicating no group differences in the strength of correlation between the displacement and velocity displayed in the video and ratings of arousal.

Table 5. Regression models predicting differences in arousal ratings between the NT and ASD group from distance to the neutral face and speed.

	Dependent variable:	
	Arousal rating difference	Arousal rating difference standardized
	(1)	(2)
Distance [Z-Score]	0.14	0.01
	[-2.97, 3.24]	[-0.25, 0.27]
Speed [Z-Score]	-1.14	-0.09
	[-4.25, 1.96]	[-0.35, 0.16]
Intercept	**4.29****	-0.00
	[1.64, 6.95]	[-0.22, 0.22]
Observations	80	80
R^2	0.01	0.01
Adjusted R^2	-0.02	-0.02
Residual Std. Error (df = 77)	12.12	1.01
F Statistic (df = 2; 77)	0.31	0.31
Note:		*p<0.05; **p<0.01; ***p<0.001

3.10.3 Specificity of predictors: analyses of valence

To gather evidence for the specificity of the tested predictors, we repeated our primary analyses with models as specified in Equation (12) and (13), but with valence ratings instead of the arousal ratings for the dependent variable. In these two within-group models and one between-group model (Supplementary Table 7 and Supplementary Table 8) none of the coefficients were significant.

3.11 Exploratory Results

3.11.1 Individual examination of predictors

Because we suspected the two predictors to share variability in the dependent variable, we estimated coefficients for them in separate models. Table 6 shows models for the NT and ASD group which only contain a predictor for distance to the neutral face. Here, the distance coefficient increased to 7.21 and 7.66 respectively and was significant in the NT (p = .00208) as well as the ASD model (p = .00133). Both of these models explain 12% of the variance in the dependent variable.

Table 7 shows models for the ASD and NT groups, which contain the speed measure as the only predictor. When not controlling for the effects of distance, the speed coefficients increased to 5.55 in the NT model and 6.62 in the ASD model, which were also significant (NT: p = .0194, ASD: p = .00596). Here, the NT model explains 7% of the variance in the dependent variable and the ASD model 9%.

Table 6. Regression models for the NT and ASD group predicting arousal ratings from distance to the neutral face only.

	Dependent variable:			
	Arousal rating NT	Arousal rating ASD	Arousal rating NT standardized	Arousal rating ASD standardized
	(1)	(2)	(3)	(4)
Distance [Z-Score]	**7.21****	**7.66****	**0.34****	**0.35****
	[2.78, 11.65]	[3.15, 12.16]	[0.13, 0.55]	[0.15, 0.56]
Intercept	47.60***	43.31***	-0.00	0.00
	[43.19, 52.01]	[38.83, 47.79]	[-0.21, 0.21]	[-0.21, 0.21]
Observations	80	80	80	80
R^2	0.12	0.12	0.12	0.12
Adjusted R^2	0.10	0.11	0.10	0.11
Residual Std. Error (df = 78)	20.13	20.44	0.95	0.94
F Statistic (df = 1; 78)	10.15**	11.08**	10.15**	11.08**
Note:				*p<0.05; **p<0.01; ***p<0.001

Table 7. Regression models for the NT and ASD group predicting arousal ratings from speed only.

	Dependent variable:			
	Arousal rating NT (1)	Arousal rating ASD (2)	Arousal rating NT standardized (3)	Arousal rating ASD standardized (4)
Speed [Z-Score]	**5.55*** [0.99, 10.10]	**6.62**** [2.03, 11.21]	**0.26*** [0.05, 0.48]	**0.30**** [0.09, 0.52]
Intercept	47.60*** [43.07, 52.13]	43.31*** [38.75, 47.87]	0.00 [-0.21, 0.21]	0.00 [-0.21, 0.21]
Observations	80	80	80	80
R^2	0.07	0.09	0.07	0.09
Adjusted R^2	0.06	0.08	0.06	0.08
Residual Std. Error (df = 78)	20.66	20.81	0.97	0.96
F Statistic (df = 1; 78)	5.70*	7.99**	5.70*	7.99**
Note:				*p<0.05; **p<0.01; ***p<0.001

3.11.2 Replication of results on a sample of individuals with high autistic traits

The sample size of the ASD group was small with N=19 individuals. We repeated calculation of the within-group model with both predictors (Equation (12)) and the models with individual predictors on data from the High Autistic Traits (HAT) group, who self-reported to have been diagnosed with autism. In this group the distance coefficient is also significant ($b_1 = 6.15$, p = 0.00916) in the model with both predictors and the coefficients for the distance and speed predictors are significant and increase when tested individually as previously seen in the ASD and NT group (Supplementary Table 9). We also calculated the difference model (Equation (13)) between the NT and the HAT group. We observe the same pattern of significance as found in the difference model between the NT and the ASD group, with only the intercept being significant ($b_0 = 4.64$, p = 0.00047) and none of the predictors (Supplementary Table 10).

3.12 Discussion

3.12.1 Predictors of arousal

We found the average distance to the neutral face to be significantly predictive of the average arousal ratings in the NT and the ASD group, whereas speed showed no significance as a predictor in any of the confirmatory models. An additional exploratory analysis with a bigger sample (N = 41) of individuals with high autistic traits (HAT) who self-identified as autistic reproduced the significance pattern of the NT and ASD group with only the distance coefficient being significant.

Taken together, these results confirm distance to the neutral face as an important predictor for arousal in neurotypical individuals and individuals with autism. Arousal ratings can be and are frequently given to static facial expression stimuli. Our results might explain how this is possible even in the absence of movement information. The results suggest that observers, at least in part, rely on the displacement of the face, a feature that we captured with the distance to the neutral face measure.

Furthermore, the greater-than-expected correlation between the distance and speed measures (r = .51) highlights an issue of shared variability between the predictors, which complicates the simultaneous estimation of the effects of both predictors. This is exemplified by the explained variances of the models containing both predictors, which are lower than the sum of the explained variances of the models containing the predictors individually. In our data, distance to the neutral face is the stronger predictor as it explains more variance in the dependent variable than speed when tested on its own. Because the correlation between measures was much higher in Study 2 than in Study 1, and because our exploratory analyses indicated that either measure is significantly predictive of arousal, however, we are cautious in interpreting the specific breakdown of predictive power provided by the regression model. In this light and despite its non-significance in our confirmatory models, facial movement speed should not be discarded as a potentially important cue for arousal ratings of facial expressions.

To properly estimate the effects of both predictors independently, one could find or artificially create facial expression data where they are uncorrelated. On one hand, if the posed nature of our video clips is the primary reason for the correlation structure between speed and distance, then a data set of spontaneous expressions might be helpful in determining the unique contributions of distance and speed in emotion processing. However, Cohn & Schmidt (2004) found that the correlation between amplitude and duration was actually stronger in spontaneous smiles (R^2 = .69) than in posed smiles (R^2 = .09). If a similar relationship exists between distance

(total change in amplitude) and speed (change in amplitude across duration), which this data suggests, efforts to research contributions of speed and distance to the neutral face independently in naturalistic data sets might actually be less effective, and researchers might be better advised to focus on artificially created stimuli. Importantly, however, Cohn and Schmidt's work focuses expressly on happiness, and the results may not be the same for other emotions. For example, expressions of surprise (which are formed rapidly) may rely more heavily on dynamic information than do smiles.

3.12.2 Group differences in arousal perception

We found the same pattern of significance in within-group models for the NT and ASD group. However, because this does not exclude the possibility of a significant difference in the size of the effects tested in the within-group models, we tested for group differences of distance and speed correlations with arousal. The only significant term in this model of difference scores between NT and ASD participants was the intercept. This means that participants with autism rated clips on average 4 points lower on the arousal scale, but it did not provide evidence for a difference in the strength of the distance and speed predictors between the groups. These results were replicated with the HAT sample, where the significance pattern stayed the same and the difference in arousal ratings as shown by the model intercept even increased slightly. These results indicate that individuals with autism make use of the same displacement and movement information as neurotypical individuals to judge arousal from facial expressions, and is consistent with the interpretation that arousal perception is qualitatively similar between groups.

The current data are not sufficient to understand the precise nature of the change in intercept. On the one hand, it could be that people with autism in general show more cautious rating behavior. However, that no difference in valence ratings were found between groups constitutes evidence against this explanation. On the other hand, it is possible that while there is no qualitative difference in arousal perception, there may be a quantitative difference, such that individuals with autism are biased towards perceiving less arousal overall. This idea is consistent with prior work showing that autism is accompanied by aberrant empathy and theory of mind (Baron-Cohen & Wheelwright, 2004; Dziobek et al., 2008). If these empathy processes enhance the apparent arousal of an expression, we would expect a shift of this sort. Future work is required to determine whether such an enhancement exists.

Individuals with ASD tend to avoid the eyes and focus instead more on the mouth when looking at faces (W. Jones et al., 2008; Klin et al., 2002). Similarly, evidence for a reduced

integration of facial feature information has been found for individuals with autism for moving face stimuli, even when they attended to the eye region (Shah, Bird, & Cook, 2016), which could also contribute to the lower arousal ratings of the ASD group. Assuming a model in which arousal information from different parts of the face is additive, either of these could explain lower arousal ratings in ASD, since in either case individuals with ASD do not attend to all the information present in the face. Both possible explanations and their respective contributions could be investigated in a study that correlates eye tracking data with arousal ratings. Another approach would be to systematically occlude parts of the faces that participants have to rate. Alternatively, measures similar to the ones used in our study could be calculated separately for upper and lower face regions.

3.12.3 Specificity of arousal predictors

Models that tested if the same predictors also predicted valence ratings within the NT and ASD groups or differences in valence ratings between the groups did not yield any significant coefficients. Also no mean difference in valence ratings between the groups was found. This suggests that the chosen predictors indeed have some specificity for arousal perception instead of capturing some general feature of facial affect.

3.12.4 Outlier videos

Three video clips were identified as outliers in terms of the relationship between the distance measure and arousal. Two of the corresponding three video clips show the same actor, which could suggest that some idiosyncratic facial expression patterns of this actor are responsible for the extreme values on the distance to the neutral face measure for these videos. In fact, the majority of both clips show the actor's head turned and eyes closed. Even though all tracking data were normalized in relation to a common coordinate system (Study 1; Methods: Face tracking data) artifacts from out-of-plane rotations of the head may still bias the computation of distance to the neutral face due to the focal length of the camera and the projection of the face onto the 2D plane captured by it.

Similarly, our facial expression tracking model has six points per eye, actively upweighting the eyes' influence. While this upweighting is effective at capturing the importance of the eyes in the understanding of facial expressions, it may also provide some inconsistency here. Specifically, closed eyes may increase the distance of a clip from the neutral expression while *reducing* the apparent arousal by implying sleepiness or lethargy. While we do not wish to draw inferences from two outliers, future work should examine the specific effects of head movements, and the locations of movements that can be related to arousal in

different situations. For example, eye-closing may carry different meaning than eye-opening for arousal. Similarly, it is plausible that head movements amplify the apparent arousal of an expression only when they are congruent with the expression. For example, turning ones gaze aside enhances the intensity of fear expressions, but reduces the intensity of anger expressions (R. B. Adams & Kleck, 2005). A similar relationship of perceived intensity and the interaction of head movement and emotional expression is conceivable.

Finding an explanation for the outlier video closest to the main point cloud is not as straight forward. It shows the emotion category *enthusiasm* and differences from the three other videos displaying the same emotion in the data set are not immediately obvious. However, given its high average RMSD values it would be predicted to have higher arousal ratings. The expression of the actor seems less natural, i.e., more staged, compared to the other videos of the same emotional category, which might have led raters to assume that much of the activation displayed was a result of emotional masking (Brick, Staples, & Boker, submitted).

3.12.5 Design choice and future work

Our selected stimuli showed facial expressions without any distractors—the background is solid gray and each video contains only the actor. Yet distractors frequently appear in situations when the interpretation of facial expressions is crucial, for example in social interactions, and might alter their perception. Distractors may be especially problematic for individuals with autism, who may have difficulty separating out distracting stimuli (Adams & Jarrold, 2012; Christ, Kester, Bodner, & Miles, 2011) and focusing their attention towards facial expressions (Klin et al., 2002; Riby & Hancock, 2009). Future work including distractors might find stronger differences between ASD and NT groups.

Although our results provide insight into the predictors of arousal, it has to be noted that all presented models only explain a small part of the variance of the arousal ratings. This poses the question how the unexplained part of the variance can be accounted for. It is likely that some part will be due to intersubjective differences in arousal perception and due to noisy measurements. However, features of facial expressions predictive of arousal other than the ones presented in this study are conceivable. For example, autonomic blood responses, as seen in skin tone might be such a feature as it is frequently paired with high-arousal states in everyday language (for example: “red with anger”, “to pale with fear”, “to blush with shame”). The recognition of certain emotional categories might themselves lead to an inference on arousal. Other features from the tracking data, such as differences in face size and rotation that were

removed by the normalization procedure may also influence arousal perception on their own. Future studies should investigate these variables separately and with appropriate stimuli.

Our study investigated facial correlates of perceived arousal as rated by the observers. Another important question is how the self-rated arousal of a person relates to the measures calculated from the same person's face. This research requires a strong emotion induction in the participants, but can be conducted according to our design otherwise. It has been suggested (Schimmack & Grob, 2000) that up to three separate arousal dimensions might be necessary to account for the full breadth of theoretical accounts on arousal. Differences between these and how they relate to facial movement might be teased apart by using our design with rating instructions that reflect each of these theoretical accounts of arousal.

3.12.6 Limitations

Our study has some limitations due to its design. Although our participants were randomly assigned to videos, the design does not allow isolation of possible causes. As a result, no definite statement of causality can be made. Our sample also shows an average age difference (with the ASD group 6.97 years older on average) and a differential sex breakdown, with 29% male in the NT group and 53% in ASD. Both sex and age have known associations with general emotion perception (Isaacowitz & Stanley, 2011; Ruffman et al., 2008; A. E. Thompson & Voyer, 2014), and it is possible they have some effect on the group differences. However, conceptually it is not clear if and how these effects would be related to the task of arousal rating. At the individual rating level, however, the correlation of arousal rating with age is r = -.004 for the NT sample and r = -.02 for the ASD sample, and the correlation with sex is r = -.006 and r = -.10 respectively. Given such low correlations, we suspect that any bias induced by these demographic differences would be quite small.

4. Project 3: Relationship Between Facial Expressions and Emotion Perception

The following study was conducted in collaboration with Timothy R. Brick and Isabel Dziobek. I designed the study, collected and analyzed the data as well as described and visualized the results under their supervision and with the help of their advice. As this was a joint effort, I use the first person plural ("we") in the following.

4.1 Research Motivation

Facial expressions have been studied in the context of emotion recognition abilities and differences thereof between populations, such as, for example, age groups (Ruffman et al., 2008), neurotypical and clinical populations (Harms et al., 2010; Kohler et al., 2003) or between males and females (Kret & De Gelder, 2012). Some studies find a sex difference in both emotion recognition accuracy and sensitivity (Montagne et al., 2005), whereas others only find a difference in sensitivity, i.e. an advantage of women to recognize even subtle emotional expressions (Hoffmann, Kessler, et al., 2010). These findings hint at a small effect, ceiling effects or precision limitations in the employed paradigms. In fact, a meta-analysis of 551 effect sizes concluded that women have a small advantage over men in the recognition of non-verbal displays of emotion (mean Cohen's d= 0.19) (A. E. Thompson & Voyer, 2014). The estimated effect size varied with specific emotion categories (for example d=0.15 for surprise and d=0.25 for sadness) and sensory modalities of the stimuli (for example d=0.17 visual-only and d=0.38 for both auditory and visual information presented subsequently). Interestingly, it also has been shown that women exhibit higher variability on emotion rating scales than men, while still being more accurate in their emotion attributions (Hall & Matsumoto, 2004). This could potentially indicate that women perceive additional variation in facial expressions that men do not. While these results indeed indicate a difference in the emotion recognition capabilities between men and women, they do not reveal the nature of these differences.

Several possibilities exist that would be consistent with the current state of knowledge. For example, the difference in emotion recognition abilities could be purely quantitative in nature, meaning that men perceive emotional facial expressions in the same fashion but are simply less sensitive to them. Alternatively, the differences could be of qualitative nature, meaning men and women interpret the same facial expression signals in a systematically different way with regard to their emotional content. Finally, the differences between the sexes

could be some combination of quantitative and qualitative, possibly to varying degrees across emotion categories. The current emotion recognition paradigms might hamper detection and investigation of these differences for a number of reasons related to the representations that they use for emotions. Research on emotion perception differences is therefore closely linked to the problem of optimal representation as introduced in Section 1.5.

In this project, we propose to distinguish between emotion perception and emotion recognition and suggest to investigate the former instead of the latter. We define perception as the process that translates sensory information acquired from an object of the real world, i.e. a stimulus, into a mental representation, as described by Ernst and Bülthoff (2004). This is in contrast to recognition, which refers to matching the mental representation of a stimulus to a pre-existing mental representation in a pattern recognition sense as described by Newen, Welpinghus and Juckel (2015). Therefore, perception necessarily precedes recognition because only after a stimulus is translated into a mental representation, it can be compared to other representations. If emotion perception is concerned with the relationship between facial expressions and the mental representation they evoke, then research has to be careful in considering how these two endpoints are investigated. In the following, we explain why traditional emotion recognition paradigms may not be ideal to measure quantitative or qualitative differences in emotion perception between populations. We introduce a method from the neuroscience literature called Partial Least Squares analysis, which we put forward as a general solution to the problem of optimal representation. We demonstrate this method by investigating emotion perception of facial expressions within males and females as well as perception differences between the sexes. The technique also allows us to use face tracking in order to provide a detailed quantification of facial expressions over time.

4.2 Emotion Representations

Emotion representations are central to the research of emotion perception from facial expressions. The following will examine the representations relevant to both ends of the emotion perception process, the face and the impression of the observer, and identify problematic as well as desirable properties of these representations.

4.2.1 Representations of facial expressions

An important aspect for the research on perceived emotion is the conceptualization of emotional facial expressions. Facial expression stimuli used in emotion research can be static (pictures) or dynamic (video). Dynamic facial expression stimuli offer a range of benefits over static

stimuli, such as faster and more accurate emotion recognition, higher emotional intensity and better distinction between fake and genuine expressions (Calvo et al., 2016; Krumhuber et al., 2013) apart from the obvious greater natural validity. Traditionally, facial expressions in emotion research are either investigated as a whole or in respect to their features. If expressions are investigated as a whole, different facial expression stimuli are usually distinguished and labeled according to an emotion framework. Here, the aforementioned frameworks are usually employed again. However, facial expressions can and are also investigated according to the features that constitute them. Feature-oriented facial expression representation systems offer the advantage that they provide a more detailed quantification of the face which can be mapped back to specific facial expressions. The most popular example of a feature-oriented system is the Facial Action Coding System (FACS) (Ekman & Friesen, 1978). The FACS describes the face in terms of the activity of groupings of facial muscles, so-called Action Units (AUs), and their respective intensity. With this system, an expression of, for example, happiness would be described as 6B, 12D indicating slightly (B) raised cheeks (6) and extremely (D) elevated lip corners (12).

FACS coding is a manual process that requires multiple certified encoders, which investigate facial expression pictures or videos. Naturally, this is a very time-consuming and expensive procedure, especially for videos which often have to be examined frame by frame. However, specific computer vision software enables automated FACS coding (Baltrusaitis, Mahmoud, & Robinson, 2015), although usually only some AUs are supported while others are not. Recent advancements in computer vision software also allow frame-by-frame tracking of the entire face, so-called face tracking (Zadeh, Lim, Baltrušaitis, & Morency, 2018). Here, the face is captured as a set of landmark coordinates over time. This provides an efficient and automatic way to measure facial movements. Moreover, this whole-face representation is largely free of the human bias of an encoder or the assumptions of a specific emotion representation system. An apparent problem with face tracking is the high-dimensionality of the data it produces. This might be the reason why, although the technology has been available for some time, no wide-spread use in emotion research can be observed. High-dimensional data can be aggregated into simple measures as exemplified by Project 2 (Section 3.6.2), however this comes at the cost of a loss of information. Thus, it would be advantageous to directly use these high-dimensional face representations.

4.2.2 Representation of emotion impression

The other side of emotion perception from facial expressions concerns the emotional impression that is elicited in the observer. Major theories of emotion and their specific limitations have already been discussed in detail (Section 1.3, 1.5 and 2.2). Basic emotion theory is prominently employed to investigate emotion recognition differences. However, recognition differences are not ideal to find differences in the perception of emotions between populations for a number of reasons. Recognition differences between populations can hint at perceptual differences between populations in the best case but they cannot inform us about their nature—that is which specific facial features cause different perceptions.

It is important to notice that discretization to a limited number of categories (basic emotions) or reduction to a few dimensions (core affect) results in a loss of precision. This will be exemplified by the following example. Assuming that there are real perception differences between groups, the two extreme outcomes of a research endeavor determined to find them are: measuring differences of the size of the real differences or not detecting any differences at all even though they exist. In the former case, the measured differences can only equate to the real differences if the categories or dimensions used in the task aligned perfectly with the real differences. On the other hand, the measured differences would equate to zero, if the true differences lay outside the range covered by these categories or dimensions. Directly looking for continuous differences in perception across a large number of dimensions greatly increases the chance that the true differences can be detected. This applies to the representation chosen for the participant's emotion response as well as the representation for the emotion stimulus. Facial expressions that show the six basic emotions or expressions of varying valence and arousal, might not even contain the features which evoke a different emotional perception for the sexes, in particular if these expressions are presented as static stimuli. Similarly, groups of participants might perceive certain stimuli differently, but might not be able to indicate this, if their choice in response options is too limited. Hence, the search for group differences in perception should preferentially be carried out in a multivariate fashion over continuous dimensions.

4.2.3 Finding suitable representations

Picking an optimal emotion framework for research on emotion perception is a two-fold challenge, since a representation system has to be selected for internal emotion representations of facial expression observers and for the presented facial expression stimuli, which are both highly complex phenomena. As we have discussed, the available frameworks are debatable in

terms of validity and might be too restrictive to allow for a precise investigation of emotion perception and differences thereof.

As abstractions of real-world phenomena, emotion frameworks do not capture the phenomenon in its entirety and therefore might be well-suited to answer some questions, while being insufficient to answer others (i.e. "all models are wrong but some are useful" (Box & Pelham, 1979)). However, if the perception differences one is looking for are unknown, the suitable representation also cannot be known a priori. Simply investigating a large number of categories or dimensions individually for such differences is not only infeasible in experiments with participants but inevitably leads to the statistical issue of multiple testing. Here, we put forward an approach that seeks to solve this problem. We record high-dimensional data on the emotion response with seven continuous rating dimensions as well as on the facial expression side with face tracking data over time and present a method that by design finds the optimal representation for each side to relate it to the respective other side as explained below. Thereby, the method covers a multivariate and continuous search space for perception differences and is largely independent from the prevailing emotion frameworks. Additionally, the technique yields highly interpretable results that can be tested for significance and allows straight-forward visualization.

4.3 Introduction to Partial Least Squares Analysis

Partial least squares (PLS) analysis is a latent variable modeling approach most commonly used in the field of neuroimaging (McIntosh, Bookstein, Haxby, & Grady, 1996; McIntosh & Lobaugh, 2004). Formally a PLS analysis of two standardized data sets, X and Y, is a Singular Value Decomposition (SVD) of their product $R = X \cdot Y$ and is very similar to a Canonical Correlation analysis (see Krishnan, Williams, McIntosh, & Abdi, 2011 for details of the method). Importantly, whereas most traditional GLM-type analyses (e.g. multiple regression) allow only a single dependent variable and imply a specific direction of prediction, PLS provides a way to investigate the correlation between two high-dimensional sets of related data. Similar to Principal Component Analysis or factor analysis, PLS reduces the dimensionality of the data by forming new variables from the combination of the correlated original variables.

PLS computes pairs of data-driven latent variables (LVs), defined so that the correlation between elements of the pair are as highly correlated as possible. The LVs extracted from a given data set are orthogonal and are ordered according to the amount of explained covariance, with the first LV explaining the most covariance and the last LV the least. Thus, pairs of LVs

link variation across data sets, and are ordered with the most explanatory LVs first. As a result, by examining only the first few LVs, information irrelevant to the relationship between the two sets of data can be disregarded. For example, human faces vary in their shape and appearance. Much of that variation is related to a person's age, sex, ethnicity and other idiosyncratic characteristics. Although it has been shown that some of these features can influence emotion recognition processes, such as age-specific features (Fölster, Hess, Werheid, et al., 2014; Freudenberg et al., 2015), they may carry little additional information about the emotional state displayed on a face. Hence, only the amount of variation in faces which has an influence on emotional perception processes should be investigated. By design, PLS focuses only on this relevant part of the variance in both data sets.

4.3.1 Statistical inference with PLS

Other methods exist to process high-dimensional data, especially in the field of machine learning. One noteworthy example is *deep learning*, which can also be used to relate pairs of high-dimensional data to each other and which can automatically separate relevant from irrelevant variation in the data. While deep learning can approximate arbitrarily complex non-linear relationships, it has the downsides of a demand for large quantities of data and, more severe, a lack of interpretability that render its use in psychological research problematic. In contrast, models used in the field of psychology are usually employed to find clearly defined relationships in a data sample of limited size that can be generalized to the population from which the sample was drawn by means of statistical testing, and which can be easily interpreted in the light of existing theory. The PLS framework provides a compromise between the transparency and interpretability of multiple regression and the multivariate power of more intricate approaches. It both permits sensible statistical hypothesis testing and provides an easy mapping back to the natural scales of the measures, which in turn eases interpretation.

4.3.2 PLS as a tool for research on emotional facial expression perception

One particular point of utility for PLS in the study of emotion is the ability to directly model the structure of dynamic facial expressions, and its relationship to emotional perception. Facial expressions can be quantified through face tracking, which produces a set of landmark coordinates over time. Continuous ratings on emotion dimensions will be used to capture the emotion perception of the participants. Here, it should be noted that the specific emotion dimensions are not important as long as they sufficiently cover the space of emotion. This means that whatever one seeks to find only has to be present (in mathematical terms: as a linear combination) in the space spanned by the original dimensions. PLS will then by design identify

these linear combinations as LVs that link variation across the rating and face tracking data. The method is therefore not tied to the structure of emotion proposed by the common emotion frameworks discussed earlier and instead finds the structure most relevant to the data under investigation. Another advantage of the method is that the effect of each LV can be mapped back to the emotion ratings and the facial expressions. In particular, the mapping to facial expressions provides an intuitive visualization. For example, facial expressions for which the emotional perception differs between men and women can be directly generated, which enables a qualitative interpretation.

Our research question was whether there are quantitative or qualitative differences (or potentially a mixture of both) in the emotion perception from facial expressions between men and women. The literature shows evidence for higher accuracy and sensitivity in emotion recognition tasks for women as compared to men (Hoffmann, Kessler, et al., 2010; Montagne et al., 2005; A. E. Thompson & Voyer, 2014) and higher variability in ratings (Hall & Matsumoto, 2004) indicating quantitative differences in perception potentially due to a greater female sensitivity to the variability in facial expressions. On the contrary, qualitative differences in perception seem to be, to the best of our knowledge, under-researched. Based on this, we formulated the following hypotheses:

Hypothesis 1: We expect quantitative differences between the sexes and therefore expect women to systematically perceive more variation from facial expressions than men do.

Hypothesis 2: Additionally, we expect qualitative differences between the way that men and women perceive emotions from facial expressions and expect that at least part of the variation in facial expressions and in emotion ratings is linked in a systematically different way for men and women.

4.4 Methods

4.4.1 Participants

The data of the 441 participants (129 males) from Project 1, who remained after several filtering steps, was used again in this study. A full description of the sample can be found in Section 2.8.1.

4.4.2 Materials

4.4.2.1 Video data sets

The same data set of 480 video clips as described in Project 1 (Section 2.8.2) was used in this study. The data set featured 12 actors (6 male, 6 female, age range: 21-64), which showed expressions from 40 emotion categories including Ekman's six basic emotions and 34 complex emotions.

4.4.2.2 Face tracking data

Tracking data for all videos was acquired using the software OpenFace (Baltrusaitis, 2018). OpenFace provides the x- and y-coordinates of 68 landmarks that are placed on the face for each frame of a facial expression video. All expressions were then normalized to a common frame of reference using Generalized Procrustes Analysis to remove differences due to the location on the frame or the overall size of the face in the video. The final face tracking data set contained 13600 data points for 306 videos, as all videos with less than three female or male raters were removed to ensure that meaningful averages could be calculated in the procedures described in Section 4.4.4.1 and 4.4.4.2. The number of 13600 per video originates, because each video had a length of 4 seconds and used a framerate of 25 frames per second. Therefore, each video produces tracking data for 100 frames, each of which contains 68 X- and 68 Y-coordinates. Frames were concatenated clip-wise, which enables the possibility of finding correlations across time within facial expressions, therefore allowing the analysis of dynamic facial expressions instead of static ones.

4.4.3 Procedure

Testing was carried out in German on the soscisurvey.de platform (Leiner, 2014). Each participant rated 12 videos randomly chosen from the pool of 480 videos. First, participants rated valence and arousal on scales anchored by pictures of the respective Self-Assessment-Manikin (Bradley & Lang, 1994). Second, participants provided ratings on the Basic Emotions and Interest (BEI) scales: happiness, sadness, fear, anger, surprise, disgust, interest, and then rated the subjective difficulty of making the BEI ratings. Ratings were given on continuous scales ranging from "gar nicht X" ("not X at all") to "sehr X" ("very X") with X being one of the rating dimensions. All rating responses were encoded between 1 and 101.

4.4.4 Data analysis

A PLS model takes the rating data and the face tracking data of the video on which the rating data was collected as inputs. It results in pairs of LVs, which describe how some of the rating dimensions correspond to some of the movements of the tracking landmarks from the facial expression data. This can be imagined akin to PCA or factor analysis, where certain original

variables are said to load on the principal components or underlying factors. Similarly, in PLS the original variables from the two input data sets load onto the LVs. Because LVs come in pairs in PLS (with each pair containing one LV for the first and one for the second data set) it is possible to see which variables are related within and across data sets. The variables related within one data set load onto the same LV and the variables related across data sets load onto the same pair of LVs. For example, it could be that the tracking landmarks for the right and left side of the corners of the mouth load onto one LV for the tracking data and that the happy rating dimension loads onto the one LV in the rating data that is paired with that first LV. In that way, this pair of LVs would reveal that an upwards movement of the corners of the mouth is related to happiness ratings. Therefore, PLS can capture the perceptual link between facial expression and emotional impression evoked in the observer.

Three PLS models were computed. One within-group model for male and female participants each and one between-group model to explicitly examine the differences between men and women. Quantitative differences as predicted by Hypothesis 1 could be observed in the within-group models in a difference in the number of significant pairs of LVs and in the cumulative variance that they explain. Qualitative differences as predicted by Hypothesis 2 would be observed in differing loading patterns onto the significant LVs of same rank in the within-group models as well as significant LVs in the difference model.

4.4.4.1 Within-group analysis

To account for interpersonal rating style differences and dependencies in the data, emotion ratings for each participant were centered by subtracting the mean rating across all of the participant's ratings. For the analyses within groups data was then sorted into two matrices X and Y for each group so that each row in X contained the centered seven BEI ratings and Y contained the face tracking data for the video clip that these ratings were given to. This resulted in two matrices $X_{female,}$ with dimensions 2253×7 and Y_{female} with dimensions 2253×13600 for the female group and matrices $X_{male,}$ with dimensions 1264×7 and Y_{male} with dimensions 1264×13600 for the male group. Two PLS models were then computed; one using X_{male} and Y_{male} and one using X_{female} and Y_{female}.

4.4.4.2 Between-group analysis

For the analysis between groups, group mean ratings were calculated for each video clip. Subsequently, the mean ratings of the female group were subtracted clip-wise from the male mean ratings resulting in mean rating differences. Similar to the procedure described for the within-group analyses, these mean rating differences were then arranged in a matrix Xdiff

(306×7), which was paired with a matrix Ydiff (306×13600), containing the respective face-tracking data. Xdiff and Ydiff were then used to compute a PLS model.

4.4.4.3 P-value calculation with permutation testing

P-values for the resulting LVs of the two within-group models and the difference model were obtained using permutation testing. Conventional parametric methods cannot be used to calculate the p-values of LVs, since the underlying asymptotic distribution of the LVs under the null hypothesis is unknown. Permutation testing, on the other hand, is a non-parametric method that constructs a null distribution from the data itself. In this case, we generate a distribution of the covariances of the pairs of LV under the null hypothesis that there is no relationship between the X and Y data sets—that is, that the ratings attributed to a given clip are independent of the facial movements displayed in that clip.

Permutation testing is an iterative process. In each iteration the order of the rows of the emotion ratings in X is shuffled, therefore breaking the association between rows in X and Y. Then, the PLS model is calculated on this *surrogate* data set, and its resulting LVs are rotated to align with the initial PLS solution following (McIntosh & Lobaugh, 2004). This is crucial as permuting the original data can result in PLS solutions where latent variables occur in a different order or with a different sign than the ones of the original solution or even point in different directions. After rotation the singular values of the LVs are recorded and the next iteration is carried out. The resulting distribution is the distribution of LVs when the expressions made do not correspond to the ratings on the same data row, and is therefore an empirical approximation of the distribution of the LVs under the null hypothesis that there is no relationship between the two datasets X and Y. The (one-tailed) p-value of an LV can then be estimated by simply calculating the percentage of elements in the null distribution greater in value than the LV observed in the true data set.

All analyses in this paper were conducted using R and the package "plsr" developed by Schneider and Brick (2019), which provides the needed functionality in an easy to use interface (see Section 4.7.1). The number of permutation iterations was set to 10000 for all analyses.

4.5 Results

4.5.1 Descriptive statistics

Figure 8 shows the Pearson correlation of the rating dimensions for the female group in the upper triangle of the matrix and the rating correlations for the male group in the lower triangle. The diagonal shows the standard deviations for each rating dimension for both groups. All

correlations have the same directionality and similar correlation patterns are visible for men and women. Negative emotion dimensions (sad, fearful, disgusted and angry) exhibit a positive correlation with one another, meaning they were frequently used together in rating the stimuli. In a similar fashion, the positive dimensions (surprised, happy, interested) show positive correlations.

On the other hand, happy and interested are negatively correlated with the negative dimensions. Although, the overall pattern seems to be similar in both groups, some differences in the strength of correlations are apparent, especially among the correlations between the negative emotion dimensions. For example, the correlation between fearful and disgusted (women: r=.22, men: r=.34) and the correlation between angry and sad (women: r=.23, men: r=.32) are stronger in the male group than in the female group in a descriptive sense, indicating that men in our sample used these rating dimensions more frequently in a similar way (both low or both high) than women did. On the other hand, interested and angry exhibit a stronger negative correlation in the female group (r=-.31) than in the male group (r=-.2), meaning women tended to use these rating dimensions in an opposing way (one high when the other is low) more often.

The standard deviation of the rating dimensions (Figure 8, diagonal) is similar between rating dimensions with values ranging from 27 to 32 rating points. The standard deviation of the rating dimensions in the female group is increased by one point (interested, happy, surprised, sad) or equal (angry, disgusted) to the male group. Only the fearful dimension shows a one point increase in the standard deviation for the male group compared to the female group.

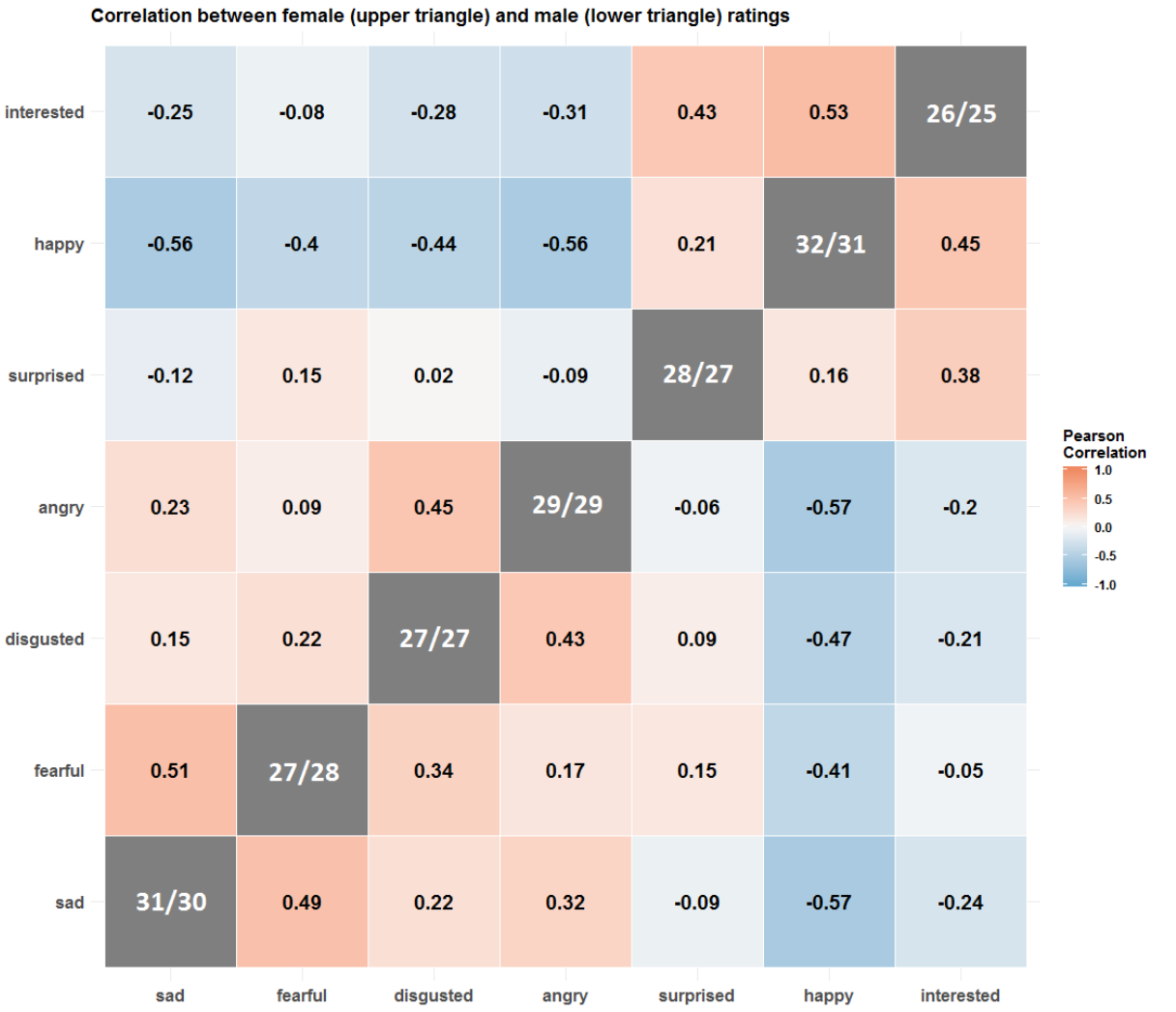


Figure 8. Correlation matrix for ratings of women (upper triangle) and men (lower triangle). The standard deviation of the rating dimensions is shown on the diagonal for both groups (women left, men right). Positive correlations (red) can be seen among negative emotion categories (angry, disgusted, fearful, sad) and between positive emotion categories (interested, happy, surprised), whereas negative correlations (blue) appear between interested and happy and the negative emotions in both women and men.

4.5.2 Number of significant latent variables

Figure 9 shows the p-values for the LVs of the within-group models for men and women and for the difference model, which were computed with permutation testing. The number of significant LVs of the within-group models pertain to Hypothesis 1, as a difference in significance pattern could hint at a difference in sensitivity towards emotional facial expressions. Significant LVs of the difference model, on the other hand, pertain to Hypothesis 2, as they indicate qualitative differences between the groups.

The model for male participants has three LVs below the significance threshold (red horizontal line) of $\propto = 0.05$ (p_1, p_2, $p_3 < 0.001$), whereas the model for female participants has six LVs below the significance threshold (p_1 ,p_2 , p_3, p_4 , $p_5 < 0.001$, $p_6 = 0.02$). The difference model exhibited one significant LV (p_1 = 0.004). The null distributions of corresponding singular values generated through permutation of the data as described in the Methods section (Section 4.4.4.3) can be seen in Supplementary Figure 1, Supplementary Figure 2 and Supplementary Figure 3. Here, the location of the actual value within the distribution is indicated by a vertical red line.

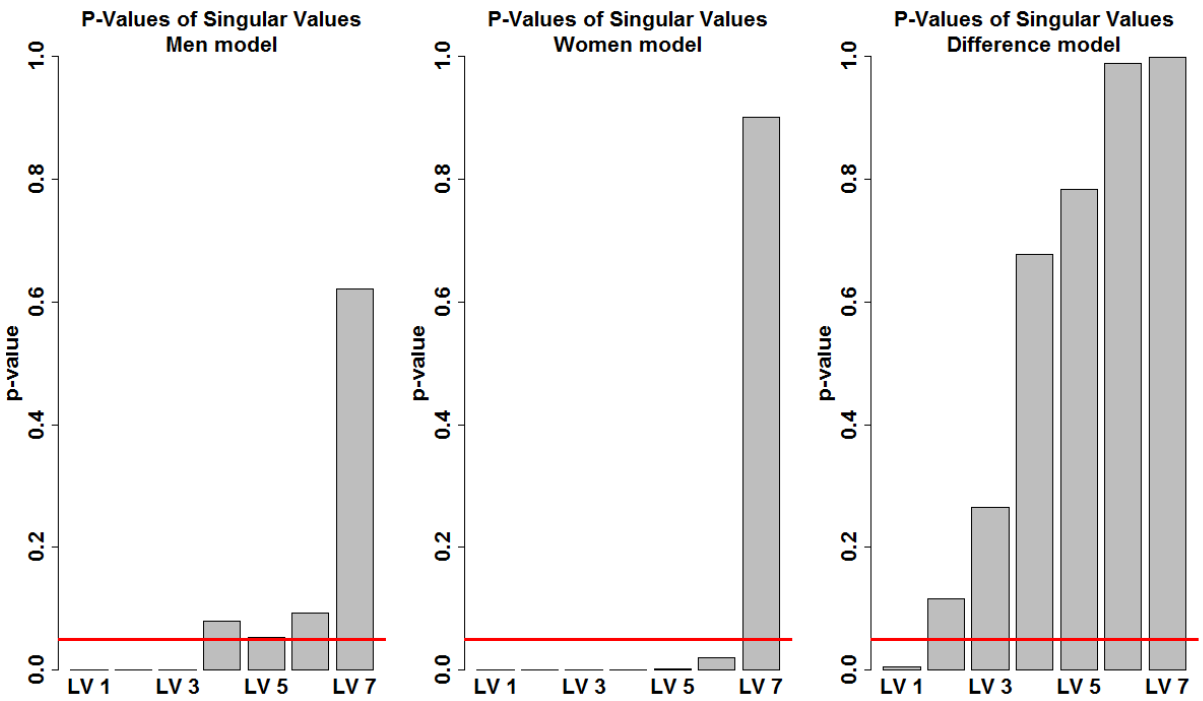


Figure 9. P-Values of the singular values of the latent variables (LV) of the men (left), women (middle) and difference model (right) determined through permutation testing with 10000 permutations.The significance threshold of 0.05 is depicted by a red line. The first three latent variables are significant in the model for men, whereas the first six latent variables are significant in the model for women and only the first latent variable is significant in the difference model.

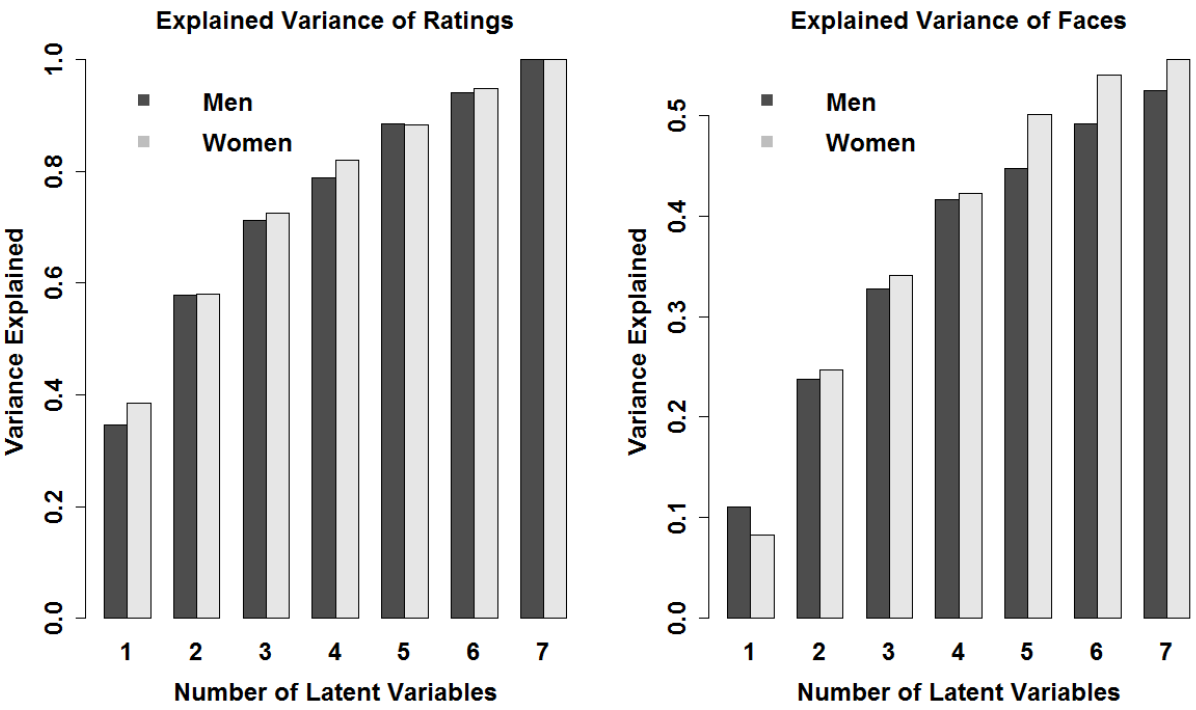


Figure 10. Explained variance of rating (left) and facial expression datasets by number of latent variables for the men (dark grey) and women (light grey) model.

4.5.3 Explained variance

4.5.3.1 Male and female model

The amount of variance of the original rating and facial expression data that can be explained by a certain number of LVs is depicted in Figure 10 for the male and female model. These results therefore also relate to Hypothesis 1. The explained variance in the rating and facial expression data increases with the number of LVs in both models. The proportion of variance explained in the rating data is similar across the number of LVs for the male and female model, with the female model explaining between 0.2% to 3.8% more variance in all cases except with 5 LVs where the male model explains 0.27% more variance.

The amount of variance explained in the facial expression data is also largely similar across models and number of latent variables for the facial expression data. However, there is an additional increase in explained variance in the female model for LVs five to seven that goes beyond of what would be expected from the progression of variance across the previous LVs. In total, the female model consistently explains between 0.6% to 5.4% more variance when using two to seven LVs, whereas the male model explains 2.7% more variance when only one

LV is used. Therefore, across number of LVs and data types, the female model explains slightly more variance than the male one, with only one exception.

With all seven LVs 100% of the variance of the rating data and 52 - 56% of the facial expression data could be explained by the male and female models respectively. This results from the higher dimensionality of the face data (13600 dimensions) compared to the rating data (7 dimensions). With its three significant LVs the male model explains 71% of the variance in the rating data and 33% of the variance in the face data. Accordingly, the female model explains 95% of the variance in the rating data and 54% of the variance in the facial expression data with its six significant components.

4.5.3.2 Difference model

The significant LV of the difference model explains 20% of the variance of the differences in ratings between the groups, corresponding to 16% of the variance in the facial expression data. These values correspond to r = .45 and r = .4 and are therefore medium effect sizes according to Cohen (Cohen, 1992). These results give an estimate of the hypothesized qualitative emotion perception differences between men and women, as postulated by Hypothesis 2.

4.5.4 Loading patterns of within-group models

The following results describe how the loading patterns of the within-group models and therefore how emotion perception space is laid out for men and women. These results therefore describe qualitative similarities and differences and are therefore relevant for Hypothesis 2.

4.5.4.1 Rating dimensions

Table 8 displays the loadings of the original rating dimensions onto the LVs for the female and male model. These can be read as correlations of the original variables, here the rating dimensions, with the newly established latent variables. Similar to correlations, they range between -1.0 and +1.0. However, since the direction of an LV is arbitrary, the sign of a loading is only meaningful in comparison to the sign of the other loadings on the LV. If groups of loadings of opposite sign appear on the same LV, then this LV is said to separate between these loadings.

Differing loading patterns can be observed for the rating dimensions in the male and female model. Individual loadings on the same LV cannot be compared statistically between models unless the latent spaces spanned by the rotation are aligned. However, qualitative comparison of LVs as a whole is meaningful in relation with the order of LVs.

LV1 separates between the rating dimensions of happy, surprised and interested on one hand and sad, disgusted and angry on the other hand in both models of mostly comparable size. However, the surprised loading is 1.7 times higher in the male group (0.25 compared to 0.41).

LV 2 shows loadings of sad, surprised, fearful and interested with the same sign in the female group. In the male group LV2 separates between happy on one side and surprised, disgusted, angry, fearful and interested on the other side. Loadings of surprised (0.74 and -0.71) and interested (0.24 and -0.22) are of comparable size, whereas the other loadings differ between the groups.

LV3 separates sad and fearful from surprised, disgusted and angry in the female models and from disgusted and angry in the male model. The sad, disgusted and angry loadings are of comparable size for both models (-0.52 and 0.54; 0.26 and -0.30; 0.68 and -0.62), whereas the fearful rating differs (-0.24 and 0.43) between the groups.

LV4, LV5 and LV6 are only significant in the female group and therefore will only be described for the female model. Here, LV4 separates sad, angry and interested from disgusted. LV5 separates surprised from all other dimensions. LV6 separates sad, disgusted and interested from fearful.

Table 8. Loadings of the PLS models for female and male group on the rating data side. Loadings are boldfaced if they are greater than 0.2 in absolute value and belong to a significant latent variable (LV).

	Women							**Men**							**Difference**
	LV1	LV2	LV3	LV4	LV5	LV6	LV7	LV1	LV2	LV3	LV4	LV5	LV6	LV7	LV1
happy	**0.64**	-0.17	0.02	-0.05	**0.59**	-0.08	-0.46	**0.61**	**0.26**	-0.12	0.54	-0.02	0.47	0.20	**-0.27**
sad	**-0.34**	**0.27**	**-0.52**	**0.22**	**0.28**	**0.54**	-0.34	**-0.30**	-0.16	**0.54**	0.01	-0.24	0.69	-0.23	**0.36**
surprised	**0.25**	**0.74**	**0.33**	-0.08	**-0.36**	0.09	-0.37	**0.41**	**-0.71**	-0.09	-0.42	0.24	0.23	0.20	**0.72**
disgusted	**-0.28**	0.11	**0.26**	**-0.80**	**0.39**	**0.22**	0.12	**-0.21**	**-0.33**	**-0.30**	0.49	0.50	0.06	-0.52	-0.05
angry	**-0.44**	-0.07	**0.68**	**0.45**	**0.29**	-0.07	-0.23	**-0.43**	**-0.20**	**-0.62**	0.10	-0.44	0.21	0.37	0.07
fearful	-0.13	**0.53**	**-0.24**	0.07	**0.37**	**-0.68**	0.20	-0.02	**-0.46**	**0.43**	0.54	-0.15	-0.39	0.37	**0.48**
interested	**0.36**	**0.24**	0.19	**0.32**	**0.28**	**0.42**	0.65	**0.38**	**-0.22**	-0.13	0.01	-0.65	-0.20	-0.57	**0.22**

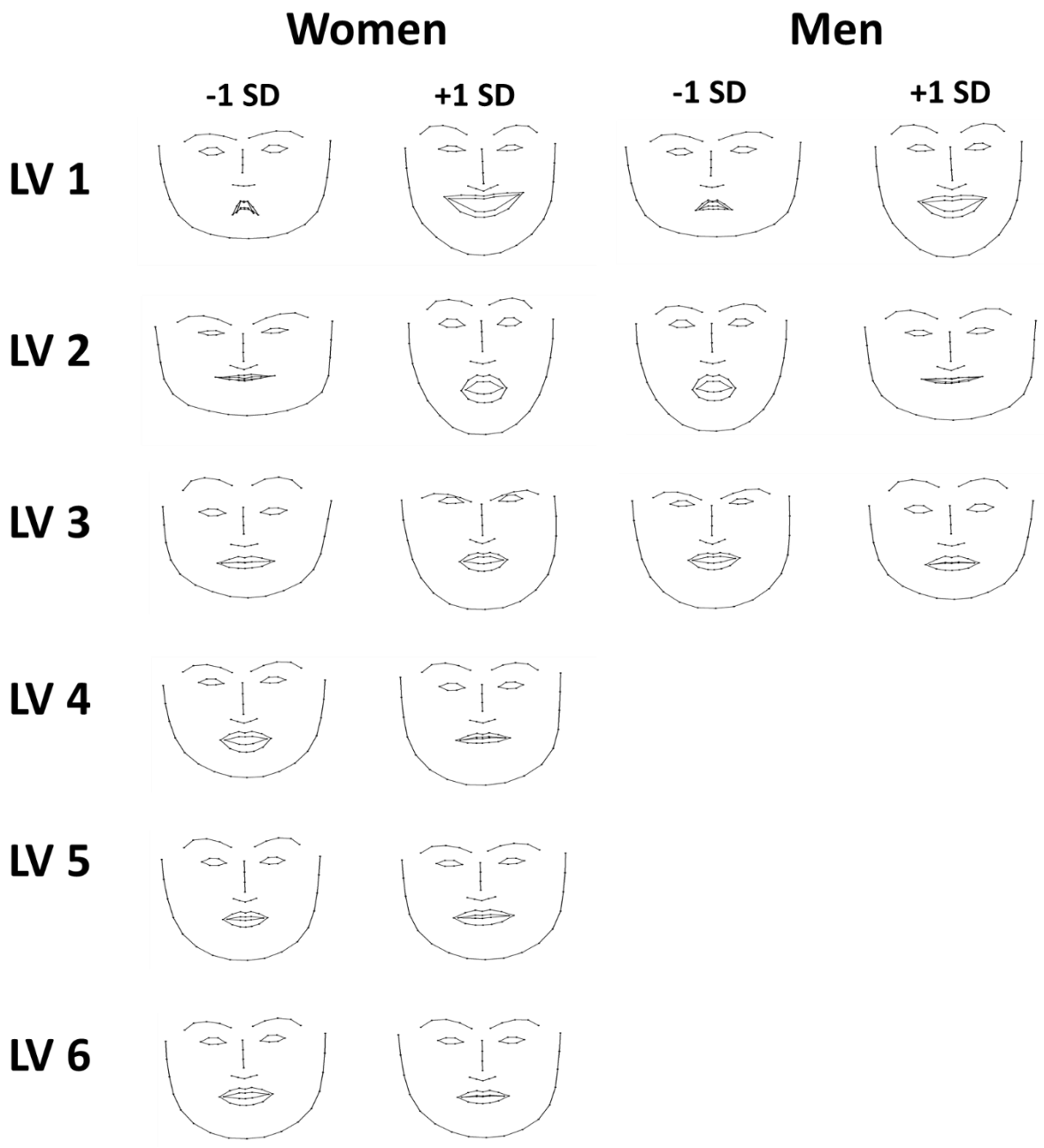


Figure 11. Effect of the significant latent variables (LV) on the face side. Each row shows the effect of a specific LV on the facial expression side of the PLS model for the female and male group. The facial expressions are generated by setting each LV to -1 SD and +1 SD for women and men. Shown is only the 50th frame of the sequence of frames that are generated by the model.

4.5.4.2 Facial expression dimensions

Figure 11 displays the effect of LVs on the facial expression side for the female and male model by showing the result of setting the respective LV to +1 SD or -1 SD. Here it can be seen that

effects are largely consistent between the models as similar faces appear. It should be noted that LV2 and LV3 have similar loadings but with opposite sign in the two groups, which results in similarity between the plots for +1 SD and -1 SD (second and third row) and vice versa. Although faces appear to have similar emotional content, it can be observed that the faces generated by the female model are of greater intensity. This is apparent, for example, in a wider smile for the facial expressions generated by a positive LV1 (first row, column two) or more strongly lowered eyebrows generated by a positive LV3 (third row, column two).

4.5.5 Loading patterns of between-group model

The loading patterns of the between-group model will be described in the following. These results are essential for Hypothesis 2 as they detail the emotion perception differences between men and women.

LV1 in the difference model mainly separates differences in happy from differences in sad, surprised, fearful and interested ratings (Table 8). Figure 12 shows the effect of the significant LV onto the rating and facial expression data simultaneously when the LV is set to -3 and +3 SD. The figure shows, that setting this LV to -3 SDs produces an elongated and narrow face with widened eyes, raised eyebrows, slightly lowered corners of the mouth and a slight head-tilt backwards (Figure 12, top left) compared to the neutral expression (Figure 12, top middle). This expression is associated with higher ratings from men on the happy (11 rating points) and disgusted (7 rating points) dimensions and higher ratings on the surprised (31 rating points), fearful (18 rating points), sad (16 rating points), interested (9 rating points) and angry (1 rating point) dimension from women.

Setting the LV to +3 SD produces a short, wide face with narrow eyes, lowered eyebrows and a slight grin (Figure 12, top right). This expression is associated with higher ratings of happiness (8 rating points) from women and higher ratings of surprised (37 rating points), fearful (25 rating points), sad (19 rating points), interested (13 rating points), angry (6 rating points) and disgusted (2 rating points) from men.

Figure 12. Visualization of the first difference LV. The effect of the difference LV on the facial expression side (top) and rating side (bottom) is visualized for -3 SD and + 3 SD.The bar plots show the differences in ratings to the facial expressions displayed on top of them. Bars in the positive region correspond to higher ratings of men, whereas bars in the negative region correspond to higher ratings of women.

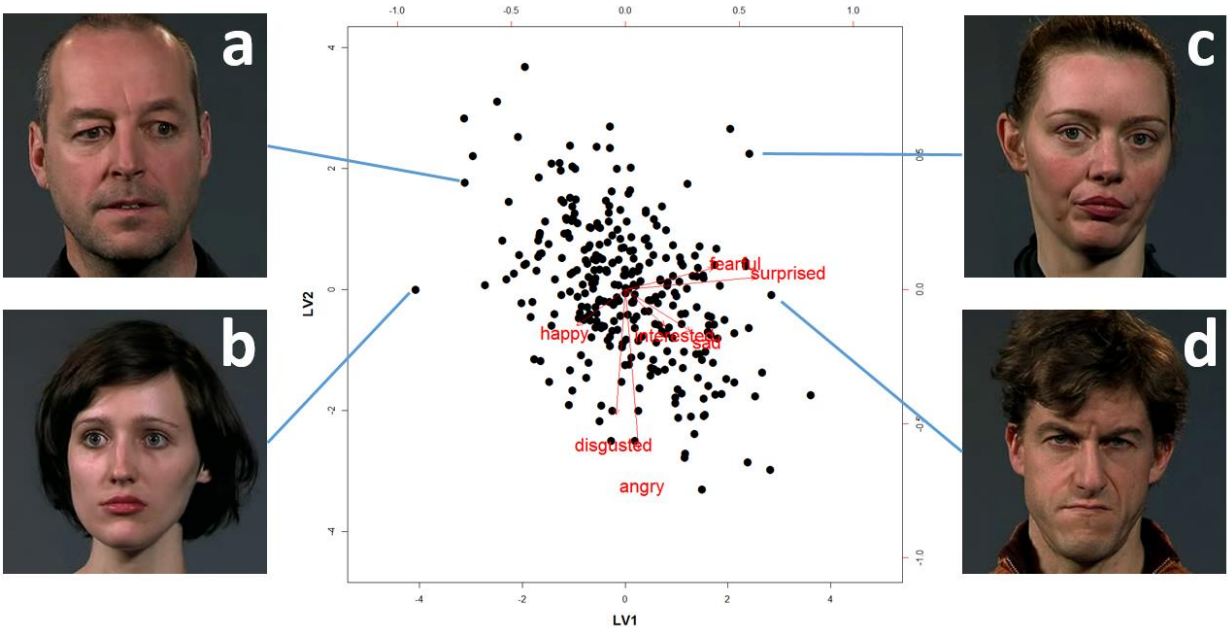


Figure 13. Biplot for the difference PLS model. The original rating differences (black) and rating dimensions (red arrows) are plotted in LV space.Still frames from the videos corresponding to the rating differences are shown for two strong loadings on the negative (a, b) and positive (c, d) side of LV1.

Figure 13 shows the rating differences (black) projected into the latent variable space of the difference model. Red arrows indicate how the differences in the original rating dimensions align with LV1 and LV2. Arrows point into the direction of greater male ratings. The smaller the angle between a rating dimension and an axis, the more this dimension loads onto the respective LV. The arrows show that the rating differences of fearful and surprised are almost parallel to LV1, meaning they are captured almost exclusively by it, whereas the differences in interested, sad and happy ratings stand at an angle of roughly 45° to LV1 and LV2, meaning both LV1 and LV2 capture part of their variance. The differences in disgusted and angry ratings, however, are almost parallel to LV2 and therefore their variance is captured almost exclusively by LV2. The figure also shows four exemplary still frames from the videos on which differences in ratings occurred. Two of the depicted video examples (Figure 13: b and d) have rating differences that load strongly on LV1 and not at all on LV2. Therefore, the expressions displayed in these videos correspond strongly to the expressions shown in the top row of Figure 12, and correspond to rating differences between the groups similar to the ones shown in Figure 12 in the bottom row. Videos a and c, however, have a considerable influence of LV2 and therefore deviate from this pattern. Interestingly, although all four videos show expressions that

differ from a neutral expression, only very subtle and localized facial movement is visible in the clips, resulting in the appearance of a stiff expression. Additionally, all videos show movement other than facial expressive movements in the form of swallowing (b), head shaking (d) or head tilting (a, c).

4.6 Discussion

This study aimed to investigate whether the female and male perception of emotional facial expressions differ in a quantitative (Hypothesis 1) or qualitative way (Hypothesis 2). For this purpose, we employed PLS analysis on emotion ratings and face tracking data in two within-group and one between group models.

4.6.1 Hypothesis 1: quantitative differences in emotion perception

Hypothesis 1 predicted greater sensitivity towards the variation in emotional facial expressions for women than for men. The model for female raters had six significant LVs and thus three more significant LVs than the model for male raters, which indicates a greater complexity of female emotion perception. That is, a greater number of dimensions systematically related to variation in ratings and facial expressions could be distinguished statistically for the female group. With the six LVs the model for women explained 95% of the variance in the rating data and 54% of the variance in the facial expressions, whereas the model for men only explained 71% and 33% respectively with its three significant LVs. The difference in number of significant LVs could be due to lower statistical power for the male group resulting from a lower number of male participants. However, the explained variance of the facial expression data per number of LVs (Figure 3, right) shows that even with the same number of LVs for men and women, women consistently explain more variance, which is still consistent with the explanation of greater complexity of the female emotion perception.

It has been previously shown that women exhibit higher variability on emotion rating scales than men, while still being more accurate in their emotion attributions (Hall & Matsumoto, 2004). Also in our study women exhibited greater variance in their ratings on four of the seven rating dimensions. Our research confirms that at least part of this additional variance is significantly tied to variance in facial expressions as indicated by the results on explained variance and significant LVs, which might explain why females are more accurate in their emotion judgments: simply, because they can utilize more of the information present in facial expressions.

These results hint at quantitative differences between the sexes in the direction of increased female sensitivity towards facial expression information as predicted by Hypothesis 1. However, it is still possible that additional qualitative differences exist. For this reason, we discuss the emotion structure laid out by the within-group models as well as the significant LV of the difference model in the following.

4.6.2 Hypothesis 2: qualitative differences in emotion perception

Hypothesis 2 predicted qualitative differences in the perception of emotion from facial expressions between men and women. This should be reflected in a difference in the way that variation in facial expressions and emotion ratings are linked for these groups.

4.6.2.1 Female and male emotion space

Both, male and female models seem to partition the emotion space of ratings along three dimensions with high similarity to the Pleasure-Arousal-Dominance (PAD) model (Mehrabian, 2007) , as will be explained in the following. This dimensional emotion model describes the space of emotion by a valence scale ranging from unpleasant to pleasant, an arousal scale ranging from calm to excited and a dominance scale, ranging from submissive to dominant. According to Mehrabian (2007) this last scale describes "a feeling of control and influence over one's surroundings and others". The scale extends the well-known core-affect model (Russell & Barrett, 1999), although its underpinnings are different, and enables differentiation between elementary emotional states such as anger and fear, which occupy the same region in core-affect space (low valence, high arousal).

The loading patterns (Table 1, Figure 4) of the PLS models show that LV1 corresponds to the pleasure dimension in both groups, because it separates the positive emotion rating dimensions happy, surprised and interested from the negative dimensions sad, disgusted and angry. LV2 seems to correspond to the arousal dimension with high loadings of surprise and fearful in both models. LV3 is anchored by sadness and fear on one side —emotions associated with helplessness— and disgust and anger in the male model, and disgust, anger and surprise in the female model on the other side. Disgust expression has been linked to moral and social judgments (Schnall, Haidt, Clore, & Jordan, 2008), and is closely related to expressions of contempt (Ekman & Friesen, 1986). Being able to judge someone or something implies a state of power and control and might plausibly be considered a "dominant" emotion. It seems therefore possible to interpret LV3 of the male model as similar to the dominance dimension of the PAD model. However, in the female model surprised loads on the same side as disgusted and angry, which violates a clear separation between dominant and submissive, or emotions of

control and helplessness, as surprise clearly indicates a lack of control. But the female model includes LV5, which separates surprised from all other rating dimensions. Therefore, LV3 and LV5 can be used in combination to achieve the separation between dominant and submissive emotions, although at the cost of additional complexity. LV4 of the female model more distinctly separates disgusted from sad, angry and interested. This hints at a greater ability in the female model to distinguish among the negative emotions. Consistent with this explanation is the lower correlation observed in the female group between some of the negative emotions (Figure 1). One straightforward explanation is that women are more adept at distinguishing these emotions perceptually, although these data do not rule out the possibility that men are equally adept, but less motivated and thus less careful, because the emotion perception task might not align with the stereotypical interests of their gender (Meece, Glienke, & Burg, 2006).

Although similarity to the PAD model exists, the female and male PLS models diverge in various ways from it. This could also be due to the fact, that the PAD model was the result of a factor analysis which includes a rotation step that produces high and clearly separated loadings after a number of latent dimensions has been chosen. In the present study, however, no such rotation to simple structure is performed.

Facial expressions corresponding to a given rating set on the female model are of greater intensity than those corresponding to identical ratings on the male model. This reveals a difference in the strength of the association between facial expressions and the ratings that they elicit. For women to give ratings of the same level as men they require more intense facial expressions. Theoretically, this could be due to either of two reasons: i) women are not as sensitive to facial expressions as men are and thus require expressions of higher intensity, or ii) women in general are more cautious in attributing strong ratings and thus rate these expressions as less intense. This latter variant is more consistent with the literature on confidence, response style and on the sensitivity to emotional signals (Montagne et al., 2005) in men and women. Specifically, it has been shown that men exhibit overconfidence across domains (Barber & Odean, 2001; Lundeberg et al., 1994) and suggested that men tend towards more extreme ratings in affective rating tasks (Marshall & Lee, 1998). A differential use of rating scales by men and women thus seems to be the better explanation.

4.6.2.2 Size of differences in emotion perception between women and men

The difference model exhibits one significant LV, which indicates that women and men perceive certain dynamic facial expressions presented to them in a qualitatively different way as was predicted by Hypothesis 2. This difference LV accounts for 20% of the variance of the

differences in ratings between the groups. Our method finds a medium-sized effect, much larger than previous findings of female advantage in visual-only emotion recognition (A. E. Thompson & Voyer, 2014), which constitutes a small effect (d=0.17, or approximate R^2 of .007). This is not surprising, as traditional emotion recognition paradigms investigate accuracy differences on separate emotion categories, i.e. univaritate differences, whereas the method that we presented allows to find differences in emotion perception independent from predefined categories in a multivariate sense.

4.6.2.3 Qualitative characteristics of emotion perception differences

The loading pattern of the difference LV on the rating side closely resembles the one for LV2 in the female model, which could be seen as analogous to the arousal dimension of the PAD model. This might hint at qualitative emotion perception differences between men and women specifically for emotional states situated along the arousal axis in PAD space. In fact, faces generated by the extreme negative range of the difference LV appear to be subtle expressions of fear or distress (Figure 5, top left). In accordance with that, videos which produce rating difference that have a high loading on the difference LV exhibit signs of emotional masking as the actors appear to have stiff and seemingly unmoving faces. Here, women rated these videos as higher in fear and surprise, although the rapid movements traditionally associated with these emotions are not present in these expressions. Such movements might be consciously masked by the person expressing the emotion, if the social situation would make it disadvantageous to show the full extent of the felt emotion, for example, because of imminent threat. One explanation might be that women are more sensitive to social cues, and interpret the stiff or awkward nature of these expressions as an indicator that negative affect is being felt and masked, whereas men accept the display at face value. Men rated the same expression as higher in happiness, which might indicate that they are successfully deceived by the masking.

Expressions generated by the extreme positive range of the difference LV (Figure 5, top right) appear sort of mischievous by combining lowered eyebrows, typically a sign of anger, with raised corners of the mouth, typically indicating happiness. Combining contradictory positive and negative emotion indicators might again be a sign of emotional masking. Men rate these faces mainly to be more surprised, fearful, sad and interested. However, none of these emotions seem to be displayed in the generated facial expressions. Women rate these facial expressions to be happier than men. While signs of happiness are present in the generated expressions, watching the video clips whose rating differences load strongly on the positive side of the LV, rather gives the impression of masked anger. Masking the unpleasant and

potentially socially undesired emotion anger by displaying signs of a positive emotion also seems more likely than vice-versa. In this way, the positive range of the first difference LV would reveal misattributions of both sexes for expressions that mix typical signs of anger and happiness. One has to be careful in over-interpreting these results. While certainly interesting, our findings need to be validated in independent studies.

4.6.3 Limitations and outlook

Although PLS provides greater insight into emotion perception processes than traditional paradigms do, it comes with specific limitation that need to be acknowledged. PLS is a method that uses linear combinations of the original data dimensions to create latent variables. As such, non-linearities in the data can only be modeled by linear approximation, as with most other GLM-based methods. However, the perception of facial expressions might be influenced by non-linear processes. For example, dynamic information such as velocity and acceleration have been shown to carry important information for facial expression classification (Brick et al., 2009; Pollick, Hill, Calder, & Paterson, 2003). To alleviate this potential weakness of the PLS method a non-linear embedding of the data could be created either manually for selected features of the data or completely automated by machine-learning algorithms such as autoencoders.

Some videos in our data set show expressions with unilateral movement. Facial expressions generated from the PLS models still show signs of unilateral facial movement and asymmetry, however, it cannot be ruled out that some of these movements averaged out due to model computation, when instead they should have added up. A potential solution could be to only analyze one half of the face and to project all movements of the other side onto this half before analysis begins.

The present study has shown how shape information of facial expressions, represented by face tracking data, can be used with the PLS method. However, information about emotional states is most certainly also present in the appearance of the face. Autonomic blood responses, as seen in changes of the skin tone, for example while blushing and changes in face illumination due to wrinkles and frowns may also provide emotional information. These features could be integrated into the PLS technique quite easily as work on combination of shape and appearance features have been carried out, for example in so-called Active Appearance Models (Cootes, Edwards, & Taylor, 1998), which have been also used to model faces (Theobald et al., 2009).

In the present study over 300 facial expression videos of a wide array of emotional expressions displayed by female and male actors of different age groups were used. Although

this is a diverse emotion stimulus set, we cannot claim to have captured all or even most of the variation present in facial expressions. Perception differences between populations can only be found if the expressions eliciting these differing perceptions are contained in the analyzed data set. We are convinced that additional difference dimensions will be found with a larger and potentially more diverse data set. A potential candidate would be a dimension that explains differences in disgust and anger perception, which already appears as a second difference LV in the difference model, even though it is not statistically significant. The presented method is designed in a way that allows rapid data collection through online surveys only. Additional research with additional sets of facial videos may help to understand the emotional expressions not captured in this study.

4.7 Dissemination of Methodological Developments

4.7.1 Partial least squares statistic library for R

I wrote a package for the statistical environment R under the supervision of Timothy R. Brick, which provides the functionality to conduct PLS analyses in a straightforward fashion. The package was submitted to the "Comprehensive R Archive Network" (CRAN) (Hornik, 2012) and accepted for publication on the online platform after two rounds of reviews (Schneider & Brick, 2019). It is available online free of charge (https://cran.r-project.org/package=plsr).

The main function of the package computes PLS models and performs significance testing of resulting latent variables via permutation testing as described in Section 4.4.4.3. Furthermore, the package provides functionality for bidirectional prediction from computed PLS models and various ways of visualization of the model and permutation results. Among these, users can find functions to plot the loadings onto the latent variables, the distribution of singular values generated through permutation, the size of the p-values computed from these distributions and a function to automatically create a Shiny app (Winston Chang, 2018) from the computed models, which allows interactive exploration of the relationships between the two input datasets connected through the latent variables.

4.7.2 Interactive website

To call attention to the developed PLS method and its usefulness for emotion research I designed an interactive website (thisemotiondoesnotexist.com) in collaboration with Christian Knauth (Figure 14). The website is built with the JavaScript library React (Facebook, 2013). It allows users to explore a PLS model of emotion ratings and facial expression data similar to the ones used in Project 3. This model, however, contains data from both men and women.

Users can manipulate ratings of basic emotion and interest ratings on sliders (Figure 14a), which will result in instant generation of the respective facial expression represented as face tracking data (Figure 14b). The generated facial expressions can be played with a button click just like a video (Figure 14c). Additionally, emotion labels which were gathered simultaneously with the emotion rating data are displayed on the right side (Figure 14d). Blue and red bars signify the distance of each label in standard deviations to the centroid of the label in the emotion space spanned by the seven rating dimensions. Clicking on a label will set the rating dimensions to the centroid of the respective label and thereby also display the according facial expression. The website contains a link to an explanation page that contains further details on the method and its aim. This page also contains a link to the R package library.

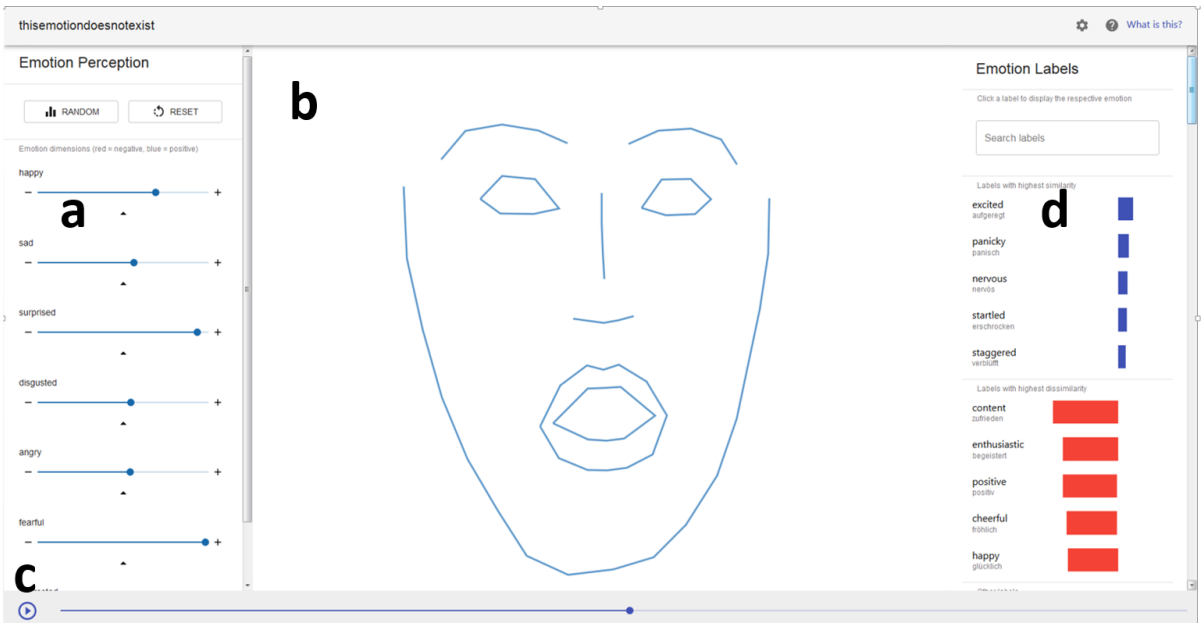


Figure 14. View of the interactive website thisemotiondoesnotexist.com. Users are able to manipulate sliders of emotion dimensions (a), which generates a matching facial expression (b). Each facial expression can be played like a video (c). Additionally, emotion labels (d) with the highest (blue) and lowest (red) similarity to the emotion dimension settings are displayed.

5. Thesis Conclusion

This thesis identified three fundamental problems of contemporary emotion research: the problem of ground truth, the problem of incomplete constructs and the problem of optimal representation. I proposed to find solutions for these problems that rely on data-driven and computer-aided methods. This approach emphasizes detailed measurement of the manifestations of emotions which enables research largely independent of the prevalent theories of emotion and their limitations. Three research projects developed methodological solutions to the three problems and demonstrated their usefulness by applying them to psychological research questions. In the following I discuss the methodological contribution of this work in light of the addition to the substantive literature on emotion generated with them. Furthermore, I evaluate the reusability of the methods and their potential for future research.

5.1 Contribution to the Methods of Emotion Research

5.1.1 Problem of ground truth

The problem of ground truth describes the assumption of a particular structure of emotion as the a priori ground truth even though the validity of that structure is debatable. This is problematic, because commonly used emotion representations are simplistic and restrictive reductions of an unknown emotion space which might lead to biased results or might not even allow to investigate certain research questions. On the other hand, ground truth is handy when comparing abilities between groups, because it provides a way to define correctness.

In Project 1 we have outlined a potential solution to this issue. We have shown how ground truth can be empirically derived as the perceptual population consensus. This circumvents the issue of a fixed a priori definition of ground truth and allows to estimate ground truth for different populations and on an unlimited number of stimulus categories. The described study employed a novel emotion perception paradigm which used dynamic facial expression stimuli from 40 emotion categories. It allowed the observers to rate stimuli on continuous dimensions rather than forcing them to make discrete choices. With this paradigm we investigated self-rated and objective difficulty of emotion perception. To the best of my knowledge, our study represents the first to examine the influence of valence and arousal on the difficulty of facial expression perception. We showed quadratic relationships between valence and arousal on one hand, and both difficulty measures on the other hand. Here, valence had a manifold stronger effect than arousal. The objective difficulty measure in particular showed a strong relationship with squared valence. Such an inference would not have been

possible with common emotion recognition paradigms because those do not allow to quantify difficulty in an objective manner, that is, independent of static assumptions of correctness. Further exploratory analyses gave strong evidence for a higher predictive importance of valence for both difficulty measures in contrast to arousal and all person-specific predictors. This predominant role of valence for the difficulty of emotion recognition falls in line with a functional account of emotions. Consistent with the literature on emotion recognition, older actors were found to be more difficult to judge as indicated by an increase on both difficulty measures when actor age increased. Similarly, the age difference between observer and actor was predictive of the objective difficulty. The study also provided evidence to suggest that certain groups overestimate their emotion recognition capabilities, in particular: young adults, the elderly and men, which is consistent with the development, aging and confidence literature. While most of our analyses reflected known effects supported by the literature, we also found unexpected results. In particular, the increased difficulty on the objective difficulty measure for female actors and observers stood in contrast to reported findings. Thereby, we showed that the proposed method of estimating ground truth is able to generate results that confirm known or suspected relationships but which also add novel results to the substantive literature on emotion. The two difficulty measures seem to provide access to different processes of emotion perception and might be regarded as a valuable contribution to the repertory of emotion researchers. Moreover, the results generated with our paradigm informed the development of an adaptive difficulty algorithm for a social cognition training system targeted at individuals with autism (Moebert et al., 2019). In summary, the presented work highlights the need for novel experimental approaches and paradigms that abandon the concept of accuracy in favor of measures of equal simplicity which allow a more nuanced view onto emotion perception processes.

5.1.2 Problem of incomplete constructs

The problem of incomplete constructs refers to psychological constructs which are vaguely defined in theory and whose practical implications are uncertain. A solution to this problem could be the use of better quantifications of emotion manifestations, which can be tied to the constructs in order to characterize those through these manifestations. To demonstrate this approach, Project 2 investigated how arousal, a long-standing construct of emotion psychology, relates to various measures derived from face tracking data in individuals with and without autism. Arousal is a concept frequently used in emotion research and described to be related to physical activation and alertness (Russell & Barrett, 1999). In the core affect construct (Russell,

2003b; Russell & Barrett, 1999) arousal is used as a dimension to describe emotional states and, among other applications, used to rate emotional facial expressions. However, the perception of arousal from facial expressions seems to be understudied and it is unknown which features of facial expressions have an influence on the arousal perception. In Project 2, we described characteristics of facial expressions, such as the distance to the neutral face, speed and acceleration magnitude, that are likely to be used for arousal estimation by the observer. We described in detail how these can be quantified with measures computed from face tracking data and discussed different methods of aggregating these measures over the course of a facial expression. We found that all measures for distance from a baseline face, and measures for speed and acceleration magnitude respectively, showed moderate to strong correlations. Two of these measures were then used to predict arousal ratings from neurotypical individuals and from individuals with autism, and additionally to predict the differences in arousal ratings between those groups. Distance and speed each have some power to predict arousal, although the effect of speed disappears when controlling for distance in the NT and ASD sample. We found a statistically significant difference in the intercept of arousal ratings between the groups, but found no influence of the selected measures on the difference in arousal ratings. We reproduced these within and between-group findings with an exploratory group of participants with high autistic traits who self-identified as having ASD. The results suggest that arousal perception in individuals with ASD is qualitatively similar to neurotypical individuals, and that differences in ratings may merely be a matter of degree. The predictors distance and speed seem to be specific predictors for arousal to some extent, as they do not predict valence ratings.

As a methodological contribution, the project has described in great detail how to compute simple measures of distance, speed and acceleration magnitude from high-dimensional face tracking data. Furthermore, we showed through correlation analyses that different methods of aggregation for static and dynamic measures respectively result in similar aggregates and can therefore be used interchangeably. We then demonstrated how the measures can be used in psychological research by applying them to investigate arousal perception in two populations. This has added to the substantive literature on arousal by identifying the manifestations that can be specifically tied to the construct and by illuminating differences and similarities in arousal perception between populations. In summary, we showed that deriving simple measures from face tracking data is a valuable and useful method for psychological research that can generate novel results. Moreover, the project is as an example on how to extend incomplete

psychological constructs with modern methods of data collection, which can be easily transferred to other constructs than arousal.

5.1.3 Problem of optimal representation

The problem of optimal representation refers to the search for the best representation of emotion, which has been a central concern of the field of emotion research. However, common emotion recognition paradigms use debatable and overly simplistic emotion representations that do not allow the investigation of emotion perception processes in detail and which are one-fits-all solutions. In contrast, I proposed to find detailed representations that are optimal only in regard to a specific context.

Project 3 introduced PLS analysis as a technique that allows to find such an optimal representation in an automatic and data-driven way and used this technique to investigate emotion perception differences between men and women. Using this method with emotion rating data and face-tracking data of facial expression videos, we showed that women's emotional perception reliably captures more of the variance in facial expressions. To the best of our knowledge, this has not been shown before, because it was not methodologically feasible. These results, for the first time, provide a possible cause for a greater female sensitivity and accuracy in recognizing emotions that has been previously described in the literature (Kessels et al., 2014; Kret & De Gelder, 2012; Montagne et al., 2005). Additionally, we could show that significant differences exist in the way that women and men perceive some facial expressions. Importantly, the method allowed us to visualize concrete facial expression sequences that correspond to the discovered rating differences, and draw qualitative conclusions about their content and meaning. These expressions suggest differing perceptions of masked and ambiguous facial expressions between the sexes. In particular, women seem to perceive apparently masked expressions as fear and surprise more often than men. Thus, our results indicate that quantitative as well as qualitative differences in the perception of emotional facial expressions exist between the sexes.

The PLS method has several advantages. These are: the possibility of directly using high-dimensional data without the prior assumption of a certain emotion framework, enabling statistical inference, high interpretability of results and the possibility for visualization. Our results have demonstrated the usefulness of the PLS method and present it as a tool to investigate differences in emotion perception. Such differences have often been investigated in various clinical groups in comparison to neurotypicals with the common emotion recognition paradigm, for example in individuals with autism (Pelphrey, Morris, Mccarthy, & Labar, 2007),

schizophrenia (Kohler, Walker, Martin, Healey, & Moberg, 2010) or bipolar disorder (Derntl, Seidel, Kryspin-Exner, Hasmann, & Dobmeier, 2009; Martino, Strejilevich, Fassi, Marengo, & Igoa, 2011). Although a general emotion recognition impairment was found for these clinical populations, emotion-specific differences are regularly discussed but seem to be difficult to detect with emotion recognition paradigms. Here, the increased sensitivity of the PLS method could confirm and visualize specific differences, similar to our study on perception differences between men and women. The method could also be used to investigate inter- and intraindividual emotion perception differences by building models for individual subjects and across different points in time. Moreover, this technique could be applied to research the perception of faces in terms of a domain other than emotion, for example, in terms of attractiveness or aggressiveness. Even more general, it could be applied to research the perception of other objects or phenomena, such as, for example, music or art. In short, it is a general technique to investigate perception that can potentially be used with any kind of high-dimensional data. Based on the novel results attained with the method and the numerous potential applications for future research, I deem this method to be not only a potential solution to the problem of representation but also a substantial contribution to the methodology of psychological research in general.

5.1.4 Method reusability

All methods described in this thesis are highly reusable in the sense that they are clearly documented and independent of specific context. All methods were described in detail with the aim to enable their use by researchers without a computer science background. They can be employed to investigate research questions other than the ones which were used to demonstrate their utility. They could potentially also find employment in research or applications outside of emotion research or even psychology.

Project 1 featured a simple, yet effective method to calculate ground truth and a difficulty measure in relation to it. It can be easily transferred to any context in which a ground truth for a perceptual phenomenon is needed. Project 2 described how measures of distance, speed and magnitude of acceleration can be computed from face tracking data and how they can be examined within and between groups. These measures can be calculated on face tracking data, and potentially even other tracking data, from any kind of source and are therefore not restricted to the format of the specific face tracking software we used. As these measures describe important high-level features of dynamical objects they could be relevant for a number of perceptual processes. These measures are easy to compute and can be directly used in the

statistical models common in psychology. Therefore, they offer a balance between simplicity and information, which might render them attractive for other researchers.

The method developed in Project 3 is arguably the most complex of the proposed methods, as it combines techniques from computer science with advanced statistical methods. Researchers might be less inclined to use complex novel methods, because their use might seem laborious and difficult and their applicability to concrete research questions might not be immediately obvious. Hence, we implemented and published an R package for the method (Schneider & Brick, 2019). The package is available online, free of charge, open source and facilitates PLS analyses with permutation testing. Furthermore, I developed a website that calls attention to the method and demonstrates its application interactively.

5.2 Outlook: Quantification – the Way Forward in Psychology?

This thesis has proposed computational methods to solve problems of contemporary psychological research on emotions and demonstrated their usefulness by direct application. The general approach common to all methods was a focus on detailed measurement of concrete manifestations of emotion independent of the prevalent theories of emotion. At this point in time, it is difficult to say precisely what emotions are and what they are not and many related concepts seem vague. However, I am convinced that measurement development and empirical research will pave the way forward so that we may one day know about these things with more certainty. I believe that this approach is not only useful for the field of emotion research but in general for psychological research. Compared to other scientific fields like, for example, physics, the history of empirical research in psychology is short. As such, psychology is a young science that has yet to develop much of the theoretical foundations as well as the empirical practices which were established in other fields over the course of centuries. As exemplified by the history of the measurement of temperature, the evolution of a developing field of science can start out with nothing more than vague notions about a phenomenon and then, through empirical research, give rise to new insights and theories in an iterative process until a somewhat stable foundation of theory and practice is reached. As I have demonstrated for three different problems of current emotion research, computational solutions can be a valuable asset for this endeavor. As a consequence of novel empirical results, researchers might have to accept that some current psychological models are insufficient and need to be replaced by more useful ones, which might be increasingly multivariate, context-specific and data-driven. Then, perhaps one day in the distant future someone in a university seminar will wonder what emotions are

and will find a clear scientific definition in an online encyclopedia—one that is based on precise measurement.

6. References

Adams, N. C., & Jarrold, C. (2012). Inhibition in autism: Children with autism have difficulty inhibiting irrelevant distractors but not prepotent responses. *Journal of Autism and Developmental Disorders*, *42*(6), 1052–1063. https://doi.org/10.1007/s10803-011-1345-3

Adams, R. B., & Kleck, R. E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*. https://doi.org/10.1037/1528-3542.5.1.3

American Psychiatric Association. (2013). 299.00 Autism Spectrum Disorder: DSM-5 Diagnostic Criteria. In *Diagnostic and Statistical Manual of Mental Disorders*.

Anusic, I., Schimmack, U., Pinkus, R. T., & Lockwood, P. (2009). The Nature and Structure of Correlations Among Big Five Ratings: The Halo-Alpha-Beta Model. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/a0017159

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, *338*(6111), 1225–1229. https://doi.org/10.1126/science.1224313

Aviezer, H., Zangvil, S., Messinger, D. S., Mattson, W. I., Gangi, D. N., & Todorov, A. (2015). Thrill of victory or agony of defeat? Perceivers fail to utilize information in facial movements. *Emotion*, *15*(6), 791–797. https://doi.org/10.1037/emo0000073

Baltrusaitis, T. (2018). OpenFace: an open source facial behaviour analysis toolkit. Retrieved from https://github.com/TadasBaltrusaitis/OpenFace

Baltrusaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. https://doi.org/10.1109/fg.2015.7284869

Barber, B. M., & Odean, T. (2001). Boys will be Boys: Gender, Overconfidence, and Common Stock Investment. *The Quarterly Journal of Economics*, *116*(1), 261–292. https://doi.org/10.1162/003355301556400

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient (EQ): An investigation of adults with Asperger Syndrome or High Functioning Autism and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175.

https://doi.org/10.1023/B:JADD.0000022607.19833.00

Baron-Cohen, Simon, Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The ' ' Reading the Mind in the Eyes ' ' Test Revised Version : A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *J. Child Psychol. Psychiat. Association for Child Psychology and Psychiatry*, *42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Baron-Cohen, Simon, Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/A:1005653411471

Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science*. https://doi.org/10.1111/j.1745-6916.2006.00003.x

Barrett, L. F., Bliss-Moreau, E., Quigley, K. S., & Aronson, K. R. (2004). Interoceptive sensitivity and self-reports of emotional experience. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.87.5.684

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The Experience of Emotion. *Annual Review of Psychology*. https://doi.org/10.1146/annurev.psych.58.110405.085709

Bates D, Maechler M, Bolker B, & Walker S. (2015). Package"lme4". *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspectives on Psychological Science*. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Bechara, A. (2000). Emotion, Decision Making and the Orbitofrontal Cortex. *Cerebral Cortex*. https://doi.org/10.1093/cercor/10.3.295

Bertone, A., Mottron, L., Jelenic, P., & Faubert, J. (2003). Motion Perception in Autism: A "Complex" Issue. *Journal of Cognitive Neuroscience*, *15*(2), 218–225. https://doi.org/10.1162/089892903321208150

Blair, R. J. R., Colledge, E., Murray, L., & Mitchell, D. G. V. (2001). A selective impairment

in the processing of sad and fearful expressions in children with psychopathic tendencies. *Journal of Abnormal Child Psychology*, *29*(6), 491–498. https://doi.org/10.1023/A:1012225108281

Blake, R., Turner, L. M., Smoski, M. J., Pozdol, S. L., & Stone, W. L. (2003). Visual recognition of biological motion is impaired in children with autism. *Psychological Science*, *14*(2), 151–157. https://doi.org/10.1111/1467-9280.01434

Boker, S. M., Cohn, J. F., Theobald, B.-J., Matthews, I., Mangini, M., Spies, J. R., … Brick, T. R. (2011). Something in the Way We Move: Motion Dynamics, not Perceived Sex, Influence Head Movements in Conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 631–640.

Bölte, S., Feineis-Matthews, S., Leber, S., Dierks, T., Hubl, D., & Poustka, F. (2002). The development and evaluation of a computer-based program to test and to teach the recognition of facial affect. *International Journal of Circumpolar Health*.

Box, G., & Pelham, rge E. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics*.

Bradley, M., & Lang, P. J. (1994). Measuring Emotion: The Self-Assessment Semantic Differential Manikin and the. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(I), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Brick, T. R., Hunter, M. D., & Cohn, J. F. (2009). Get the FACS fast: Automated FACS face analysis benefits from the addition of velocity. In *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*. https://doi.org/10.1109/ACII.2009.5349600

Brick, T. R., Koffer, R. E., Gerstorf, D., & Ram, N. (2017). Feature Selection Methods for Optimal Design of Studies for Developmental Inquiry. *The Journals of Gerontology: Series B*, *73*(1), 113–123. https://doi.org/10.1093/geronb/gbx008

Brick, T. R., Staples, A. D., & Boker, S. M. (n.d.). Attributed Emotions from Thin Slices of Natural Conversation are Primarily Mixed Emotions.

Britton, J. C., Taylor, S. F., Sudheimer, K. D., & Liberzon, I. (2006). Facial expressions and complex IAPS pictures: Common and differential networks. *NeuroImage*, *31*(2), 906–919. https://doi.org/10.1016/j.neuroimage.2005.12.050

Cabanac, M. (2002). What is emotion? *Behavioural Processes*.

Calvo, M. G., Avero, P., Fernández-Martín, A., & Recio, G. (2016). Recognition thresholds for static and dynamic emotional faces. *Emotion*, *16*(8), 1186–1200. https://doi.org/10.1037/emo0000192

Campos, J. J., Mumme, D., Kermoian, R., & Campos, R. G. (1994). A Functionalist Perspective on the Nature of Emotion. *THE JAPANESE JOURNAL OF RESEARCH ON EMOTIONS*, *2*(1), 1–20. https://doi.org/10.4092/jsre.2.1

Chang, H. (2005). *Inventing Temperature: Measurement and Scientific Progress*. *Inventing Temperature: Measurement and Scientific Progress*. https://doi.org/10.1093/0195171276.001.0001

Christ, S. E., Kester, L. E., Bodner, K. E., & Miles, J. H. (2011). Evidence for Selective Inhibitory Impairment in Individuals With Autism Spectrum Disorder. *Neuropsychology*, *25*(6), 690–701. https://doi.org/10.1037/a0024256

Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., & Zafeiriou, S. (2018). A Comprehensive Performance Evaluation of Deformable Face Tracking "In-the-Wild." *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-017-0999-5

Cohen J. (1988). Statistical Power Analysis for the Behavioural Science (2nd Edition). In *Statistical Power Anaylsis for the Behavioural Science (2nd Edition)*.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

COHN, J. F., & SCHMIDT, K. L. (2004). THE TIMING OF FACIAL MOTION IN POSED AND SPONTANEOUS SMILES. *International Journal of Wavelets, Multiresolution and Information Processing*. https://doi.org/10.1142/S021969130400041X

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998). Active appearance models. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/BFb0054760

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*. https://doi.org/10.1037/emo0000302

Core Team R. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://doi.org/ISBN 3-900051-07-0

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1702247114

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*. https://doi.org/10.1037/h0040957

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*. https://doi.org/10.1145/2682899

Damasio, A. R. (1998). Emotion in the perspective of an integrated nervous system. In *Brain Research Reviews*. https://doi.org/10.1016/S0165-0173(97)00064-7

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. *D. Appleton and Company*. https://doi.org/10.1007/s11664-006-0098-9

Darwin, C. (1871). The descent of man and selection in relation to sex, in Charles Darwin, The origin of species and The descent of man (combined volume). *Journal of Anatomy and Physiology*. https://doi.org/10.1017/CBO9780511703829

Darwin, C. (1872). The Expression of Emotion in Man and Animals. *Animals*, *1227*, 372. https://doi.org/10.5962/bhl.title.4820

Derntl, B., Seidel, E. M., Kryspin-Exner, I., Hasmann, A., & Dobmeier, M. (2009). Facial emotion recognition in patients with bipolar I and bipolar II disorder. *British Journal of Clinical Psychology*. https://doi.org/10.1348/014466509X404845

Detenber, B. H., Simons, R. F., & Bennett, G. G. (1998). Roll 'em!: The effects of picture motion on emotional responses. *Journal of Broadcasting and Electronic Media*, *42*(1), 113–127. https://doi.org/10.1080/08838159809364437

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1322355111

Duchowski, A. T. (2017). *Eye Tracking Methodology*. *Eye Tracking Methodology*. https://doi.org/10.1007/978-3-319-57883-5

Duffy, E. (1957). The psychological significance of the concept of "arousal" or "activation." *Psychological Review*, *64*(5), 265–275. https://doi.org/10.1037/h0048837

Dupré, J. (2002). Is "Natural Kind" a Natural Kind Term? *The Monist*.

Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T., & Convit, A. (2008). Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET). *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-007-0486-x

Eisenbarth, H., & Alpers, G. W. (2011). Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion*, *11*(4), 860–865. https://doi.org/10.1037/a0022758

Ekkekakis, P. (2012). Affect, Mood, and Emotion Choosing a Measure: AThree-Step Process Understanding the Differences Between Affect, Emotion, and Mood. *Measurement in Sport and Exercise Psychology*.

Ekman, P. (1992a). An Argument for Basic Emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. (1992b). An argument for basic emotions An Argument for Basic Emotions. *Cognition and Emotion*, *6*(3–4), 169–200.

Ekman, P. (1992c). Are There Basic Emotions? *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Ekman, P. (2016). What Scientists Who Study Emotion Agree About. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691615596992

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*. https://doi.org/10.1177/1754073911410740

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/h0030377

Ekman, P., & Friesen, W. V. (1986). A new pan-cultural facial expression of emotion. *Motivation and Emotion*. https://doi.org/10.1007/BF00992253

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. *CA: Consulting Psychologists Press. Ellsworth, PC, & Smith, CA (1988). From Appraisal to Emotion: Differences among Unpleasant Feelings.*

Motivation and Emotion. https://doi.org/10.1007/s10751-008-9818-2

Ekman, P., & Heider, K. G. (1988). The universality of a contempt expression: A replication. *Motivation and Emotion*. https://doi.org/10.1007/BF00993116

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*. https://doi.org/10.1126/science.164.3875.86

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*. https://doi.org/10.1037//0033-2909.128.2.203

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2004.02.002

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research*. https://doi.org/10.1108/10662240510590360

Evans, J. R., & Mathur, A. (2018). The value of online surveys: a look back and a look ahead. *Internet Research*. https://doi.org/10.1108/IntR-03-2018-0089

Facebook. (2013). React - A JavaScript library for building user interfaces. Retrieved October 13, 2019, from https://reactjs.org/

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/0096-3445.113.3.464

Fischer, A., & Lafrance, M. (2015). What drives the smile and the tear: Why women are more emotionally expressive than men. *Emotion Review*, *7*(1), 22–29. https://doi.org/10.1177/1754073914544406

Fölster, M., Hess, U., & Werheid, K. (2014). Facial age affects emotional expression decoding. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2014.00030

Fölster, M., Hess, U., Werheid, K., Isaacowitz, D., Komes, J., & Schiller, F. (2014). Facial age affects emotional expression decoding, *5*(February), 1–13. https://doi.org/10.3389/fpsyg.2014.00030

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, *18*(12), 1050–1057.

https://doi.org/10.1111/j.1467-9280.2007.02024.x

Foss-Feig, J. H., Tadin, D., Schauder, K. B., & Cascio, C. J. (2013). A Substantial and Unexpected Enhancement of Motion Perception in Autism. *Journal of Neuroscience*, *33*(19), 8243–8249. https://doi.org/10.1523/JNEUROSCI.1608-12.2013

Frank, M. G., & Stennett, J. (2001). The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology*, *80*(1), 75–85. https://doi.org/10.1037//0022-3514.80.1.75

Freitag, C. M., Konrad, C., Häberlen, M., Kleser, C., von Gontard, A., Reith, W., … Krick, C. (2008). Perception of biological motion in autism spectrum disorders. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2007.12.025

Freudenberg, M., Adams, R. B., Kleck, R. E., & Hess, U. (2015). Through a glass darkly: Facial wrinkles affect our processing of emotion in the elderly. *Frontiers in Psychology*, *6*(OCT). https://doi.org/10.3389/fpsyg.2015.01476

Fried, E. I. (2017). What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review*. https://doi.org/10.1080/17437199.2017.1306718

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricature. *Memory and Cognition*. https://doi.org/10.3758/BF03194377

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X10000300

Gross, J. J., & Feldman Barrett, L. (2011). Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion Review*. https://doi.org/10.1177/1754073910380974

Grühn, D., & Scheibe, S. (2008). Age-related differences in valence and arousal ratings of pictures from the International Affective Picture System (LAPS): Do ratings become more extreme with age? *Behavior Research Methods*. https://doi.org/10.3758/BRM.40.2.512

Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism

spectrum disorder: Insights from eye tracking studies. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2014.03.013

Hacking, I. (1999). Why ask what? In *The Social Construction of What?* https://doi.org/10.2307/3005999

Haeffel, G. J., & Howard, G. S. (2010). Self-report: Psychology's four-letter word. *American Journal of Psychology*.

Hall, J. A., & Matsumoto, D. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion*, *4*(2), 201–206. https://doi.org/10.1037/1528-3542.4.2.201

Hargittai, E., & Shafer, S. (2006). Differences in actual and perceived online skills: The role of gender. *Social Science Quarterly*, *87*(2), 432–448. https://doi.org/10.1111/j.1540-6237.2006.00389.x

Harms, M. B., Martin, A., & Wallace, G. L. (2010). Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychology Review*. https://doi.org/10.1007/s11065-010-9138-6

Hemenover, S. H., & Schimmack, U. (2007). That's disgusting!..., but very amusing: Mixed feelings of amusement and disgust. *Cognition and Emotion*, *21*(5), 1102–1113. https://doi.org/10.1080/02699930601057037

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X0999152X

Hepach, R., Kliemann, D., Grüneisen, S., Heekeren, H. R., & Dziobek, I. (2011). Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency-mplications for social-cognitive tests and training tools. *Frontiers in Psychology*, *2*(OCT). https://doi.org/10.3389/fpsyg.2011.00266

Herba, C. M., Landau, S., Russell, T., Ecker, C., & Phillips, M. L. (2006). The development of emotion-processing in children: Effects of age, emotion, and intensity. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *47*(11), 1098–1106. https://doi.org/10.1111/j.1469-7610.2006.01652.x

Hess, U., Adams, R. B., & Kleck, R. E. (2009). The face is not an empty canvas: How facial

expressions interact with facial appearance. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2009.0165

Hess, U., Blairy, S., & Kleck, R. E. (1997). The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, *21*(4), 241–257. https://doi.org/10.1023/a:1024952730333

Hess, U., & Bourgeois, P. (2010). You smile-I smile: Emotion expression in social interaction. *Biological Psychology*, *84*(3), 514–520. https://doi.org/10.1016/j.biopsycho.2009.11.001

Hlavac, M. (2015). stargazer: Well-Formated Regression and Summary Statistics. *R Package Version 5.2.*, 1–11. Retrieved from http://cran.r-project.org/package=stargazer

Hoffmann, H., Kessler, H., Eppel, T., Rukavina, S., & Traue, H. C. (2010). Expression intensity, gender and facial emotion recognition: Women recognize only subtle facial emotions better than men. *Acta Psychologica*, *135*(3), 278–283. https://doi.org/10.1016/j.actpsy.2010.07.012

Hoffmann, H., Traue, H. C., Bachmayr, F., & Kessler, H. (2010). Perceived realism of dynamic facial expressions of emotion: Optimal durations for the presentation of emotional onsets and offsets. *Cognition & Emotion*, *24*(8), 1369–1376. https://doi.org/10.1080/02699930903417855

Hornik, K. (2012). The Comprehensive R Archive Network. *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.1212

Isaacowitz, D. M., Löckenhoff, C. E., Lane, R. D., Wright, R., Sechrest, L., Riedel, R., & Costa, P. T. (2007). Age differences in recognition of emotion in lexical stimuli and facial expressions. *Psychology and Aging*, *22*(1), 147–159. https://doi.org/10.1037/0882-7974.22.1.147

Isaacowitz, D. M., & Stanley, J. T. (2011). Bringing an Ecological Perspective to the Study of Aging and Recognition of Emotional Facial Expressions: Past, Current, and Future Methods. *Journal of Nonverbal Behavior*. https://doi.org/10.1007/s10919-011-0113-6

Izard, C. E. (1968). Cross-cultural research findings on development in recognition of facial behavior. *Proceedings of the 76th Annual Convention of the American Psychological Association*, *3*, 727.

Izard, C. E. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review*. https://doi.org/10.1177/1754073910374661

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244. https://doi.org/10.1073/pnas.1200155109

Jack, Rachael E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural Confusions Show that Facial Expressions Are Not Universal. *Current Biology*. https://doi.org/10.1016/j.cub.2009.07.051

Jones, D. (2010). A WEIRD View of Human Nature skews Psychologists' Studies. *Science*.

Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry*, *65*(8), 946–954. https://doi.org/10.1001/archpsyc.65.8.946

Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., & Akamatsu, S. (2001). Dynamic properties influence the perception of facial expressions. *Perception*, *30*(7), 875–887. https://doi.org/10.1068/p3131

Kelly, K. J., & Metcalfe, J. (2011). Metacognition of Emotional Face Recognition. *Emotion*, *11*(4), 896–906. https://doi.org/10.1037/a0023746

Keltner, D, & Buswell, B. N. (1996). Evidence for the distinctness of embarrassment, shame, and guilt: A study of recalled antecedents and facial expressions of emotion. *Cognition & Emotion*.

Keltner, Dacher. (1995). Signs of Appeasement: Evidence for the Distinct Displays of Embarrassment, Amusement, and Shame. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.68.3.441

Keltner, Dacher, & Bonanno, G. A. (1997). A study of laughter and dissociation: Distinct correlates of laughter and smiling during bereavement. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/0022-3514.73.4.687

Keltner, Dacher, & Gross, J. J. (1999). Functional accounts of emotions. *Cognition and Emotion*, *13*(5), 467–480. https://doi.org/10.1080/026999399379140

Kessels, R. P. C., Montagne, B., Hendriks, A. W., Perrett, D. I., & de Haan, E. H. F. (2014). Assessment of perception of morphed facial expressions using the Emotion Recognition Task: Normative data from healthy participants aged 8-75. *Journal of Neuropsychology*, *8*(1), 75–93. https://doi.org/10.1111/jnp.12009

Kilts, C. D., Egan, G., Gideon, D. A., Ely, T. D., & Hoffman, J. M. (2003). Dissociable neural pathways are involved in the recognition of emotion in static and dynamic facial expressions. *NeuroImage*, *18*(1), 156–168. https://doi.org/10.1006/nimg.2002.1323

Kim, A. (2016). Wilhelm Maximilian Wundt. In *The Stanford Encyclopedia of Philosophy* (Fall 2016). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/fall2016/entries/wilhelm-wundt/

Kirchner, J. C., Hatri, A., Heekeren, H. R., & Dziobek, I. (2011). Autistic symptomatology, face processing abilities, and eye fixation patterns. *Journal of Autism and Developmental Disorders*, *41*(2), 158–167. https://doi.org/10.1007/s10803-010-1032-9

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*. https://doi.org/10.1007/BF00992553

Kliemann, D., Dziobek, I., Hatri, A., Steimke, R., & Heekeren, H. R. (2010). Atypical Reflexive Gaze Patterns on Emotional Faces in Autism Spectrum Disorders. *Journal of Neuroscience*, *30*(37), 12281–12287. https://doi.org/10.1523/JNEUROSCI.0688-10.2010

Kliemann, Dorit, Rosenblau, G., Bölte, S., Heekeren, H. R., & Dziobek, I. (2013). Face puzzle-two new video-based tasks for measuring explicit and implicit aspects of facial emotion recognition. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00376

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, *59*(9), 809–816. https://doi.org/10.1001/archpsyc.59.9.809

Kohler, C. G., Turner, T. H., Bilker, W. B., Brensinger, C. M., Siegel, S. J., Kanes, S. J., … Gur, R. C. (2003). Facial emotion recognition in schizophrenia: Intensity effects and error pattern. *American Journal of Psychiatry*. https://doi.org/10.1176/appi.ajp.160.10.1768

Kohler, C. G., Walker, J. B., Martin, E. A., Healey, K. M., & Moberg, P. J. (2010). Facial emotion perception in schizophrenia: A meta-analytic review. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/sbn192

Kraut, R., Olson, J., Banaji, M., Bruckman, a, Cohen, J., & Couper, M. (2004). Psychological research online: Opportunities and challenges. *American Psychologist*. https://doi.org/10.1037/0003-066x.59.2.105

Kret, M. E., & De Gelder, B. (2012). A review on sex differences in processing emotional signals. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2011.12.022

Kring, A., & Gordon, A. (1998). Sex Differences in Emotion: Expression, Experience, and Physiology. *J Pers Soc Psychol*. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=paovftc&NEWS=N&AN=00005205-199803000-00010

Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2010.07.034

Krumhuber, E. G., Kappas, a., & Manstead, a. S. R. (2013). Effects of Dynamic Aspects of Facial Expressions: A Review. *Emotion Review*, *5*(March 2016), 41–46. https://doi.org/10.1177/1754073912451349

Kuhn, M. (2008). caret Package. *Journal Of Statistical Software*, *28*(5), 1–26. Retrieved from http://www.jstatsoft.org/v28/i05/paper

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*. https://doi.org/10.1037/a0030811

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2016). lmerTest: Tests in Linear Mixed Effects Models. *R Package Version*, *3.0.0*, https://cran.r-project.org/package=lmerTest. https://doi.org/10.18637/jss.v082.i13

LaFrance, M., Hecht, M. A., & Paluck, E. L. (2003). The Contingent Smile: A Meta-Analysis of Sex Differences in Smiling. *Psychological Bulletin*, *129*(2), 305–334. https://doi.org/10.1037/0033-2909.129.2.305

Lazarus, R. S. (1991). *Emotion & Adaptation*. *Oxford University Press*. https://doi.org/10.1007/s13398-014-0173-7.2

Leighton, S. R. (1982). Aristotle and the Emotions. *Phronesis*. https://doi.org/10.1163/156852882X00104

Leiner, D. (2014). SoSci Survey. *SoSci Survey*.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X11000446

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). *Autism Diagnostic Observation Schedule (ADOS)*. *Journal of Autism and Developmental Disorders* (Vol. 30). https://doi.org/10.1007/BF02211841.

Lozier, L. M., Vanmeter, J. W., & Marsh, A. A. (2014). Impairments in facial affect recognition associated with autism spectrum disorders: A meta-analysis. *Development and Psychopathology*, *26*(4), 933–945. https://doi.org/10.1017/S0954579414000479

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010* (pp. 94–101). https://doi.org/10.1109/CVPRW.2010.5543262

Lundeberg, M. A., Fox, P. W., & Punćochaŕ, J. (1994). Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments. *Journal of Educational Psychology*, *86*(1), 114–121. https://doi.org/10.1037/0022-0663.86.1.114

Marshall, R., & Lee, C. (1998). A Cross-Cultural, Between-Gender Study of Extreme Response Style. *European Advances in Consumer Research*, *3*, 90–95.

Martino, D. J., Strejilevich, S. A., Fassi, G., Marengo, E., & Igoa, A. (2011). Theory of mind and facial emotion recognition in euthymic bipolar I and bipolar II disorders. *Psychiatry Research*. https://doi.org/10.1016/j.psychres.2011.04.033

Matsumoto, D. (1992). More evidence for the universality of a contempt expression. *Motivation and Emotion*. https://doi.org/10.1007/BF00992972

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*. https://doi.org/10.1080/02699930802204677

McDuff, D., Kodra, E., El Kaliouby, R., & LaFrance, M. (2017). A large-scale analysis of sex differences in facial expressions. *PLoS ONE*, *12*(4), 1–11. https://doi.org/10.1371/journal.pone.0173942

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*. https://doi.org/10.1006/nimg.1996.0016

McIntosh, Anthony Randal, & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. In *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2004.07.020

Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology*. https://doi.org/10.1016/j.jsp.2006.04.004

Mehrabian, A. (2007). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*. https://doi.org/10.1007/bf02686918

Milne, E., Swettenham, J., Hansen, P., Campbell, R., Jeffries, H., & Plaisted, K. (2002). High motion coherence thresholds in children with autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *43*(2), 255–263. https://doi.org/10.1111/1469-7610.00018

Moebert, T., & Lucke, U. (2019). E.V.A. – Emotionen Verstehen und Ausdrücken. In N. Pinkwart & J. Konert (Eds.), *DELFI 2019* (pp. 289–290). Gesellschaft für Informatik e.V.

Moebert, T., Schneider, J. N., Zoerner, D., Tscherejkina, A., & Lucke, U. (2019). 4. How to use socio-emotional signals for adaptive training. In M. Augstein, E. Herder, & W. Wörndl (Eds.), *Personalized Human-Computer Interaction*. https://doi.org/10.1515/9783110552485-004

Montagne, B., Kessels, R. P. C., Frigerio, E., De Haan, E. H. F., & Perrett, D. I. (2005). Sex differences in the perception of affective facial expressions: Do men really lack emotional sensitivity? *Cognitive Processing*, *6*(2), 136–141.

https://doi.org/10.1007/s10339-005-0050-6

Montirosso, R., Peverelli, M., Frigerio, E., Crespi, M., & Borgatti, R. (2010). The development of dynamic facial expression recognition at different intensities in 4- to 18-year-olds. *Social Development*, *19*(1), 71–92. https://doi.org/10.1111/j.1467-9507.2008.00527.x

Mortel, T. Van de. (2008). Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing*.

Mulligan, K., & Scherer, K. R. (2012). Toward a working definition of emotion. *Emotion Review*. https://doi.org/10.1177/1754073912445818

Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2014.2317187

Nackaerts, E., Wagemans, J., Helsen, W., Swinnen, S. P., Wenderoth, N., & Alaerts, K. (2012). Recognizing Biological Motion and Emotions from Point-Light Displays in Autism Spectrum Disorders. *PLoS ONE*, *7*(9). https://doi.org/10.1371/journal.pone.0044473

Nelson, N. L., & Russell, J. A. (2013). Universality revisited. *Emotion Review*. https://doi.org/10.1177/1754073912457227

Newen, A., Welpinghus, A., & Juckel, G. (2015). Emotion recognition as pattern recognition: The relevance of perception. *Mind and Language*. https://doi.org/10.1111/mila.12077

Norman, G. J., Necka, E., & Berntson, G. G. (2016). The Psychophysiology of Emotions. In *Emotion Measurement* (pp. 83–98). Elsevier. https://doi.org/10.1016/B978-0-08-100508-8.00004-7

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological Research on the internet. *Journal of Social Issues*. https://doi.org/10.1111/1540-4560.00254

Palermo, R., & Coltheart, M. (2004). Photographs of facial expression: Accuracy, response times, and ratings of intensity. *Behavior Research Methods, Instruments, and Computers*. https://doi.org/10.3758/BF03206544

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*(1), 151–160. https://doi.org/10.1016/j.concog.2014.06.007

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721414531598

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon mechanical turk. *Judgment and Decision Making*.

Pelphrey, K. A., Morris, J. P., Mccarthy, G., & Labar, K. S. (2007). Perception of dynamic changes in facial affect and identity in autism. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsm010

Perdikis, D., Volhard, J., Müller, V., Kaulard, K., Brick, T. R., Wallraven, C., & Lindenberger, U. (2017). Brain synchronization during perception of facial emotional expressions with natural and unnatural dynamics. *PloS One*, *12*(7), e0181225. https://doi.org/10.1371/journal.pone.0181225

Phillips, L. H., & Allen, R. (2004). Adult aging and the perceived intensity of emotions in faces and stories. *Aging Clinical and Experimental Research*, *16*(3), 190–199. https://doi.org/10.1007/BF03327383

Pollick, F. E., Hill, H., Calder, A., & Paterson, H. (2003). Recognising facial expression from spatially and temporally modified movements. *Perception*. https://doi.org/10.1068/p3319

Quine, W. V. (1969). Natural Kinds. In *Essays in Honor of Carl G. Hempel*. Springer. https://doi.org/10.1007/978-94-017-1466-2

Riby, D. M., & Hancock, P. J. B. (2009). Do faces capture the attention of individuals with Williams syndrome or autism? Evidence from tracking eye movements. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-008-0641-z

Riediger, M., Schmiedek, F., Wagner, G. G., & Lindenberger, U. (2009). Seeking pleasure and seeking pain: Differences in prohedonic and contra-hedonic motivation from adolescence to old age. *Psychological Science*, *20*(12), 1529–1535. https://doi.org/10.1111/j.1467-9280.2009.02473.x

Riediger, M., Voelkle, M. C., Ebner, N. C., & Lindenberger, U. (2011). Beyond "Happy, angry, or sad?": Age-of-poser and age-of-rater effects on multi-dimensional emotion perception. *Cognition and Emotion*, *25*(6), 968–982. https://doi.org/10.1080/02699931.2010.540812

Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., … Alcañiz, M. (2007). Affective interactions using virtual reality: The link between presence and emotions. *Cyberpsychology and Behavior*. https://doi.org/10.1089/cpb.2006.9993

Robertson, C. E., Thomas, C., Kravitz, D. J., Wallace, G. L., Baron-Cohen, S., Martin, A., & Baker, C. I. (2014). Global motion perception deficits in autism are reflected as early as primary visual cortex. *Brain*, *137*(9), 2588–2599. https://doi.org/10.1093/brain/awu189

RStudio Team. (2015). RStudio: Integrated Development for R. RStudio. Retrieved from https://www.rstudio.com/

Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2008.01.001

Rump, K. M., Giovannelli, J. L., Minshew, N. J., & Strauss, M. S. (2009). The development of emotion recognition in individuals with autism. *Child Development*, *80*(5), 1434–1447. https://doi.org/10.1111/j.1467-8624.2009.01343.x

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A. (2003a). Core affect and the psychological construction of emotion. *Psychological Review*.

Russell, J. A. (2003b). Core affect annd the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805–819. https://doi.org/10.1037/0022-3514.76.5.805

Rutherford, M. D., & Troje, N. F. (2012). IQ predicts biological motion perception in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *42*(4), 557–565. https://doi.org/10.1007/s10803-011-1267-0

Sato, W., Fujimura, T., & Suzuki, N. (2008). Enhanced facial EMG activity in response to dynamic facial expressions. *International Journal of Psychophysiology*, *70*(1), 70–74. https://doi.org/10.1016/j.ijpsycho.2008.06.001

Sato, W., & Yoshikawa, S. (2004). BRIEF REPORT The dynamic aspects of emotional facial expressions. *Cognition & Emotion*, *18*(5), 701–710. https://doi.org/10.1080/02699930341000176

Sato, W., & Yoshikawa, S. (2007). Enhanced experience of emotional arousal in response to dynamic facial expressions. *Journal of Nonverbal Behavior*, *31*(2), 119–135. https://doi.org/10.1007/s10919-007-0025-7

Saygin, A. P., Cook, J., & Blakemore, S. J. (2010). Unaffected perceptual thresholds for biological and non-biological form-from-motion perception in autism spectrum conditions. *PLoS ONE*, *5*(10). https://doi.org/10.1371/journal.pone.0013491

Scarantino, A. (2009). Core Affect and Natural Affective Kinds. *Philosophy of Science*, *76*(5), 940–957. https://doi.org/10.1086/605816

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*. https://doi.org/10.1177/0539018405058216

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*. https://doi.org/10.1080/02699930902928969

Schimmack, U. (2010). What multi-method data tell us about construct validity. *European Journal of Personality*. https://doi.org/10.1002/per.771

Schimmack, U., & Grob, A. (2000). Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*. https://doi.org/10.1002/1099-0984(200007/08)14:4<325::AID-PER380>3.0.CO;2-I

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/0146167208317771

Schneider, J. N., & Brick, T. R. (2019). plsr: Pleasure - Partial Least Squares Analysis with Permutation Testing. Retrieved from https://cran.r-project.org/package=plsr

Schneider, J. N., Brick, T. R., & Dziobek, I. (n.d.). Distance to the neutral face predicts arousal ratings of facial expressions in neurotypical and autistic individuals.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition and Emotion*. https://doi.org/10.1080/026999300402745

Shah, P., Bird, G., & Cook, R. (2016). Face processing in autism: Reduced integration of cross-feature dynamics. *Cortex*, *75*, 113–119. https://doi.org/10.1016/j.cortex.2015.11.019

Shuman, V., Sander, D., & Scherer, K. R. (2013). Levels of valence. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00261

Skidmore, S. T., & Thompson, B. (2010). Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164410379320

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment*. https://doi.org/10.1037/1040-3590.17.4.396

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science*, *16*(3), 184–189. https://doi.org/10.1111/j.0956-7976.2005.00801.x

Stemmler, G. (2004). Physiological processes during emotion. In *The Regulation of Emotion*. https://doi.org/10.4324/9781410610898

Stemmler, G., Aue, T., & Wacker, J. (2007). Anger and fear: Separable effects of emotion and motivational direction on somatovisceral responses. *International Journal of Psychophysiology*. https://doi.org/10.1016/j.ijpsycho.2007.03.019

Teglasi, H., Simcox, A. G., & Kim, N. Y. (2007). Personality constructs and measures. *Psychology in the Schools*. https://doi.org/10.1002/pits.20218

Theobald, B.-J., Matthews, I., Mangini, M., Spies, J. R., Brick, T. R., Cohn, J. F., & Boker, S. M. (2009). Mapping and manipulating facial expression. *Language and Speech*, *52*(Pt 2-3), 369–386. https://doi.org/10.1177/0023830909103181

Thomas, L. A., De Bellis, M. D., Graham, R., & LaBar, K. S. (2007). Development of emotional facial recognition in late childhood and adolescence. *Developmental Science*, *10*(5), 547–558. https://doi.org/10.1111/j.1467-7687.2007.00614.x

Thompson, A. E., & Voyer, D. (2014). Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis. *Cognition and Emotion*, *28*(7), 1164–1195. https://doi.org/10.1080/02699931.2013.875889

Thompson, V. A. (2014). Chapter Two – What Intuitions Are… and Are Not. In *Psychology of Learning and Motivation* (Vol. 60, pp. 35–75). https://doi.org/10.1016/B978-0-12-800090-8.00002-0

Trampe, D., Quoidbach, J., & Taquet, M. (2015). Emotions in everyday life. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0145450

Uljarevic, M., & Hamilton, A. (2013). Recognition of emotions in autism: A formal meta-analysis. *Journal of Autism and Developmental Disorders*, *43*(7), 1517–1526. https://doi.org/10.1007/s10803-012-1695-5

Waller, B. M., Cray, J. J., & Burrows, A. M. (2008). Selection for Universal Facial Emotion. *Emotion*. https://doi.org/10.1037/1528-3542.8.3.435

Walsh, F. A. (1927). The Logic of Modern Physics. *New Scholasticism*, *1*(4), 364–367. https://doi.org/10.5840/newscholas19271413

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect. *The PANAS Scales*, *54*, 1063-1070. https://doi.org/http://dx.doi.org/10.1037/0022-3514.54.6.1063

Winston Chang. (2018). Package “shiny” Type Package Title Web Application Framework for R.

Wisniak, J. (2000). The Thermometer?From The Feeling To The Instrument. *The Chemical Educator*. https://doi.org/10.1007/s00897990371a

World Health Organization. (1993). The ICD-10 Classification of Mental and Behavioural Disorders. Diagnostic criteria for research. *International Classification*, *10*, 1–267. https://doi.org/10.1002/1520-6505(2000)9:5<201::AID-EVAN2>3.3.CO;2-P

Wundt, W. (1908). Outlines of psychology (1897). In *Found. Psychol. thought A Hist.*

Psychol. https://doi.org/10.1007/978-1-4684-8340-6_7

Zadeh, A., Lim, Y. C., Baltrušaitis, T., & Morency, L. P. (2018). Convolutional experts constrained local model for 3D facial landmark detection. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*. https://doi.org/10.1109/ICCVW.2017.296

7. Supplementary Materials

Supplementary Table 1. Emotion categories of the video data set. The original German label (left column) and the English translation (right column) are shown for all 40 categories of the video data sets. The third column indicates if this emotion category was also used in Study 2 of Project 2.

Original German term	English translation	**Included in Study 2**
amüsiert	amused	
angeekelt	disgusted	x
angstvoll	fearful	x
ärgerlich	angry	x
beleidigt	offended	
betroffen	affected	
beunruhigt	troubled	
dankbar	grateful	
eifersüchtig	jealous	
enthusiastisch	enthusiastic	x
entschuldigend	apologetic	
entsetzt	aghast	x
enttäuscht	disappointed	x
erleichtert	relieved	
erwartungsvoll	expectant	
frustriert	frustrated	
gekränkt	aggrieved	
gelangweilt	bored	x
heiter	happy	x
interessiert	interested	x
melancholisch	melancholic	x
mitleidig	compassionate	x
neidisch	envious	x
neugierig	curious	x
schuldig	guilty	
schwärmerisch	lyrical	x
stolz	proud	
traurig	sad	x
überrascht	surprised	x
verachtend	contemptuous	x
vergebend	pardoning	x
verlegen	embarrassed	
verliebt	in love	x
verwirrt	confused	
verzweifelt	desperate	x
wehmütig	wistful	
wütend	furious	
zufrieden	content	
zuversichtlich	confident	x
zweifelnd	doubtful	

$$Rating\ points_{ijk} = \beta_0 + \beta_0 participant_i + \beta_0 video_k + \beta_1 * \left(actor\ sex_j\right) + \epsilon_{ijk} \quad (1)$$

$$Rating\ points_{ijk} = \beta_0 + \beta_0 participant_i + \beta_0 video_k + \beta_1 * \left(actor\ sex_j\right) + \epsilon_{ijk}$$

$$Rating\ standard\ deviation_{ijk} = \beta_0 + \beta_0 participant_i + \beta_0 video_k + \beta_1 * \left(actor\ sex_j\right) + \epsilon_{ijk} \quad (2)$$

$$Rating\ points\ on\ target_{ijk} = \beta_0 + \beta_0 participant_i + \beta_0 video_k + \beta_1 * \left(participant\ sex_j\right) + \epsilon_{ijk} \quad (3)$$

$$Rating\ standard\ deviation_{ijk} = \beta_0 + \beta_0 participant_i + \beta_0 video_k + \beta_1 * (participant\ sex_i) + \epsilon_{ijk} \quad (4)$$

$$SRD = \beta_0 + \beta_1 * (video\ mean\ valence) + \beta_2 * (video\ mean\ valence^2) + \epsilon \quad (5)$$

$$OD = \beta_0 + \beta_1 * (video\ mean\ valence) + \beta_2 * (video\ mean\ valence^2) + \epsilon \quad (6)$$

$$SRD = \beta_0 + \beta_1 * (video\ mean\ arousal) + \beta_2 * (video\ mean\ arousal^2) + \epsilon \quad (7)$$

$$OD = \beta_0 + \beta_1 * (video\ mean\ arousal) + \beta_2 * (video\ mean\ arousal^2) + \epsilon \quad (8)$$

Supplementary Table 2. Mixed effects models to check for a systematic effect of observer sex and observer age on rating scales with a random effect for video.

	Dependent variable:						
	angry	happy	sad	disgusted	surprised	interested	fearful
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Observer female	**3.52*****	-1.18	**-2.62****	**-4.49*****	**-2.91****	**-2.18***	**-3.31*****
	[-4.99, -2.06]	[-2.28, -0.07]	[-4.11, -1.12]	[-5.96, -3.03]	[-4.44, -1.38]	[-3.71, -0.65]	[-4.80, -1.83]
Observer age	-0.01	0.03	-0.05	-0.08	-0.10	0.04	0.06
	[-0.09, 0.07]	[-0.03, 0.09]	[-0.13, 0.03]	[-0.16, -0.002]	[-0.18, -0.02]	[-0.04, 0.13]	[-0.03, 0.14]
Intercept	36.98	31.37	38.15	32.69	48.25	51.05	31.69
	[33.69, 40.26]	[28.16, 34.58]	[34.78, 41.52]	[29.54, 35.84]	[44.95, 51.55]	[47.94, 54.16]	[28.47, 34.91]
Random effects	SD	SD	SD	SD	SD	SD	SD
Video	21.37	27.99	22.12	18.65	19.58	15.18	19.51
Residual	23.66	17.75	24.19	23.69	24.83	24.90	24.01
Observations	5,292	5,292	5,292	5,292	5,292	5,292	5,292
Log Likelihood	-24,798.57	-23,529.13	-24,922.48	-24,747.72	-24,996.59	-24,909.91	-24,831.99
Akaike Inf. Crit.	49,607.14	47,068.25	49,854.97	49,505.44	50,003.17	49,829.83	49,673.98
Bayesian Inf. Crit.	49,640.01	47,101.12	49,887.84	49,538.31	50,036.04	49,862.70	49,706.85

Note: *p<0.05; **p<0.01; ***p<0.001

Supplementary Table 3. Mixed effects models to investigate the influence of the actor sex on the total rating points assigned per video and the standard deviation across rating scales with random effects for video and observer.

	Dependent variable:	
Fixed effects	Total rating points (1)	Standard deviation across rating scales (2)
Actor sex female	23.93 [15.24, 32.61]	2.23 [1.19, 3.26]
Intercept	240.63 [232.29, 248.98]	28.58 [27.74, 29.41]
Random effects	SD	SD
Video	44.14	5.24
Observer	60.67	4.28
Residual	63.15	7.75
Observations	5,292	5,292
Log Likelihood	-30,415.14	-19,094.75
Akaike Inf. Crit.	60,840.27	38,199.51
Bayesian Inf. Crit.	60,873.14	38,232.38
Note:	p-values for these exploratory analyses are intentionally not provided	

Supplementary Table 4. Mixed effects models to investigate the influence of the observer sex on the total rating points assigned per video, the standard deviation across rating scales and the rating points spent on the target emotion with random effects for video and observer.

	Dependent variable:		
	Total rating points	Standard deviation across rating scales	Rating points spent on target emotion
Fixed effects	(1)	(2)	(3)
Participant female	-19.01	1.02	2.57
	[-31.95, -6.07]	[0.02, 2.02]	[-0.84, 5.99]
Intercept	266.08	28.97	79.51
	[254.46, 277.70]	[28.01, 29.94]	[76.02, 83.00]
Random effects	SD	SD	SD
Video	45.72	5.35	9.19
Observer	60.14	4.26	6.75
Residual	63.14	7.75	20.17
Observations	5,292	5,292	911
Log Likelihood	-30,424.79	-19,101.54	-4,116.97
AIC	60,859.57	38,213.09	8,243.95
BIC	60,892.44	38,245.96	8,268.02
Note:	p-values for these exploratory analyses are intentionally not provided		

Supplementary Table 5. Pearson correlation of individual participant ratings and self-rated difficulty (SRD).

	happy	sad	surprised	disgusted	angry	fearful	interested	valence	arousal	SRD
happy	1	-0.50	0.18	-0.38	-0.50	-0.36	0.46	0.87	0.01	-0.12
sad	-0.50	1	-0.01	0.25	0.32	0.53	-0.15	-0.53	0.03	0.12
surprised	0.18	-0.01	1	0.15	0.02	0.24	0.45	0.11	0.43	0.01
disgusted	-0.38	0.25	0.15	1	0.48	0.36	-0.13	-0.45	0.17	0.11
angry	-0.50	0.32	0.02	0.48	1	0.21	-0.15	-0.55	0.16	0.11
fearful	-0.36	0.53	0.24	0.36	0.21	1	0.01	-0.39	0.30	0.14
interested	0.46	-0.15	0.45	-0.13	-0.15	0.01	1	0.40	0.26	-0.11
valence	0.87	-0.53	0.11	-0.45	-0.55	-0.39	0.40	1	-0.06	-0.10
arousal	0.01	0.03	0.43	0.17	0.16	0.30	0.26	-0.06	1	-0.02
difficulty	-0.12	0.12	0.01	0.11	0.11	0.14	-0.11	-0.10	-0.02	1

Supplementary Table 6. Pearson correlation of video mean ratings and self-rated difficulty (SRD).

	happy	sad	surprised	disgusted	angry	fearful	interested	valence	arousal	SRD
happy	1	-0.69	0.19	-0.63	-0.73	-0.55	0.65	0.97	-0.02	-0.24
sad	-0.69	1	-0.18	0.29	0.35	0.65	-0.43	-0.70	-0.005	0.20
surprised	0.19	-0.18	1	0.08	-0.15	0.25	0.58	0.15	0.66	-0.03
disgusted	-0.63	0.29	0.08	1	0.61	0.42	-0.40	-0.69	0.24	0.13
angry	-0.73	0.35	-0.15	0.61	1	0.19	-0.46	-0.75	0.16	0.16
fearful	-0.55	0.65	0.25	0.42	0.19	1	-0.14	-0.57	0.42	0.21
interested	0.65	-0.43	0.58	-0.40	-0.46	-0.14	1	0.63	0.39	-0.19
valence	0.97	-0.70	0.15	-0.69	-0.75	-0.57	0.63	1	-0.11	-0.19
arousal	-0.02	-0.005	0.66	0.24	0.16	0.42	0.39	-0.11	1	-0.09
difficulty	-0.24	0.20	-0.03	0.13	0.16	0.21	-0.19	-0.19	-0.09	1

Supplementary Table 7. Regression models for the NT and ASD group predicting valence ratings from distance and speed.

	Dependent variable:			
	Valence rating NT	Valence rating ASD	Valence rating NT standardized	Valence rating ASD standardized
	(1)	(2)	(3)	(4)
Distance [Z-Score]	2.66	-0.19	0.09	-0.01
	[-4.61, 9.92]	[-6.51, 6.13]	[-0.16, 0.35]	[-0.27, 0.25]
Speed [Z-Score]	0.09	1.49	0.003	0.06
	[-7.18, 7.36]	[-4.84, 7.81]	[-0.26, 0.26]	[-0.20, 0.32]
Intercept	43.47***	44.12***	0.00	0.00
	[37.25, 49.68]	[38.71, 49.53]	[-0.22, 0.22]	[-0.22, 0.22]
Observations	80	80	80	80
R^2	0.01	0.003	0.01	0.003
Adjusted R^2	-0.02	-0.02	-0.02	-0.02
Residual Std. Error (df = 77)	28.37	24.68	1.01	1.01
F Statistic (df = 2; 77)	0.36	0.13	0.36	0.13
Note:				*p<0.05; **p<0.01; ***p<0.001

Supplementary Table 8. Regression models predicting differences in valence ratings between the NT and ASD group from distance to the neutral face and speed.

	Dependent variable:	
	Arousal difference (1)	Arousal difference (2)
Distance [Z-Score]	2.84 [-0.35, 6.04]	0.23 [-0.03, 0.48]
Speed [Z-Score]	-1.40 [-4.59, 1.80]	-0.11 [-0.37, 0.14]
Intercept	-0.65 [-3.38, 2.08]	-0.00 [-0.22, 0.22]
Observations	80	80
R^2	0.04	0.04
Adjusted R^2	0.01	0.01
Residual Std. Error (df = 77)	12.46	0.99
F Statistic (df = 2; 77)	1.53	1.53
Note:		*p<0.05; **p<0.01; ***p<0.001

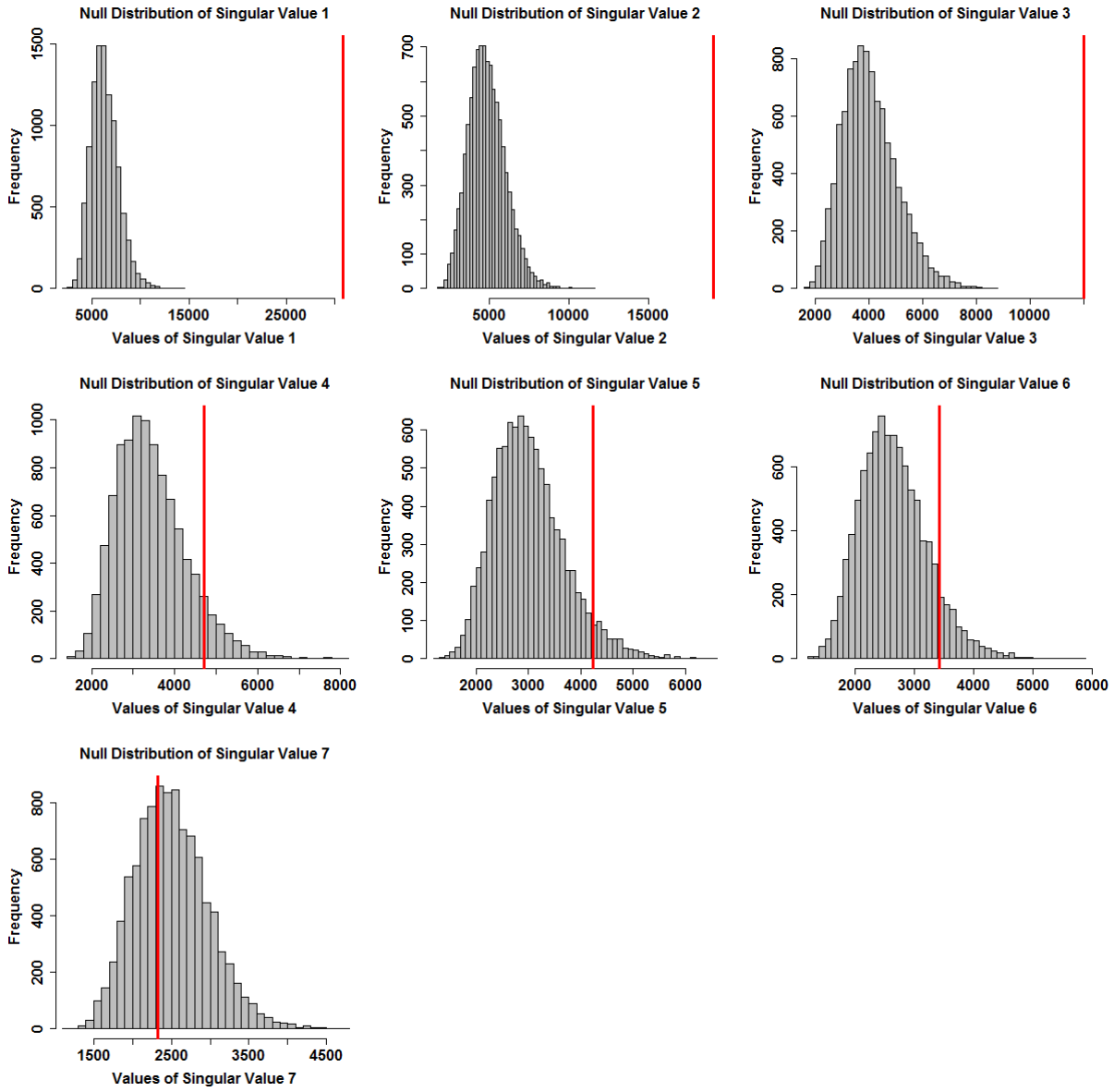
Supplementary Table 9. Regression models with distance to the neutral face and speed predictors tested together and individually for the HAT group.

	Dependent variable:					
	Arousal ratings			Arousal ratings standardized		
	(1)	(2)	(3)	(4)	(5)	(6)
Distance [Z-Score]	**6.15****	**8.24*****		**0.31****	**0.42*****	
	[1.64, 10.66]	[4.30, 12.17]		[0.08, 0.55]	[0.22, 0.62]	
Speed [Z-Score]	4.10		**7.23*****	0.21		**0.37*****
	[-0.41, 8.61]		[3.20, 11.26]	[-0.02, 0.44]		[0.16, 0.58]
Intercept	42.96***	42.96***	42.96***	-0.00	-0.00	-0.00
	[39.10, 46.82]	[39.05, 46.87]	[38.95, 46.96]	[-0.20, 0.20]	[-0.20, 0.20]	[-0.20, 0.20]
Observations	80	80	80	80	80	80
R^2	0.20	0.17	0.13	0.20	0.17	0.13
Adj. R^2	0.18	0.16	0.12	0.18	0.16	0.12
Residual Std. Error	17.66 (df = 77)	17.87 (df = 78)	18.33 (df = 78)	0.90 (df = 77)	0.92 (df = 78)	0.94 (df = 78)
F Statistic	9.77*** (df = 2; 77)	16.26*** (df = 1; 78)	11.64** (df = 1; 78)	9.77*** (df = 2; 77)	16.26*** (df = 1; 78)	11.64** (df = 1; 78)

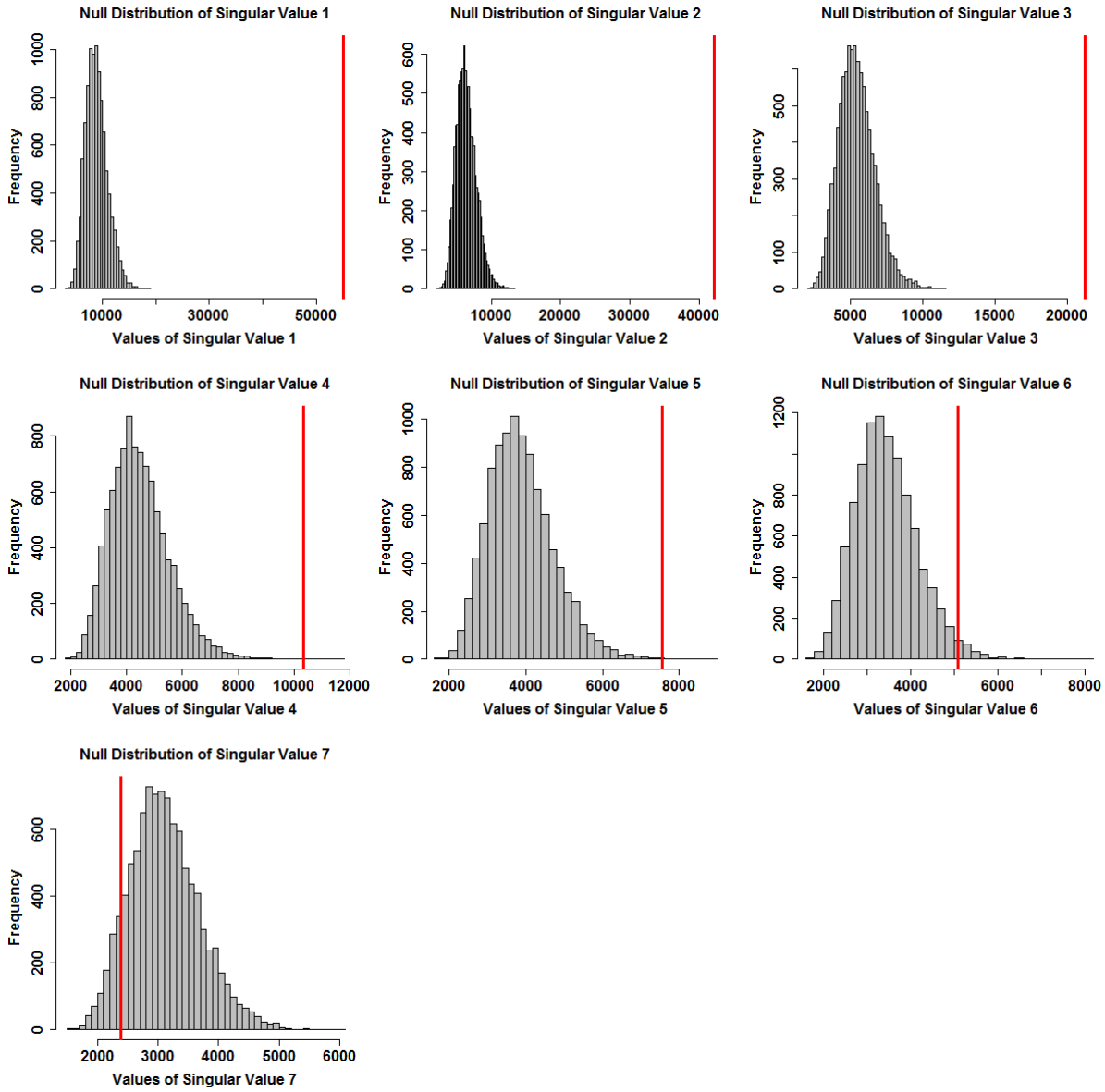
Note: *p<0.05; **p<0.01; ***p<0.001

Supplementary Table 10. Regression models predicting differences in arousal ratings between the NT and HAT group from distance to the neutral face and speed.

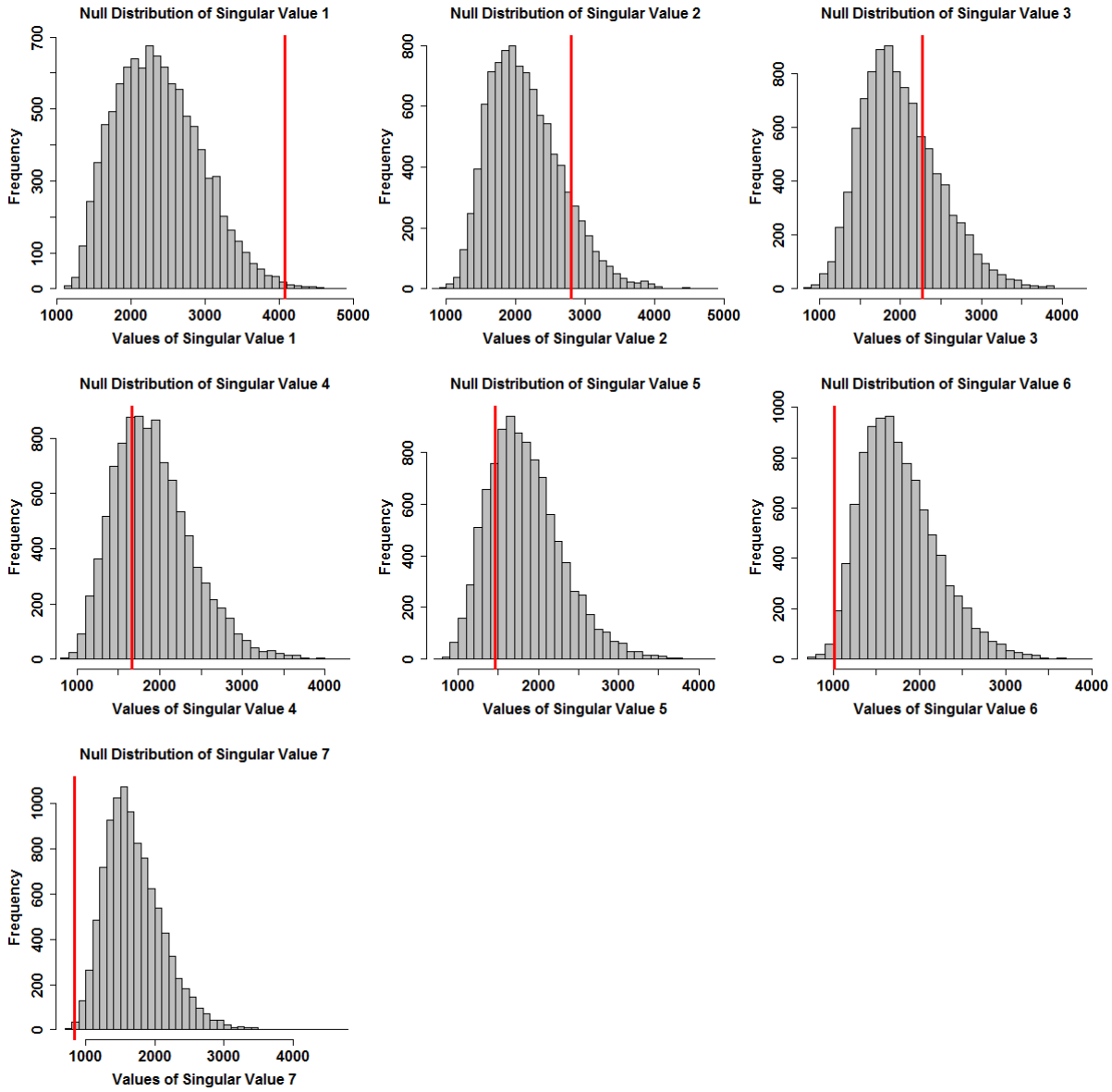
	Dependent variable:	
	Arousal difference	Arousal difference standardized
	(1)	(2)
Distance [Z-Score]	-0.22	-0.02
	[-3.13, 2.69]	[-0.28, 0.24]
Speed [Z-Score]	-1.57	-0.14
	[-4.48, 1.34]	[-0.39, 0.12]
Intercept	**4.64*****	0.00
	[2.15, 7.13]	[-0.22, 0.22]
Observations	80	80
R^2	0.02	0.02
Adjusted R^2	-0.003	-0.003
Residual Std. Error (df = 77)	11.36	1.00
F Statistic (df = 2; 77)	0.88	0.88
Note:		*p<0.05; **p<0.01; ***p<0.001



Supplementary Figure 1. Null distributions of singular values of the within-group model for the male group. The vertical red line marks the location of the actual singular value of the model within the null distributions found through permutation of the original data.



Supplementary Figure 2. Null distributions of singular values of the within-group model for the female group. The vertical red line marks the location of the actual singular value of the model within the null distributions found through permutation of the original data.



Supplementary Figure 3. Null distributions of singular values of the between-group model. The vertical red line marks the location of the actual singular value of the model within the null distributions found through permutation of the original data.

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorliegende Dissertation selbstständig und nur mit den angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Dissertation wurde bei keiner anderen Hochschule in gleicher oder ähnlicher Form eingereicht.

Jan Niklas Schneider, Potsdam, den 22. Oktober 2019